T.C.
ULUDAĞ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI

(DOKTORA TEZİ)

Sadullah ÇELİK

BURSA-2018

T.C.
ULUDAĞ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI

(DOKTORA TEZİ)

Sadullah ÇELİK

Danışman

Prof. Dr. Mustafa AYTAÇ

BURSA-2018

T.C.

ULUDAĞ ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Ekonometri Anabilim Dalı, İstatistik Bilim Dalı'nda 711317001 numaralı Sadullah ÇELİK'in hazırladığı "**BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI**" konulu Doktora Tezi ile ilgili tez savunma sınavı, 27/02/2018 günü 14:00 – 15:00 saatleri arasında yapmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin/çalışmasının ..... (başarılı/başarısız) olduğuna ..... (oybirliği/oy çokluğu) ile karar verilmiştir.

Üye (Tez Danışmanı ve Sınav Komisyon Başkanı)
Prof. Dr. Mustafa AYTAÇ
Uludağ Üniversitesi

Üye
Prof. Dr. Şahamet BÜLBÜL
Marmara Üniversitesi

Üye
Prof. Dr. Ayşe OĞUZLAR
Uludağ Üniversitesi

Üye
Prof. Dr. Aysan ŞENTÜRK
Uludağ Üniversitesi

Üye
Prof. Dr. Dilek ALTAŞ
Marmara Üniversitesi

27/02/2018

YEMİN METNİ

Doktora tezi olarak sunduğum **"BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI"** başlıklı çalışmamın bilimsel araştırma, yazma ve etik kurallarına uygun olarak tarafımdan yazıldığına ve tezde yapılan bütün alıntıların kaynaklarının usulüne uygun olarak gösterildiğine, tezimde intihal ürünü cümle veya paragraflar bulunmadığına şerefim üzerine yemin ederim.

27/02/2018

Adı Soyadı : Sadullah ÇELİK

Öğrenci No : 711317001

Anabilim Dalı : Ekonometri

Programı : İstatistik

Statüsü : Doktora

SOSYAL BİLİMLER ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA İNTİHAL YAZILIM RAPORU

ULUDAĞ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 02/02/2018

Tez Başlığı / Konusu: BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve d) Sonuç kısımlarından oluşan toplam 191 sayfalık kısmına ilişkin, 02/02/2018 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından (Turnitin)* aşağıda belirtilen filtrelemeler uygulanarak alınmış olan özgünlük raporuna göre, tezimin benzerlik oranı %5 'tir.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar dahil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Uludağ Üniversitesi Sosyal Bilimler Enstitüsü Tez Çalışması Özgünlük Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

02/02/2018 İmza

Adı Soyadı: Sadullah ÇELİK
Öğrenci No: 711317001
Anabilim Dalı: Ekonometri
Programı: Doktora
Statüsü: ☐ Y.Lisans ☒ Doktora

Danışman
Prof. Dr. Mustafa AYTAÇ
02/02/2018 İmza

*** Turnitin programına Uludağ Üniversitesi Kütüphane web sayfasından ulaşılabilir.**

ÖZET

Yazarlar	: Sadullah Çelik
Üniversite	: Uludağ Üniversitesi
Enstitü	: Sosyal Bilimler Enstitüsü
Anabilim Dalı	: Ekonometri
Bilim Dalı	: İstatistik
Tezin Niteliği	: Doktora Tezi
Sayfa Sayısı	: XIX+168
Mezuniyet Tarihi	: 27/02/2018
Tez Danışmanı	: Prof. Dr. Mustafa AYTAÇ

BÜYÜK VERİ VE İSTATİSTİKTEKİ UYGULAMALARI

Son yıllarda bilgisayar ve bulut teknolojilerinde görülen modern gelişmeler üretilen ve saklanan bilginin miktarında ve hızında büyük artışa sebep oldu. Bilgi miktarındaki bu artış "Büyük Veri" olarak adlandırılan yeni bir kavramın hayatımıza girmesini sağladı. Büyük Veri özellikle hükûmetlerin ve işletmelerin gelecekleriyle ilgili bilinçli kararları almaları konusunda büyük avantaj sağlar. Fakat ortaya çıkan bu verinin büyüklüğü ve çeşitliliği beraberinde bazı sorunları da getirdi. Geleneksel veritabanı sistemleri bu veri formatlarını işlemekte yetersiz kaldığından, bu sorunun üstesinden gelmek için yeni araç ve tekniklere ihtiyaç duyulmaktadır. Bugün çok çeşitli teknik ve teknolojiler "Büyük Veri"yi; toplamak, işlemek, analiz etmek ve görselleştirmek için geliştirilmiştir. Bu teknik ve teknolojiler; istatistik, bilgisayar bilimi, uygulamalı matematik ve ekonomi gibi birçok alanı kapsamakta ve bunlardan yararlanmaktadır. Bu çalışmada Google'ın altyapısında bulunan BigQuery'deki GDELT veri seti kullanılarak 1979-2017 yılları arasında dünyada yaşanan çatışma olayları ile Türkiye ve Ukrayna'da yaşanan protestolar SQL yardımıyla gerçek zamanlı olarak analiz edilmiştir. Analiz sonucunda elde edilen verilerden 1979-2017 yılları arasında dünyada yaşanan çatışmalar ile Türkiye ve Ukrayna'da yaşanan protestolar grafikler şeklinde sunulmuştur. Ayrıca, analiz sonucunda elde edilen çatışma ve protesto verilerinin kuvvet yasasına uygun olup olmadığı test edilmiş ve test sonuçlarında bu verilerin kuvvet yasasına uygun bir dağılım gösterdikleri bulunmuştur. Elde edilen bu bulgulardan, çatışma ve protesto gibi toplumsal olayların kuvvet yasasına uygun bir dağılım sergiledikleri söylenebilir.

Anahtar Sözcükler: Büyük Veri, Spark, Hadoop, BigQuery, GDELT, Kuvvet Yasası Dağılımı, Çatışma, Protesto.

ABSTRACT

Name and Surname	: Sadullah ÇELİK
University	: Uludağ University
Institution	: Social Science Institution
Field	: Econometrics
Branch	: Statistics
Degree awarded	: PhD
Page Number	: XIX+168
Degree Date	: 27/02/2018
Supervisor	: Prof. Dr. Mustafa AYTAÇ

BIG DATA AND APPLICATIONS IN STATISTICS

Developments in computer and cloud technologies in recent years have led to a significant increase in the amount and speed of information generated and stored. This increase in the amount of information enabled a new concept called "Big Data" to enter into our lives. The Big Data gives a big advantage especially for governments and businesses to make informed decisions about their future. But the size and diversity of the resulting data also emerged some problems. Since traditional database systems are insufficient to handle these data formats, new tools and techniques are needed to overcome this problem. Today, a wide range of techniques and technologies are being developed to manipulate, analyse and visualize “Big Data”. These techniques and technologies comprise and use statistics, computer science, applied mathematics and economics. In this study, using the GDELT dataset at BigQuery in Google's infrastructure, the protesters living in Turkey and Ukraine and the conflict events in the world between 1979-2017 were analysed in real time with SQL. The data obtained as a result of the analysis are presented in the form of graphs of conflicts in the world between 1979 and 2017 and protests in Turkey and Ukraine. In addition, it was tested if the conflicts and protest data obtained as the result of the analysis were in accordance with the power law, and it was found that these data were distributed according to the power law in the test results. From these findings, it can be said that social events such as conflict and protest have a distribution compatible with power law.

Keywords: Big Data, Hadoop, Spark, BigQuery, GDELT, Power Law Distribution, Conflict, Protest.

ÖNSÖZ

Bu çalışma değerli hocam Prof. Dr. Mustafa AYTAÇ'ın Büyük Veri konusunda çalışmamı teşvik etmesi ile hayat bulmuş ve verilen yoğun çabalar sonucunda bugünkü şeklini almıştır. Öncelikle danışmanlığımı yapan ve her konuda yardımını benden esirgemeyen Prof. Dr. Mustafa AYTAÇ'a teşekkürü bir borç bilirim. Doktora tez savunma sınavı jüri üyeleri olan Prof. Dr. Ayşe OĞUZLAR'a, Prof. Dr. Aysan ŞENTÜRK'e, Prof. Dr. Şahamet BÜLBÜL'e ve Prof. Dr. Dilek ALTAŞ'a teşekkürlerimi sunarım.

Hayatımın her anında yanımda olup benden maddi ve manevi desteklerini esirgemeyen aileme, bütün akraba ve dostlarıma çok teşekkür ederim.

Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümü Öğretim Üyeleri olan Prof. Dr. Necmi GÜRSAKAL'a, Prof. Dr. Nuran BAYRAM'a, Prof. Dr. Mustafa SEVÜKTEKİN'e, Prof. Dr. Kemal SEZEN'e, Doç. Dr. Sevda GÜRSAKAL'a, Doç. Dr. Mehmet ÇINAR'a, Arş. Gör. Dr. Kadriye Burcu ÖNGEN'e, Arş. Gör. Tuğba GÖKDEMİR'e, Ardahan Üniversitesi İktisadi ve İdari Bilimler Fakültesi Öğretim Görevlisi Murat Yusuf KIZILKAYA'ya, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümü Doktora Öğrencisi Emrah AKDAMAR'a, Munzur Üniversitesi İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümü Arş. Gör. Fatma SERT'e, Bilgisayar Teknolojileri Öğretmeni Oktay BOYNUKARA'ya, Endüstri Mühendisi İbrahim GÜROĞLU'na, Matematikçi Gökhan ZEYTİN'e, Kimya Mühendisi Ferdi ESER'e desteklerinden dolayı teşekkür ederim.

BURSA 2018 **Sadullah ÇELİK**

İÇİNDEKİLER

sayfa

BİRİNCİ BÖLÜM
BÜYÜK VERİ VE İSTATİSTİK

İKİNCİ BÖLÜM

BÜYÜK VERİ TEKNOLOJİLERİ VE TEKNİKLERİ

ÜÇÜNCÜ BÖLÜM
KUVVET YASASI DAĞILIMI

DÖRDÜNCÜ BÖLÜM
BIGQUERY İLE BÜYÜK VERİ UYGULAMALARI

TABLOLAR

ŞEKİLLER

SEMBOLLER

α: Ölçekleme parametresi
C: Normalleştirme sabiti
x_{min}: Kesme parametresi
D: Kolmogorov-Smirnov istatistiği
ζ: Zeta fonksiyonu
n: Gözlem sayısı
σ: Standart hata
μ: Ortalama
λ: Üssel parametre

KISALTMALAR

Bibliyografik Bilgiler	**Uluslararası**	**Türkçe**
Bakınız	V.	Bkz.
Cilt	Vol.	C.
Çeviren	trans. by	çev.
Editör/yayına hazırlayan	ed.	ed. veya haz.
Kesikli	disc.	kes.
Numara	No.	Num.
Sayfa/sayfalar	p./pp.	s./ss.
Sıklık	freq.	sık.
Sürekli	cont.	sür.

SÖZLÜK

İngilizce-Türkçe Terimler Sözlüğü

İngilizce	Türkçe
A Swarm of Sensors	Sensörler Kümesi
Association Rule Learning	İlişkili Kurallı Öğrenme
Big Table	Büyük Tablo
Big Transaction Data	Büyük İşlem Verileri
BigQuery	Büyük Sorgulama
Biometrics Data	Biometrik Veri
Chief Statisticians	Baş İstatistikçiler
Cloud Computing	Bulut Bilişim
Cluster Analysis	Kümeleme Analizi
Crowdsourcing	Kalabalığın Gücü
Dark Web	Karanlık Web
Data Deluge	Veri Akışı
Data Fusion and Data Integration	Veri Füzyonu (Kaynaştırma-Birleştirme) ve Veri Entegrasyonu
Data Mining	Veri Madenciliği
DataNode	Veri Düğümü
Data Science	Veri Bilimi
Data Scientist	Veri Bilimci
Data Warehouse	Veri Ambarı
Deep Web	Derin Web
Distributed System	Dağıtık Sistem
Dynamo	Dinamo
Ensemble Learning	Toplu Öğrenme
Genetic Algorithms	Genetik Algoritmalar
Google File System	Google Dosya Sistemi
Human-generated Data	İnsanların Ürettiği Veri
Hadoop Common	Hadoop Ortak
Internet of Things	Nesnelerin İnterneti

Intrinsic Sample Varians	İçsel Örneklem Varyansı
Job Tracker	İş İzleyici
Locational Data	Konumsal Veri
Unlimited Computing Power	Sınırsız İşlem Gücü
Web Data & Social Media	Web Verisi & Sosyal Medya
Machine Learning	Makine Öğrenme
Machine-to-machine Data	Makineden Makineye Veri
MapReduce	EşleİndirgeEşleİndirge
Master/Slave Node	Ana/Bağlayıcı Düğümü
Metadata	Metaveri
NameNode	Ad Düğümü
Natural Language Processing-(NLP)	Doğal Dil İşleme
Neural Networks	Sinir Ağları
Network Analysis	Ağ Analizi
Optimization	Optimizasyon
Pattern Recognition	Örüntü (Desen) Tanıma
Power Law Distribution	Kuvvet Yasası Dağılımı
Predictive Modeling	Öngörü Modellemesi
Queries	Sorgular
Secondary NameNode	İkincil Ad Düğümü
Semi-structured Data	Yarı yapılandırılmış Veri
Sentiment Analysis	Duygu Analizi
Signal Processing	Sinyal İşleme
Slots	Yuvalar
Smart Algorithms	Akıllı Algoritmalar
Social Network	Sosyal Ağ
Spatial Analysis	Mekânsal Analiz
Streaming Insert	Akış Ekleme
Structured Data	Yapılandırılmış Veri
Supervised Learning	Denetimli Öğrenme
Task Tracker daemon	Görev İzleyicisi arka plan programını

The United Nations Statistical Commission	Birleşmiş Milletler İstatistik Komisyonu
Unsupervised Learning	Denetimsiz Öğrenme
Unstructured Data	Yapılandırılmamış Veri
Value	Değer
Variety	Çeşitlilik
Veracity	Doğruluk
Velocity	Hız
Visualization	Görselleştirme
Volume	Veri Büyüklüğü
Web Data & Social Media	Web Verisi & Sosyal Medya

BİLGİSAYAR TERİMLERİ

Kısaltma	İngilizce	Türkçe
BI	Business İntelligence	İş Zekâsı
EL:DIABLO	Event/Location: Dataset In A Box, Linux-Option	Olay/Yer: Bir Kutudaki Veri Seti, Linux-Seçeneği
ETL	Extract, Transform and Load	Ayıklama, Dönüştürme ve Yükleme
GDELT	The Global Database on Events, Language and Tone	Olaylar, Dil ve Ton Küresel Veritabanı
GFS	Google File System	Google Dosya Sistemi
GKG	Global Knowledge Graph	Global Bilgi Grafiği
HDFS	Hadoop Distributed File System	Hadoop Dağıtılmış Dosya Sistemi
ICEWS	World-Wide Integrated Crisis Early Warning System	Dünya Çapındaki Bütünleşik Kriz Erken Uyarı Sistemi
Mlib	Machine Learning Library	Makine Öğrenme Kütüphanesi
NoSQL	Not Only Structured Query Language	Yapılandırılmamış Sorgu Dili
NRDB	Non-relational Database	İlişkisel Olmayan Veritabanı
ODBC	Open Database Connectivity	Açık Veritabanı Bağlantısı
Storm	Real Time Stream Processing	Gerçek Zamanlı Akış İşlemci
SQL	Structured Query Language	Yapılandırılmış Sorgu Dili
RDB	Relational Database	İlişkisel Veritabanı
RDD	Resilient Distributed Datasets	Esnek Dağıtılmış Veri Kümeleri
RFID	Radio Frequency Identification	Radyo Frekanslı Tanımlama

GİRİŞ

Endüstri Devrimi bundan 200 yıl önce, XX. yüzyılın son çeyreğine kadar, uzun soluklu olarak yaşandı. Bu dönemde insan aklı ve gelişen akıllı makinelerle fabrikalar, ulaşım, elektrik, sağlık, eğitim ve havacılık dünyada büyük çaplı gelişmelerin yaşanmasını sağladı. Bu süreçte iş ve ticaret bağları güçlendi, mesafeler kısaldı, bilgi alışverişi, sermaye hareketleri ve pazar payları büyüdü, ülkeler ve toplumlararası karşılıklı etkileşimde büyük artış yaşandı. 1990'lı yıllara gelindiğinde ise küresel bir olgu hâline gelen internet ile yeni bir devrime tanık olduk. İnternet devrimi, bilgi ve iletişim teknolojilerinin gücü, veri ağlarındaki bilgiye hızlı bir şekilde erişilebilmesini sağladı. Özellikle son 20 yılda internet iş dünyasın da ve sosyal hayatta büyük değişimler yaratmıştır. Bugün geldiğimiz noktada, Endüstri ve İnternet devrimleri bir araya gelerek teknoloji odaklı yeni bir dinamizmi ortaya çıkardı. General Electric (GE) bu yeni dönemi *"Endüstriyel İnternet Çağı"* ya da *"Endüstrinin Geleceği"* olarak adlandırmaktadır[1]. Çoğu kişi Endüstriyel İnternet Çağı'nı endüstrinin geçirdiği 3. büyük devrim olarak yorumlamaktadır. Endüstriyel İnternet, işletmeler ve büyük ölçekli endüstriler için birçok avantaj sağladı. Büyük Veri daha şimdiden binlerce kişiye yeni iş imkânı yarattı ve milyarlarca dolar tasarruf sağladı. Fütüristik düşünürler kendilerine *"robotlar insanları ele geçirecek mi?*" sorusunu sorarak, akıllanan makinelerin yeni özelliklere sahip bir iş gücüne ihtiyaç duyduğunu fark ettiler. Özellikle *"akıllı makinelerin küresel çapta bir ağa bağlanmaları, dünya çapında trilyonlarca sensörden gelen verinin internette buluşmasıyla ortaya Büyük Veri denilen bir okyanusu çıkardı"*[2]. Fakat bu kadar büyük verinin niceliği başlı başına bir problem oluşturmakta ve bu veriler arasında ilişki kurmak ve anlamlı sonuçlar elde etmek için çok fazla zaman ayırmak gerekiyor. Son beş yıldır birçok büyük şirket veri analizine olan insan ve varlık yatırımını neredeyse iki katına çıkardı. Hiç şüphesiz Büyük Veri'yi anlamlı ve işe yarar bilgiye dönüştürmek için yenilikçi yazılımlar kadar, yenilikçi bir personel gücüne de ihtiyaç vardı. Bugün artık Büyük Veri'nin analiz edilmesi, yorumlanması ve görselleştirilmesi için gelecekte yeni iş alanlarının giderek daha da büyüyeceği tahmin ediliyor. Yapılan bir

[1] ÖZSOY Canan M., *"Endüstrinin Geleceği ve Endüstriyel İnternet Devrimi"*, 25 Kaasım 2014, http://geturkiyeblog.com/endustrinin-gelecegi-ve-endustriyel-internet-devrimi, (19.09.2016).
[2] ÖZSOY Canan M., *"Endüstrinin Geleceği ve Endüstriyel İnternet Devrimi"*, 25 Kaasım 2014, http://geturkiyeblog.com/endustrinin-gelecegi-ve-endustriyel-internet-devrimi, (19.09.2016).

araştırmaya göre 2020 yılında internete bağlı nesnelerin 50 milyarı bulması bekleniyor. Bu nesneler her ne kadar akıllı da olsalar yine de kendilerinden daha akıllı olan insanlara ihtiyaç duyacaklardır[3].

Büyük Veri hâkimiyeti, toplumun yapısını anlama ve düzenleme biçimimizi değiştirirken bilgiyi analiz etme şeklimizi de değiştiriyor. Bugün artık çok büyük ve karmaşık veri setlerini analiz etmek mümkün hâle geldi. Özellikle 19. yüzyıldan itibaren büyük ve karmaşık verilerle karşılaşıldığında örnekleme yöntemine başvurulmaktaydı. Ancak örnekleme yöntemi bilginin az olduğu ve yüksek kapasitedeki dijital teknolojilerin yaygınlaşmadığı bir dönemde, genellikle doğal karşılanıyordu. Bugün artık teknolojideki gelişmeler bize verilerin tamamını (anakütleyi) kullanma imkânı sağlayarak daha önce sınırlı miktarda veri ile göremeyeceğimiz birçok ayrıntıyı görmemize olanak sağladı. Bu nedenle Büyük Veri, örneklemin erişemediği alt kategorilerin ve altyapıların çok daha net bir görüntüsünü görmemizi sağladı[4].

Büyük Veri işletmelerden tüketicilere ve bilimden hükûmete kadar hayatımızın bütün yönleriyle ilgili, devrim niteliğindeki bir konudur[5]. *"Bilgi ekonomisinin petrolü olarak tanımlanan Büyük Veri'nin özellikle beşeri bilimlerde büyük bir dönüşüm başlattığı görüşü, son 10 yılın sıkça tekrarlanan bir söylemi hâline gelmiştir"*[6]. Hükûmetler ve sosyal bilimciler daha önce elde edemedikleri birçok bilgiye bugün Büyük Veri sayesinde ulaşmaya başladı. Eğer analiz yapmak için doğru yöntem kullanılırsa Büyük Veri, dünyayı algılama ve değiştirme şeklimizi değiştirip sorunların çözülmesinde büyük kolaylıklar sağlayacaktır. Bugün başta Amerika ve Çin olmak üzere, birçok ülke en iyi uygulamaların geliştirilip belirlenmesi için Büyük Veri'yi kullanmaya başladı. Üniversitelerin araştırma programlarında analitik yöntemler üretilirken, bazı kuruluşlar da bunları teknolojiye adapte ederek hükûmet programlarında hayata geçirmiştir. Bunun en bariz örneği 2010 yılındaki Amerika-Afganistan savaşında görülmektedir. O dönemde henüz 28 yaşında olan Chris White Harvard doktorası sonrası okulunda büyük veri,

[3] GE TÜRKİYE BLOG, *"Endüstriyel İnternet, Büyük Veri ve Operasyon Optimizasyonu"*, 24 Kasım 2015, https://geturkiyeblog.com/endustriyel-internet-buyuk-veri-operasyon-optimizasyonu/, (19.09.2016).
[4] MAYER Victor S. – Kenneth CUKIER, *"Big Data Arevolution That will Transform How We Live, Work, and Think"*, 2013, p.20.
[5] JAGADISH Hosagrahar V. – Johannes GEHRKE– Alexandros LABRINIDIS – Yannis PAPAKONSTANTINOU – Jignesh M. PATEL – Raghu RAMAKRISHNAN – Cyrus SHAHABI, *''Communications of the ACM''*, 2014, 57 (7): 86-94.
[6] IŞIKLI Şevki, *''Büyük Veri, Epistemoloji ve Etik Tartışmalar''*, Online Academic Journal of Information Technology, Fall – Vol: 5 --/Num:17, 2014.

istatistik ve makine öğrenimi üzerine çalışmaktaydı. White akademik dünyada iyi bir pozisyon kazanmış, ileride profesör olmayı ve kendi alanında çalışmalar yapmayı planlıyordu. Birgün danışmanı ona DARPA'da (İleri Savunma Araştırma Projeleri Ajansı) bir konferansa katılmasını önerdi. DARPA Pentagon'un bilim ve yenilik departmanıydı. Bu departman da akıllı insanlar, büyük fikirler ve hükûmetin geniş bütçesi bir araya getirilirdi. DARPA'nın hedefi ise ülkenin teknolojik açıdan geri kalmasını önleyerek, ülkeye stratejik avantaj sağlayacak, dünyayı değiştirecek bir teknolojiyi piyasaya sürmekti. White Afganistan da devam eden savaşla ilgili bilgi alarak, Afganistan'daki karanlık güçleri öğrendi. Bu karanlık güçlerin eylemleri acımasız, fakat taktikleri ve bürokrasileri çok sofistikeydi. Bunlar öldürüyor, terör saçıyor, büyüyor ve kazanıyorlardı. ABD, bunlara karşı koymak için Büyük Veri'den yararlanma fırsatı olduğunu duymuş ve bu avantajı en kısa zamanda kullanmak istiyordu. White aslında savaşa dair hiç bir şey bilmediğini düşünüyordu. Çünkü detaylara pratik ve operasyonel bir bakış açısından bakmayı hiç denememişti. White'tan görünürde birbiriyle ilişkisiz olan, işlenmemiş (ham) devasa veri dağlarını (yığınlarını) anlamlandırması, buzdağına benzeyen istihbarat bilgi yığınlarından planlar ve politikalar geliştirmesi isteniyordu. White, Nexus 7 olarak adlandırılan gizli bir DARPA programı çerçevesinde görevlendirilmiş bir ekibin üyesiydi. ABD ordusunun Afganistan'daki veri kaynakları; CIA, Ulusal Güvenlik Ajansı (NSA), GPS uyduları, cep telefonu kayıtları, cepheden gelen bilgiler, dijital finansal kayıtlar, güvenlik kameraları ve sosyal ağlardı. O sıralar ABD istihbaratının Afganistan'daki şefi olan Tümgeneral Micheal Flynn'in söylemiyle bu bilgiler *"devasa ve kıymeti bilinmeyen bir bilgi bütününü"* oluşturuyordu. DARPA, White ve onun gibi bir grup araştırmacıyı bu verilerden anlamlı sonuçlar çıkarmaları için Afganistan'a yolladı. Grubun bazısı uydu verilerini ve karasal gözlem verilerini birleştirerek trafik akışını (veya akmadığını ki bu da Taliban'nın ele geçirdiği bir kontrol noktasını veya yola döşenmiş bir bombayı gösteriyor) belirleyecekti. White ve ekibi Taliban ve El Kaide'nin finansının dijital izlerini sürmek, kölelik, seks, silah ve uyuşturucu ticaretinin kaynağını saptamak ve bu işlemlerin nerede ve kimler tarafından gerçekleştirildiğini belirlemekle görevlendirilmişti. White'ın Afganistan'daki görev süresi dolduğunda Nexus 7 komutanlarının da saygısını kazanmıştı. White, Nexus 7 çabalarından dolayı Savunma ve Hazine Bakanlığından büyük övgü alarak madalya ve takdir belgesi aldı. Ayrıca White ve ekibi *"kilit stratejik ve operasyonel sorulara*

benzersiz ve kıymetli ışık tutan büyük bir analitik veri çerçevesi" oluşturduğu için övgü aldı. White, Afganistan dan döndükten sonra da çalışmalarına devam etti ve savaşı, savaş sınırlarının ötesine taşımaya karar verdi. White'ın izini sürdüğü veriler çocukların ve kadınların mal gibi el değiştirdiği yerleri hedef olarak gösteriyordu. Bu suçlar sadece Afganistan'a özgü değildi ve bu suçların işlendiği yere gitmek için uçak değil sadece modem yeterliydi[7].

Bugün bildiğimiz internet aslında internetin tamamı değil sadece buz dağının görünen kısmıdır. Araştırılan, okunan ve öğrenmek için kullanılan "yüzeysel" web'in ya da açık web'in, toplam internetin %5 ila 20'sine denk geldiği tahmin ediliyor. Google'ın Firefox'un ve Siri'nin olduğu Gmail hesaplarımızın ve yer imlerinin bulunduğu, pizza siparişi verdiğimiz ve günlük haberleri takip ettiğimiz internetin 200 terabayt veriye, yani ABD Kongre Kütüphanesi'nin tamamının dijitalleştirilmiş halinden fazlasına denk geldiği tahmin ediliyor. Geri kalanı ise internetin gizemli arka sokağı olarak adlandırılan Deep Web (Derin Web) ya da Dark Web (Karanlık Web) olarak adlandırılmaktadır. Dark Web, sürüngen adı verilen araçların ulaşamadığı, özel ağlarda bulunan web sitelerinin bulunduğu alandır. Burada tutulan verilerin bazısı sosyal medya veya forum gibi parola korumalı dinamik web sitelerinde tutulan "derin" web olup geri kalan ise "karanlık" web'tir. Karanlık Web, hükûmet kullanıcıları, siber özgürlükçüler, fuhuş ve eroin ticareti, IŞİD planlamacıları, hükûmet sırlarını açığa çıkaran hacktivistler ya da Arap Baharı planlamacıları arasında hiçbir fark gözetmemektedir. Dolayısıyla Karanlık Web örgütsel suçlar için çok mükemmel bir ortamdır. White'a göre aslında internetteki veriler (resimler, ses kayıtları, e-postalar, bloglar, telefon kayıtları, GPS sinyalleri, sosyal medya gönderileri) gerçek dünyada olanların tanımlarından oluşmaktadır. White bunu *"bir araştırmacının amacı gerçek dünyadaki aynı olayı bulmak için bu tanımlar içerisinden kazı yapmak ve tersine doğru çalışmak"* olduğunu söylüyor. White'ın Memex adını verdiği arama motoru projesi sayesinde bugün için bütün bunları yapmak mümkün hâle geldi. White Memex araçları sayesinde bir dedektifin iki haftada yapacağı araştırmayı çok kısa sürede yapıyor ve bu sayede seks suçları trafiğini saptayabiliyor. Bugün Amerika da Manhattan Bölge Savcılığı insan ticareti ile ilgili tüm soruşturmalarda Memex'i kullanıyor ve sadece 2016'nın ilk altı ayında 4.720 vaka savcılık tarafından incelemiştir.

[7] Popular Science Türkiye, *''Karanlık Web'i Aydınlatan Adam''*, 03.11.2016, http://www.pressreader.com/turkey/popular-scence-turkey, (20.11.2016).

Ayrıca Memex araçları sayesinde bugün IŞİD'in propagandası ve üye alım hareketlerini, naylon şirketlere para aklama arasındaki bağlantıları, yasa dışı iş gücü ya da silah kaçakçılığını, sosyal medyadaki sözcük ve fikirlerin haritalarını, bunların amacını harita üzerinde canlı olarak görmek de mümkün hâle geldi. Bugün artık hükûmetler Büyük Veri'yi kamu düzeni içerisinde etkin bir şekilde kullanıyor. Yine de bilim adamları arasında bazı görüş farklılıkları bulunmaktadır. Bazı bilim adamları kamu politikalarının kanıta dayalı olduğunu söylerken bazısı da kötü kamu politikalarından dolayı bireylerin ihtiyaç duyduğu hizmeti alamadıklarını söyleyerek yeni yöntemlerin geliştirilmesini istemektedir. Bugün bizlerden toplanan verilerin miktarı ve içerik mahremiyeti açısından da büyük bir tehlike yaratıyor çünkü bu verilerin yanlış ellerde kullanılma ihtimali bulunmaktadır. Bilim insanları bundan korunma yollarını uygulamalara geçirmek için seferber olmaktadır. Hükûmetler de bu veriye kimin ulaşabileceği ya da gerektiği gibi kullanıldığı şekilde halkı bilgilendirmeli ve güvenilir oldukları konusunda insanları ikna etmelidir[8].

Büyük Veri uygulamaları bugün birçok alanda kullanılarak büyük başarılara imza atılmasını sağlamıştır. ABD Ulusal Okyanus Ve Atmosfer Teşkilatı (NOAA), ABD Ulusal Havacılık Ve Uzay Dairesi (NASA), sağlık sektöründeki birçok ilaç firması ve büyük ölçekli enerji firmaları bugün büyük miktardaki veriyi toplayarak bu veriden günlük olarak bilgi elde etmektedir. NOAA toplanan bu veriyi; iklim, ekosistem, hava durumu tahmini ve ticari amaçlı olarak kullanırken NASA daha çok havacılık araştırmaları için kullanmaktadır. İlaç şirketleri ve enerji şirketleri ise büyük veriyi ilaç deneyleri ve jeofizik analizleri yapmak için kullanmaktadır. Bugün New York Times büyük veri analizinde yeni nesil araçları kullanarak metin analizi ve web madenciliği yapmaktadır. The Walt Disney Company ise müşteri verilerini kullanarak müşteri alışveriş eğilimini tahmin etmektedir[9]. Yine MIT ve Birmingham Üniversitesi'ndeki araştırmacıların trafik verilerini kullanarak beş şehirde yaptığı bir çalışmaya göre az sayıda sürücünün bazı alternatif yolları tercih etmesiyle trafikteki yoğunluğun %30 oranında azaltılabileceğini bulmuşlardır. Son olarak yapılan başka bir araştırmaya göre, bir insan İstanbul'da yılda yaklaşık olarak 125 saat trafikte durarak zaman kaybediyor.

[8] Popular Science Türkiye, *''Karanlık Web'i Aydınlatan Adam''*, 03.11.2016, http://www.popsci.com/man-who-lit-dark-web 20.12.2016.
[9] OHLHORST Frank, *''Big data analytics: turning big data into big Money''*, New Jersey 2013, s.19-21

Bu da günde ortalama 8 saat çalışan bir kişinin ortalama yılda 15 günlük mesaisini yolda geçirmesi anlamındadır[10].

Bugün medya da sürekli Büyük Veri'nin iş, siyaset ve ekonomi alanlarında büyük bir dönüşüm sağlayacağı yönünde haberler yapılmaktadır. "Veri Bilimci" terimi birkaç yıl önce duyulmaya başlandı. Barack Hussein Obama, 2008'deki seçim kampanyasında internet ve sosyal medyayı etkin kullanmıştır. 2012'nin sonbaharın da ABD'de yapılan başkanlık seçimlerinde gazeteciler; Obama'nın seçim kampanyasında Veri Bilimci kullanmasının ona büyük avantaj sağladığını savunuyorlar. Obama kampanyasını gençler üzerinden kurgulayarak, gençlerle sosyal ağlar üzerinden iletişime geçmiştir. Obama sosyal medya gücünün farkındaydı ve bu gücü kullanarak ve örgütleyerek büyük bir sosyal medya başarısına imza atmıştır[11]. Nate Silver, bütün kamuoyu anketlerini alıp bunlara geçmişteki seçim sonuçlarını etkileyen faktörleri de modeline ekledikten sonra sanal ortamda simülasyonlar yaparak çok sayıda seçim gerçekleştirdi ve elde ettiği sonuçların olasılık dağılımından yola çıkarak bir tahminde bulundu. Tahmin sonuçları Obama'nın %80 olasılıkla seçimi kazanacağı yönündeydi[12]. Nitekim seçim sonuçları Nate Silver'ın tahminin doğruladı ve seçimden hemen önce çıkan "Signal and Noise" adlı kitabının satışlarında büyük bir artış yaşandı.

Bu tez çalışması dört bölümden oluşmaktadır. Büyük Veri konusu çok geniş ve disiplinlerarası bir konu olup bu tezde konuların ele alınması genellikle istatistik ağırlıklı verilmiştir.

Birinci bölümde, Büyük Veri'ye giriş yapılarak Büyük Veri'nin ortaya çıkmasıyla birlikte istatistikte görülen değişim hakkında bilgi verilerek veri biliminin istatistikle birlikte tarihi gelişim süreci ve istatistiğe olan etkilerinden kısaca bahsedilerek, Büyük Veri çağı ile istatistikte risk altına giren dört temel gerçekten bahsedilmiştir. Bununla birlikte "Veri Bilimi Nedir?" ve "Veri Bilimci Kimdir?" sorularına cevap aranmış ve daha sonra Büyük Veri'nin özellikleri olan 5V kavramlarına açıklık getirerek, Büyük Veri ile ilgili yapılan bazı tanımlar verilmiştir.

---

[10] GE TÜRKİYE BLOG, *"Büyük Veri Trafiği Nasıl Alt Eder?"*, 13 Mayıs 2016, https://geturkiyeblog.com/buyuk-veri-trafigi-nasil-alt-eder, (29.07.2016).

[11] SZKOLAR Dorotea, *"Data Mining in Obama's 2012 Victory"*, 24 January 2013, http://infospace.ischool.syr.edu/2013/01/24/data-mining-in-obamas-2012-victory, (29.07.2016).

[12] TÜFENKÇİ Zeynep, *'' In Defense of Nate Silver, Election Pollsters, and Statistical Predictions''*, 11.02.2012, https://www.wired.com/2012/11/why-predictions-and-statistical-models-are-necessary-and-good-for-democracy, (23/09/2016).

İkinci bölümde, büyük veri kümelerinin analizinde yoğun olarak kullanılan bazı Büyük Veri Teknikleri ile büyük veri kümelerinin yönetilmesinde ve analizinde kullanılan Büyük Veri Teknolojileri hakkında genel bilgiler verilmiştir.

Üçüncü bölümde, kuvvet yasası dağılımı hakkında genel bilgiler verilerek bu dağılımın görüldüğü olgulardan bahsedilmiş ve ampirik bir verinin kuvvet yasası dağılımına uygun bir dağılım gösterip göstermediğini test etmek için yapılması gereken analizler hakkında bilgiler verilmiştir.

Dördüncü bölümde, dünyanın en büyük açık veritabanı olan GDELT Projesi hakkında bilgi verilmiş ve bu veritabanındaki veriler kullanılarak dünyadaki çatışmalar ile Türkiye ve Ukrayna'daki protestolar incelenerek bunların kuvvet yasası dağılımına uygun bir dağılım sergileyip sergilemediği test edilmiştir.

BİRİNCİ BÖLÜM
BÜYÜK VERİ VE İSTATİSTİK

Resmî istatistik kurumları, uluslararası kabul gören ilkeler doğrultusunda örgütlenmiştir. 1947'de kurulan Birleşmiş Milletler İstatistik Komisyonu (The United Nations Statistical Commission), küresel istatistik sisteminin en üst kademesidir ve tüm dünyadaki üye devletlerden Baş İstatistikçileri (Chief Statisticians) bir araya getirir. Komisyon, coğrafi dağılım temelinde seçilen 24 üye ülkeden oluşur: Afrika'dan beş, Asya'dan dört, Doğu Avrupa'dan dört, Latin Amerika'dan dört ve Batı Avrupa ve diğer ülkelerden yedi üye den oluşmaktadır.

1994'te Birleşmiş Milletler İstatistik Komisyonu resmî istatistiklerin temel ilkelerini benimsedi. Bu ilkeler şunlardır:

1) Uygunluk, tarafsızlık ve eşit erişim
2) Mesleki standartlar ve etik
3) Hesap verebilirlik ve şeffaflık
4) Yanlış kullanımın önlenmesi
5) Resmî istatistik kaynakları
6) Gizlilik
7) Mevzuat
8) Ulusal koordinasyon
9) Uluslararası standartların kullanımı
10) Uluslararası işbirliği

Burada beşinci ilkeye odaklanmak istiyoruz: "Resmi istatistik kaynakları". Bu ilkeye göre; istatistiksel amaçlar için kullanılan veriler, istatistiksel araştırmalar ya da idari kayıtlar, her türlü kaynaktan alınabilir. İstatistikçiler veri kaynağını; kaliteye, zamanlamaya, maliyetlere ve katılımcıların yüküne göre seçerler. Geçmişte istatistiksel kuruluşlar, veriyi anket formları aracılığıyla ankete katılanlar tarafından anket doldurularak toplardı. İstatistik anketleri görüşmeciler tarafından kâğıda daha sonra, görüşmeciler tarafından telefonla görüşme, ardından bilgisayarlar kullanılarak daha sonra da optik okuyucular kullanılmaya başlandı ve son zamanlarda Web'deki anketler kullanılmaya başlandı. Bu araştırmacılar için "istatistiksel yük" olarak adlandırılan zaman ve maliyet açısından büyük sorun teşkil etmekteydi.

90'lı yıllarda, istatistik kurumları, istatistiki üretim sürecinin verimliliğini artırmak için maliyet ve personel kaynakları açısından tasarruf sağlamak için birtakım baskılara maruz kaldı. Aynı zamanda, ankete katılanların istatistiksel araştırmalara yüklediği yükü azaltmak için artan siyasi talepler vardı. Bu baskılar göz önüne alındığında, istatistikçiler giderek veri toplama yöntemi olarak geleneksel araştırma yaklaşımına alternatifleri dikkate almak zorunda kaldılar. Çözüm açıktı: istatistiksel olmayan birçok kuruluş verileri çeşitli biçimlerde toplardı ve bu veriler nadiren istatistiksel anketler yoluyla toplananlar için doğrudan yer değiştirenler olmasına rağmen, bazen birden fazla kaynağın kombinasyonu yoluyla, tamamen veya kısmen değiştirme imkânı sunardı. Doğrudan istatistiksel veri toplama kaynakları "İdari Kaynaklar" olarak adlandırılır ve tanımlanır. "İdari kaynak, birimlerin ve işlemlerin ilgili kayıtlarının istatistiksel veri kaynağı olarak görüldüğü bir idari düzenlemeyi uygulamakla sorumlu bir organizasyon birimidir." 90'lı yıllardan başlayarak, idari kaynaklar pek çok istatistiksel sürecin merkezi hâline geldi ve tüm gelişmiş ülkeler, istatistiksel araştırmalara ayrılan maliyetleri ve kaynakları azaltmak için bunları kullanmaya başladılar[13].

Şimdi de benzer bir kaygıyla karşı karşıyayız: istatistik kuruluşları her zaman hükûmetler tarafından ve kamuoyu tarafından maliyetlerin düşürülmesi ve istatistiksel yüklerin azaltılması için baskı altına alınmaktadır. Ancak bugün yeni bir veri kaynağı olarak internet de kullanılmaktadır. İnternette giderek daha fazla veri üretiliyor. Bu veriler çevremizdeki çok sayıda elektronik cihaz tarafından desteklenen sensörler tarafından üretiliyor. Üretilen bu verilerin miktarı, hızı ve çeşitliliği, "Büyük Veri" kavramının ortaya çıkmasına yol açmıştır.

Bugün birçok istatistik kuruluşu, resmî istatistikleri tamamlamak ve desteklemek için Büyük Veri'yi bir kaynak olarak kullanma olasılığını araştırmaya başlamıştır.

Resmî istatistiklerde Büyük Verilerin kullanımı aşağıdaki gibi birçok zorluğu beraberinde getirir:

1) Yasama yetkisi, veri erişimine ve kullanımına ilişkindir.

2) Gizlilik, yani kamu güveninin yönetimi ve verilerin tekrar kullanılması kabulü ve diğer kaynaklara bağlantısı.

[13] WALLGREN Anders – Britt WALLGREN, *"Register-based Statistics: Administrative Data for Statistical Purposes"*, Wiley Series in Survey Methodology, John Wiley & Sons, New York, 2007, ISBN: 978-0-470-02778-3-0.

3) Maliyet, diğer bir deyişle, verilerin sağlanmasıyla elde edilen verilerin potansiyel maliyetler.

4) Yönetim, verilerin yönetimi ve korunması ile ilgili politikalar ve yönergeler.

5) Metodolojik, yani veri kalitesi ve istatistiksel yöntemlerin uygunluğu.

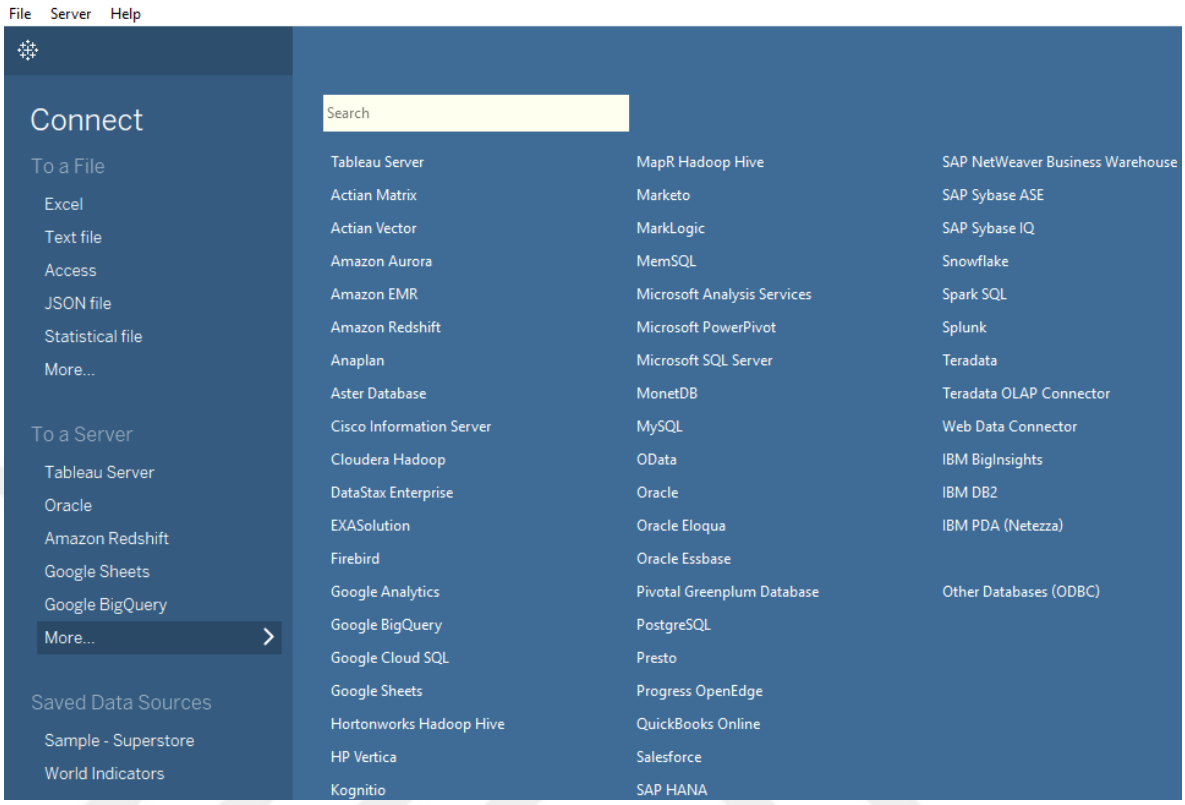
6) Teknolojik, bilgi teknolojisi ile ilgili konular.

İnsanlık tarihi boyunca sürekli yeni teknolojik gelişmeler yaşanmıştır. Özellikle son 100 yılda teknolojide yaşanan gelişmeler diğer zamanlara göre daha çok ve daha hızlı olmuştur[14]. Teknolojide yaşanan bu gelişmeler bir devrim niteliğinde olup, bu bilginin her geçen gün daha da hızlı artmasına neden olmuştur. Bilgi miktarındaki bu hızlı artışın temel nedeni ise sosyal ağ etkileşiminin giderek büyümesi, lokasyona duyarlı cihazların artması ve akıllı sensörlerin sayısındaki artıştan kaynaklanmaktadır. Bugün sosyal paylaşım sitelerinden sadece bir tanesi günlük 10 terabayt veri üretebilmekte, bir jet uçağı her 30 dakikada 10 terabyt'lık veriyi toplayabilmekte, bazı kurum ve kuruluşlar ise her saat 10'larca terabayt veri saklayabilmektedir. "Örneğin, Google, Facebook, Linkedln, Amazon, Microsoft gibi şirketler petabyte düzeyindeki veriyi bir günde maniple edip paraya dönüştürebilmektedirler. Bu şirketlerin arasında Uber ve Airbnb gibi iki şirketi daha dahil etmek gerekir. Bunlardan ikisi de küresel anlamda taksi ve otel/ev kiralama hizmetlerinden para kazanan iki platformdur. Bu iki şirketin de ne taksileri ne de otelleri var ama ellerindeki veriden inanılmaz paralar kazanıyorlar"[15].

Bilgisayar ve bulut teknolojileri gibi yeni teknolojiler sayesinde veri toplama ve saklama her geçen gün daha da kolaylaşmaktadır. Hiç kuşkusuz toplanan ve saklanan bu veriler istatistiği de büyük ölçüde etkiledi ve gelecekte de etkilemeye devam edecektir. Eskiden istatistik yöntemlerinde kullanılan paket programları veri tabanlarına bağlanarak veri çekme özelliğine sahip değildi. Oysaki bugün, Tableau gibi gelişmiş yeni nesil programlar sayesinde çok sayıda veritabanına bağlanmak mümkün hale geldi. Şekil 1.1'de görüldüğü gibi yeni nesil istatistik paket programları sayesinde veri tabanlarının çoğuna bağlanmak mümkün hale geldi. Ayrıca, programın sağladığı imkânlar yetersiz kalıyorsa, bu takdirde ODBC (Open Database Connectivity-Açık Veritabanı Bağlantısı) ile veri tabanlarına bağlanmak mümkündür. Yine, Tableau programının server versiyonu

[14] BilgiUstam, *''Bilim ve Teknoloji Tarihi – Son Yüzyıldaki Gelişmeler''*, 2017, http://www.bilgiustam.com/bilim-ve-teknoloji-tarihi-son-yuzyildaki-gelismeler, (10.01.2017).
[15] GÜRSKAL Necmi, *''R İle Betimsel İstatistik''*, 2. Baskı, Bursa, Dora Yayınevi, 2016, s. 3.

"Tableau Server" sayesinde büyük veri tabanlarına bağlanarak veri analizleri yapmakta çok kolay bir hale gelmiştir.



Şekil 1.1: Tableau Programının Veri Bağlantısı Sağlayan Özellikleri

Bugün hala istatistikte yoğun bir şekilde kullanılan SPSS, SAS, STATISTICA ve MINITAB gibi istatistik programlarının internete bağlanıp veri tabanlarından veri çekme özellikleri sınırlıdır. Bu programlarda analiz yapmak için programa öncelikle veri girişinin yapılması gerekir. Ancak analizi yapılacak verinin hacmi ve çeşitliliği arttıkça verinin programa girilmesi büyük bir problem teşkil etmektedir. Tableau (Bkz. Şekil 1.1.) ve NodeXL gibi programlar sayesinde bugün bu problem ortadan kalkmıştır. Otomatik olarak çok basit bir şekilde veri çeken bu programlar sayesinde para, işgücü ve zamandan büyük tasarruf sağlanmaktadır.

1.1. Veri Bilimi Tarihi

İnsanoğlu var olduğu günden bu yana sürekli bir değişimin ve dinamizmin içinde olmuştur. İlkel toplumda insanlar doğanın kendilerine sunduklarıyla yetinirken, tarım toplumunda ekip-biçerek üretim yapmışlardır. Tarım toplumunda insanlar için en değerli madde toprak iken, sanayi toplumunda toprağın yerini makinalar almıştır. Özellikle

sanayi devrimi ile birlikte bilişim teknolojilerinde hızlı bir gelişim yaşanmış ve bu hızlı gelişim bilgi toplumunun ortaya çıkmasına sebep olmuştur. Bilgi toplumuna geçiş ile birlikte sanayi toplumunda üretilen mallar yerine bilgi toplumunda bilgi üretilmekte ve üretim makinalar yerine iletişim ağları ile yapılmaktadır.

1960'li yıllardan sonra bilgi teknolojilerinde hızlı bir gelişme sağlanmıştır. Bu gelişmeler iletilebilen ve erişilebilen bilginin miktarında çok büyük bir artış yaşanmasına sebep olmuştur. Bütün bu gelişmeler istatistik çalışmalarını da büyük ölçüde etkiledi. John W. Tukey 1962'de yazdığı "*Veri Analizinin Geleceği*" yazısında şunları söylüyor; "Uzun bir süre özelden genele doğru çıkarımlarla ilgilenen bir istatistikçi olduğumu düşündüm. Fakat Matematiksel İstatistik'teki gelişmeleri merak ve şüphe ile izledim. Veri analizinin ilgi alanımın merkezine yerleştiğini fark ettim. Veri analizi ve istatistik mutlaka birbiriyle ilişkili olmalıdır. Ele alınan özellikler matematikten ziyade bilimin özellikleridir. Veri analizi özünde deneye dayalı bir bilim dalıdır. Veri analizi hayati bir öneme sahiptir. Elektronik Bilgisayar artışı saklanan veriyi arttırmıştır. Diğer taraftan hiç şüphesiz veri analizinde bilgisayarın büyük öneme sahip olduğunu söylemek mümkündür."[16]

1974'te Peter Naur, İsveç ve Amerika Birleşik Devletleri'nde "*Concise Survey of Computer Methods"* adlı kitabını yayınlandı. Peter Naur'un kitabı, çağdaş veri işleme yöntemlerinin geniş çapta uygulamaları ile birlikte kullanıldığı bir araştırma niteliğindedir[17]. 1977 de Uluslararası İstatistik Hesaplama Kuruluşu (IASC), Uluslararası İstatistik Enstitüsünün (ISI) bir bölümü olarak kuruldu. IASC, bu alanda uzmanlaşmış kişiler sayesinde modern bilgisayar teknolojisi ile istatistiği birleştirerek enformasyon ve bilgiyi veriye dönüştürmeyi amaçlamıştır[18].

1996'da Japonya'nın Kobe şehrin de yapılan Uluslararası Sınıflandırma Derneği Federasyonu (IFCS) konferansında "veri bilimi" terimi ilk kez bu konferansın başlığında yer aldı ("Veri Bilimi, sınıflandırma ve metodlar"). Kasım 1997' de Professor C. F. Jeff Wu Michigan Üniversitesin'deki "*H. C. Carver Chair in Statistics*" dersinin açılış

[16] TUKEY John W., *"The Future of Data Analysis"*, The Annals of Mathematical Statistics, Vol. 33, No. 1, Mar. 1962, pp. 1-67.
[17] NAUR Peter, *"Concise Survey of Computer Methods"*, 397 p., Student litteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974 ISBN/Petrocelli 0-88405-314-8, 1975.
[18] PRESS Gil, *"A Very Short History Of Data Science"*, 9 May 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#14fa, (10.10.2016).

konuşmasında istatistiği, veri bilimi ve istatistikçiyi, veri bilimcisi olarak yeniden tanımlamıştır[19].

"Veri bilimi" deyimini istatistikçi William Cleveland ilk kez 2001 yılında, "*Veri Bilimi: İstatistik Alanının Teknik Alanlarını Genişletmek İçin Bir Eylem Planı*" kullandı. William Cleveland'in makalesinde şunlar yazılıyordu: "Bu çalışma, istatistiğin temel alanlarını genişletmek amacıyla bir plan önermektedir. Bu plan temel bir değişimi belirtmekte ve değişen alan "veri bilimi" olarak adlandırılacaktır."[20]

XXI. yüzyıla gelindiğinde bilginin önemi yeni bir boyut kazanmıştır. Bilişim çağı olarak adlandırılan bu çağda PC, cep telefonu ve internetin yaygınlaşması ile birlikte insanlar işlerinin çoğunu çevrimiçi yapar hâle geldi; bu da ortaya çıkan bilgi miktarında çok hızlı ve büyük bir artışa sebep oldu. 2007 yılında Çin'in Şanghay şehrindeki Fundan Üniversitesi'nde Dataloji ve Veri Bilimi için Araştırma Merkezi (Research Center for Dataology and Data Science) kurulmuştur. 2009'da bu araştırma merkezindeki Yangyong Zhu ve Yun Xiong adlı iki araştırmacı "Dataloji ve Veri Bilimine Giriş" adlı çalışmalarını yayınladılar. Yangyong Zhu ve Yun Xiong çalışmaların da "Dataloji ve Veri Biliminin doğa bilimlerinden ve sosyal bilimlerden farklı olarak sanal alemde üretilen verileri ele aldığını" söylemişlerdir [21]. Bugün yaşadığımız dijital çağda, internette her gün devasa büyüklükte yapısal ve yapısal olmayan veri yığınları üretilmektedir. Bu derecede büyük verinin artmasıyla birlikte bu verilerin analizinin yapılması için gerekli olan iş gücü ihtiyacı da artmıştır[22]. Bu bağlantılılık olgusu ve verinin büyüklüğünde, hızında ve yapısındaki değişim istatistiğin karşısına veri bilimi olarak adlandırılan yeni bir alanın ortaya çıkmasını sağladı.

1.2. Veri Bilimi Nedir?

Veri Bilimi; çok büyük miktardaki bilginin, toplanması, hazırlanması, analiz edilmesi, görselleştirilmesi, yönetilmesi ve sunulması ile ilgilenen bir alandır. Veri biliminde temel amaç veriden anlamlı bilgilerin çıkarılmasıdır. Bu amaçla değişik

---

[19] PRESS Gil, *"A Very Short History Of Data Science"*, 9 May 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#14fa, (10.10.2016).

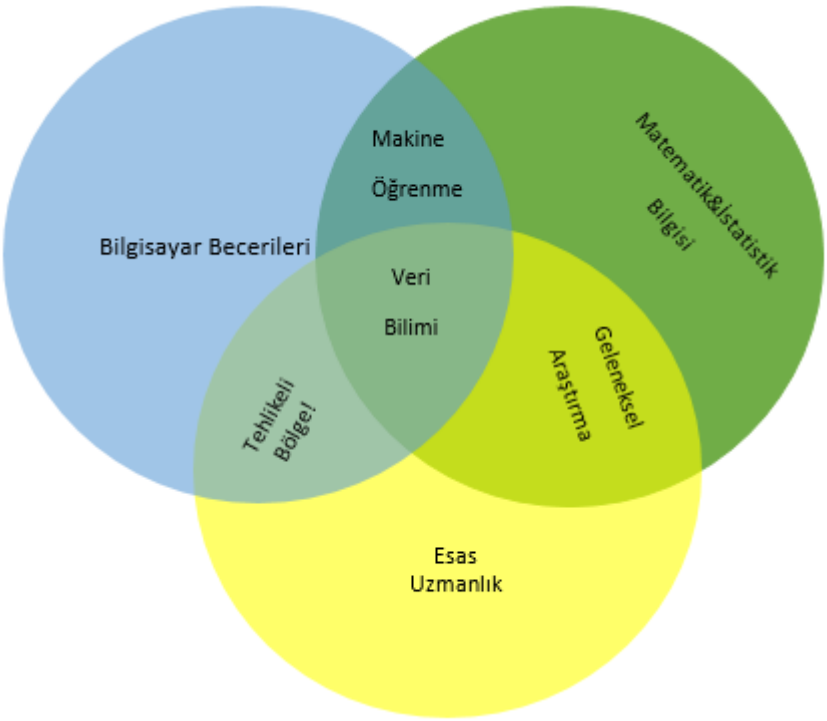
[20] CLEVELAND William *S., "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics"*, Bell Labs, 2001, http://www.stat.purdue.edu/~wsc/papers/datascience, (10.09.2016).

[21] PRESS Gil, *"A Very Short History Of Data Science"*, 9 May 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#14a1, (10.10.2016).

[22] PRESS Gil, *"A Very Short History Of Data Science"*, 9 May 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#14a1, (10.10.2016).

bilimsel alanlardan birçok teori ve teknik kullanılır. Veri Bilimi multi-disipliner bir alan olup, programları arasında farklılıklar olabilmektedir (Bkz. Şekil 1.2). Veri bilimi; sinyal işleme, olasılık modelleri, makine öğrenmesi, istatistiksel öğrenme, veri madenciliği, veritabanı, veri mühendisliği, model de dâhil geniş alanlardaki matematik, istatistik, kimya bilimi, bilgi bilimi ve bilgisayar bilimlerinden birçok alandaki teknik ve teorileri kullanır. Bu alanda modelleme, veri ambarı, veri sıkıştırma, bilgisayar programlama, yapay zekâ ve yüksek performanslı bilgi işleme kullanılmaktadır. Büyük bilgiyi ölçeklendirme yöntemleri, Veri Bilimi için özel bir önem taşır, ancak Veri Bilimi genellikle bu kadar büyük verilerle sınırlandırılamaz. Özellikle makine öğrenmesinin gelişimi, Veri Bilimi büyümesini ve önemini arttırmıştır[23]. 2002 yılı pek çok açıdan "Dijital Çağ'ın" başlangıcı olarak değerlendirilir. Çünkü bu dönemde dijital olarak depolanan veriler geleneksel olarak depolanmış verileri ilk kez geçmiştir. Özellikle son on yılda yalnızca veri bilimini akılda tutmak için bir takım teknolojiler geliştirilmiştir. Sonuç olarak, bir yandan eşi benzeri görülmemiş hızlarda veri üretmeye devam edilmekte ve diğer yandan, veriden bir katma değer elde etmek için Hadoop, Spark, Python, R, SQL ve Tableau gibi araçlar sürekli gelişmektedir.

[23] Ratheesh's- Tech Blog, *"Data Science"*, 2016, http://rathishnair.com/techblog/data-science-machine-learning/, (10.11.2016).



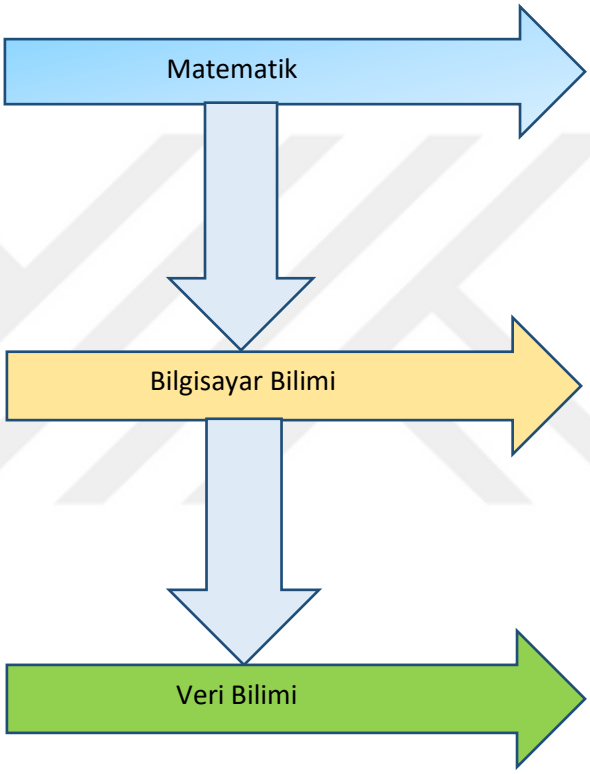
Şekil 1.2: Veri Bilimi Ven Şeması
Kaynak: CONWAY Drew, *"The Data Science Venn Diagram"*, 30 September 2010.

Veri Bilimi; tarım, pazarlama optimizasyonu, dolandırıcılık tespiti, risk yönetimi, pazarlama analitiği, kamu politikası gibi çeşitli alanlardaki sorunları araştırmak için veri hazırlama, istatistik, tahmini modelleme ve makine öğrenimi gibi yöntemleri kullanır. Veri Bilimi, makine öğrenmesi gibi genel yöntemlerin kullanımını vurgulayarak birden fazla alanda değişiklik yapmadan öğrenme gerçekleştirilir. Bu yaklaşım, alana özgü bilgi ve çözümlere vurgu yapılmasıyla geleneksel istatistiklerden farklıdır (Bunun nedeni, özelleştirilmiş çözümlerin geliştirilmesinin ölçeklenmemesidir.)[24].

Veri Bilimi; makine çevirisi, konuşma tanıma, robotik, arama motorları, dijital ekonomi, aynı zamanda biyolojik bilimler, tıp bilişimi, sağlık, sosyal bilimler ve beşeri bilimler gibi akademik ve uygulamalı araştırmaları da etkilemektedir. Bununla birlikte Veri Bilimi; ekonomiyi, iş dünyasını ve finansmanı da büyük ölçüde etkilemektedir. İş dünyası açısından, Veri Bilimi; veri madenciliği ve veri analizi gibi bir dizi etkinliği kapsar ve yeni ortaya çıkan rekabetçi zekanın ayrılmaz bir parçası olmaktadır[25].

[24] Ratheesh's- Tech Blog, *"Data Science"*, 2016, http://rathishnair.com/techblog/data-science-machine-learning/, (10.11.2016).
[25] Ratheesh's- Tech Blog, *"Data Science"*, 2016, http://rathishnair.com/techblog/data-science-machine-learning/, (10.11.2016).

Veri Bilimi'nin gerçekten yeni bir alan olduğunu fark etmek çok önemlidir. Özellikle 1980'ler de bilgisayarın yaygınlaşmasıyla birlikte matematik, bilgisayar biliminin ortaya çıkmasını sağlamıştır. Bilgisayarların ortaya çıkması da daha çok ve hızlı bilgi üretilmesini, veri tsunamisinin (fırtınası) yaşanmasını sağlamıştır. İşte bugün ortaya çıkan bu devasa büyüklükteki veri ile ilgilenen alan olan "Veri Bilimi" ortaya çıkmıştır. Veri Bilimi, büyük hacimdeki verinin hızlı ve etkin bir şekilde incelenerek anlamlandırılması ile ilgilendiği için bugün özellikle büyük miktarlarda veri üreten büyük firmalar için, Veri Bilimi gelecek adına hayati öneme sahip olmaya başlamıştır.



Şekil 1.3: Veri Biliminin Ortaya Çıkmasında Etkili Olan Bilimler
Kaynak: AALST Wil M. P. van der, *"Data Scientist: The Engineer of the Future"*, In book: Enterprise Interoperability VI, Proceeding of the I-ES Conferences 7, DIO:10.1007/978-3-319-04948-9_2, 2014, pp. 13-26.

Şekil 1.3.'te görüldüğü gibi Veri Bilimi birçok bilim dalından ortaya çıkmakla birlikte, geleneksel olarak ortaya çıkmasında matematik ve bilgisayar bilimleri daha etkin rol oynamıştır.

Veri Bilimi, aslında istatistiğe çok benzemektedir fakat, bu alan istatistik adı altında tanımlanmayıp "Veri Bilimi" olarak adlandırılmaktadır. Burada akla hemen şu soru geliyor; Veri Bilimi alanını istatistikten ve diğer bilim alanlarından ayıran nedir? Öncelikle Veri Bilimi'nde kullanılan veri heterojen bir yapıdadır. Yani yapılandırılmış,

yarı yapılandırılmış ve yapılandırılmamış türdeki veriler mevcuttur. Örneğin, sanal tıklamalar, fotoğraf, ses, video, e-posta… vb. birden fazla türdeki veriyi analiz etmek ve anlamlı hâle getirmek için bilgisayar bilimine ihtiyaç duyulur. Yine mühendislik açısından bakıldığında verinin keşfedilmesinde geleneksel veritabanları sistemlerinin yetersiz kaldığı görülmektedir. Bunun yerine daha gelişmiş veritabanları sistemlerine ve teknolojilere ihtiyaç duyulmuştur. İşte Veri Bilimi bu sebeplerden dolayı istatistikten ve diğer alanlardan farklılaşmaktadır[26].

1.3. Veri Bilimi ve İstatistik

Son yıllarda, "Büyük Veri" teriminin ortaya çıkmasıyla birlikte birçok bilimsel çalışmada "Veri Bilimi" terimi daha fazla kullanılmaya başlandı. Bilgisayar teknolojilerinde yaşanan büyük gelişmeler verinin büyüklüğünde, hızında ve yapısında büyük değişimlere yol açtı. Bu değişimler istatistiğin karşısına "Veri Bilimi" olarak adlandırılan yeni bir alanı çıkardı. "İstatistik" ve "Veri Bilimi" o kadar birbiri ile ilişkilidir ki, bunların birini diğerinden ayırmak mümkün değildir. Öyle ki, bugün "İstatistik" sözcüğünün kullanıldığı her yerde "Veri Bilimi" de geçiyor veya bunun tersi de doğrudur. Hiç kuşkusuz önümüzdeki süreçte de "İstatistik" ve "Veri Bilimi" birlikte kullanılmaya devam edecektir. Bizler istatistikçiler olarak bu sürenin çokta kısa olmayacağı, istatistiğin veri bilimi için çok önemli olduğunu düşünmekteyiz.

Veri bilimi multi-disipliner bir alan olup, istatistik bu alanlar içerisinden en önemli olanıdır. Kirk Borne'e göre etkin veri bilimci ve bazı büyük veri kullanıcıları istatistiksel çıkarımlarda kullanılan temel varsayımlara gerek duymazlar[27]. Bunun nedeni Büyük Veri'nin istatistikteki katı varsayımlara alternatif bir yol sunmasıdır. İstatistik matematikteki karmaşıklığı kullanmadan büyük veri yığınlarının keşfedilmesini sağlar. Eğer elimiz de yeterli miktarda veri olsaydı, 1000 kez doğrulama ya da milyonlarca örnekleme birimlerine sahip 1000 değişkenli modeller kurabileceğimiz için o zaman istatistiğin gereksiz olduğunu düşünebilirdik.

Borne büyük veri çağında risk altında olan dört temel istatistiksel gerçekten (açık, apaçık gerçekler) bahsetmiştir:

---

[26] AKÇAY Mustafa, *"Veri Bilimi"*, 2016, http://mustafaakca.com/veri-bilimi/, (02.10.2016).
[27] BROWN Brad – Michael CHUI – James MANYIKA, *"Are you ready for the era of 'Big Data'?"*, Article McKinsey Quarterly, October 2011: 24-35, http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data, (10.12.2016).

1. **Korelasyon nedensellik anlamına gelmez:** Aslında genelde herkes bunu biliyor; fakat çoğu kişi bunu görmezden geliyor. İnsanlar, Büyük Veri ile çalıştıklarında istatistiğin temel varsayımlarının çok önemli olmadığını düşünebilir. Çünkü şu an büyük veri setlerinin toplanması büyük korelasyonların bulunmasını sağlamıştır. Bu korelasyonlar nedenselliğe sebep olur ki, bu yeterli sayılır. Büyük verinin en çok kullanıldığı alanlar; şekillerin, eğilimlerin, korelasyonların ve birlikteliklerin olmadığı yanlı modellerdeki korelasyonların incelenmesi ve bulunmasıdır[28]. Aslında gözlenen etkilerin nedenlerini Büyük Veri'yi analiz ederek bulmak herhangi bir iş, bilim, devlet, sağlık veya güvenlik sektörü için bugün altın kadar değerlidir.
2. **Büyük Veri de bile örneklem varyansı sıfıra gitmez:** Araştırmacılar genellikle örneklem büyüklüğünü arttırdıklarında istatistiksel karmaşıklığın (noise) azalacağını sanırlar. Fakat örneklem varyansı ile istatistiksel karmaşıklık aynı şey değildir[29]. Örneklem varyansı anakütlenin temel bir özelliği iken istatistiksel karmaşıklık ölçüm sürecinin bir özelliğidir. Sonuç olarak tahmin modellerimizdeki hata belirli bir eşik değerin altına indirgenemez ki, buna içsel örneklem varyansı (intrinsic sample varians) denir. Çok değişkenli kompleks modeller için büyük örneklem kullanıldığında anakütleyi temsil eden farklı parametrelerin varyanslarının tahmini daha doğru olacaktır. Bu anakütlenin temel özelliklerinden biri olabilir. Eğer anakütlenin farklı birimlerinden daha fazla veri toplanılırsa anakütlenin temel istatistiksel özelliklerinin daha iyi bir tahmini yapılabilir.
3. **Büyük Veri'de bile örneklem sapmasının sıfıra gitmesi gerekmez:** Sapma (eğilim), yanlı veri toplama metodunda ya da modeller uygun şartlar altında düzeltilmiş olduğunda istatistikte sapma özelliği göz ardı edilir. Sapma genellikle zayıf model tasarımının bir sonucu olup, eldeki verinin büyüklüğünden bağımsızdır. Albert Einstein'nin söylediği gibi: "modeller mümkün olduğunca basit olmalı, ama daha basit değil." Büyük Veri çağında basit bir öngörü modeli kullanmak hala mümkünken, toplanan verideki

[28] MCAFEE Andrew – Erik BRYNJOLFSSON, *"Big Data: The Management Revolution."*, Harvard Business Review, 90, 10 October 2012, pp. 60-66.
[29] ALLISON Paul D., "*Missing Data*", Thousand Oaks, CA: Sage University Paper No. 136, 2002.

ilişkili desenlerin çoğu ihmal (göz ardı) edilir. Diğer taraftan örneklem hacmi büyük alınırsa yanlılık ortadan kalkar ve korelasyonlu (ilişkili) olan faktörler analizde mevcut ise istatistiksel olarak bağımsızlığın olmadığı varsayılır. İstatistiksel olarak 3. gerçek (Büyük Veri'de bile örneklem sapmasının sıfıra gitmesi gerekmez) bizi şöyle uyarıyor: Şu an büyük veriye sahip olsak bile modelleme yapmak için kullandığımız bu veriler Büyük Veri anlamına gelmez.

4. **Kanıtın Yokluğu, Yokluğun Kanıtı ile aynı şey değildir**: Büyük Veri çağında, henüz herşeyi ölçemediğimizi kolayca unuttuk. Her yerde veri yığını olsa bile, hala belirli bir konu hakkında mümkün olan yeterli veriyi toplamış değiliz. Sonuçta istatistiksel analizde peşin kararlardan kaçınmak için eksik verilerin (absence of evidence) olabileceğinin farkına varılmalıdır. Aksine, eğer "yokluğun kanıtı" ispat edilirse, "yokluğun kanıtı" bilginin çok değerli bir parçası olur. Bu istatistiksel kavramın değerini anlayamadığımız dramatik bir örnek 1986'daki Shuttle Challenger faciasıdır. Mühendisler O-halkalarının soğuk havada başarısız olduklarına dair kanıtı olmamasını, bu halkaların başarısız olmayacağını farzettiler[30]. Bu olay uç bir durum olmakla birlikte, Büyük Veri çağında istatistiksel olarak 4. gerçeğin (Kanıtın Yokluğu, Yokluğun Kanıtı ile aynı şey değildir**)** bugün hala ihmal edilmesi kaçınılmazdır.

1.4. Veri Bilimci Kimdir?

Veri bilimciler, zengin veri kaynaklarını bulmak ve yorumlamak için verilerini ve analitik kabiliyetlerini kullanırlar. Veri bilimciler; donanım, yazılım ve bant genişliği kısıtlamalarına rağmen büyük miktarda veri yönetmek; veri kaynaklarını birleştirme; veri kümelerinin tutarlılığını sağlamak; verileri anlamaya yardımcı olmak için görselleştirmeler yaratmak; verileri kullanarak matematiksel modeller kurmak; veri bulgularını sunmak ve iletmek ile uğraşırlar. Veri bilimciler, istatistikçilerin aylarca cevap aradıkları sorulara günler içerisinde ulaşmaktadır. Bununla birlikte bu kişiler,

[30] CASELLA George – Robert CHRISTIAN, *"Monte Carlo Statistical Methods"*, Second Edition, Springer Verlag, 2004.

analizler ve hızlı iterasyonlarla çalışır ve kâğıtlar/raporlar yerine gösterge tablolarıyla sunumlarını yapmaları beklenilir[31].

Google'da ekonomist olan, Hal Varian 2009'da şunları söylemiştir: "Gelecek 10 yıl da istatistik en popüler meslek olacaktır. İnsanlar şaka yaptığımı sanıyorlar, fakat kim 1990'ların en popüler mesleğinin bilgisayar mühendisliği olacağını tahmin edebilirdi?" Varian, daha sonraki makalesinde "Veri Bilimci: 21. yüzyılın en popüler mesleği"[32] olacağını idda etmiştir. Bu iddia bilim adamları arasında yeni bir tartışmaya sebep olmuştur. Özellikle Fidelity Bank, Uber, Edmunds.com, Pinterest, Facebook, LinkedIn ve Twitter'dan ötürü "Veri Bilimci" kavramını daha sık duyulmaya başlandı. Bugün birçok medyada ve iş ilanlarında "Veri Bilimci" talebinin giderek daha da arttığı görülmektedir (Bkz. Şekil 1.4).

Şekil 1.4: Veri Bilimici İş Trendindeki Yüzde Artış
Kaynak: https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html, (10.08.2017).

Peki, Veri Bilimci kimdir? Bununla ilgili birçok tanım yapılmakla birlikte, Hal Varian şöyle tanımlamaktadır: *"Veri Bilimci; bugün ki yapısal olmayan bilgi*

[31] Ratheesh's – Tech Blog, *"Data Science"*, 2016, http://rathishnair.com/techblog/data-science-machine-learning, (01.01.2016).
[32] DAVENPORT Thomas H. – D.J. PATIL, *"Data Scientist: The Sexiest Job of the 21st Century"*, October 2012, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, (10.05.2016).

tsunamisinden (selinden, fırtınasından) faydalanarak önemli iş sorunlarına cevap bulmak için bilgiyi nasıl kullanabileceğini bilen kişidir"[33].

Umut A. Yıldız ve Selçuk Topala göre: *"Veri Bilimci; herhangi bir yazılım mühendisinden daha iyi istatistik ve herhangi bir istatistikçiden daha iyi yazılım mühendisliği bilen kişidir. Bu da yazılımcıların veri bilimi için ilk sırada tercih edilmediğini gösteriyor. Bilgisayarcılar yazılımı ne kadar çok bilirse de, veri bilimcinin öncelikli görevi "çok iyi istatistik bilgisiyle verileri kullanarak sorunları çözmek hatta daha önceden sorun olduğu bile tahmin edilemeyen şeyleri öngörerek çözüm üretmek" olarak tanımlamışlardır"*. Kısacası Veri Bilimci; bilgisayar bilimi, istatistik, analitik modelleme ve matematik alanında becerileri gelişmiş, sezgileri kuvvetli ve iletişim yeteneği gelişmiş olan kişidir[34].

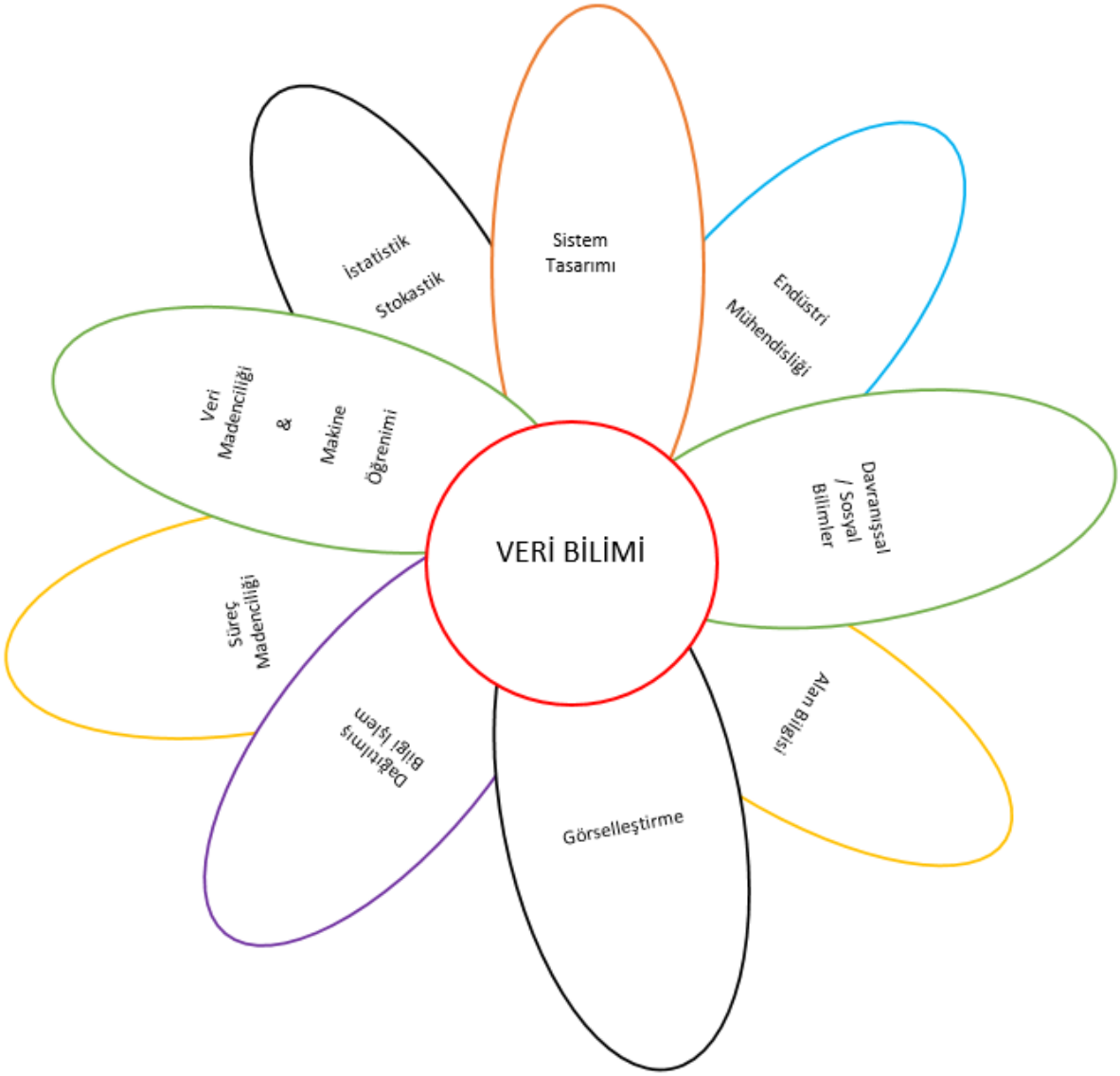
Şekil 1.5 ideal bir Veri Bilimci'nin profili tanımlanmaktadır. Şekil 1.5'te görüldüğü gibi veri bilimi multi-disipliner bir alandır. Dahası Şekil 1.5'ten de görüleceği üzere veri bilimi analitikten/istatistikten çok daha fazlasıdır. Aynı zamanda davranışsal/sosyal bilimleri, (Örneğin, etik ve insan davranışını anlamak), endüstri mühendisliğini (Örneğin, veri değeri ve yeni iş modelleri bilmek) ve görselleştirmeyi içermektedir. Büyük veride sadece EşleİndirgeMapReduce) olmayıp, aynı zamanda veri madenciliğinden daha fazlası da vardır. Veri Bilimci, analiz yöntemlerinin teorik bilgisine sahip olmanın yanında yaratıcı entelektüel merak ve güçlü bilişim teknolojilerini kullanarak etkin çözümler yapar. Birçok işveren, veri bilimcilerini, bir organizasyonun her düzeyindeki insanlara veri bilgisini sunmayı bilen veri hikayecileri olmasını ister. Ayrıca, bir organizasyonda veri odaklı karar verme süreçlerini yönlendirmek için liderlik becerilerine ihtiyaç duyulur.

Veri bilimcilerin eğitim gereksinimleri tipik olarak istatistik, veri bilimi, bilgisayar bilimi veya matematik alanında lisans derecesini içermektedir. İş için gereken zor beceriler arasında veri madenciliği, makine öğrenimi ve yapısal ve yapılandırılmamış verileri bütünleştirme özelliği bulunmaktadır. Bununla birlikte veri bilimcilerin,

[33] AALST Wil M. P. van der, '' Data Scientist: The Engineer of the Future'', In book: Enterprise Interoperability VI, Proceeding of the I-ES Conferences 7, DIO:10.1007/978-3-319-04948-9_2, 2014, pp. 13-26.

[34] YILDIZ Umut A. – Selçuk TOPAL, "Büyük Veri Kahramanı Veri Bilimci" Bilim ve Teknik, Nisan 2015, ss.76-79.

modelleme, kümeleme ve segmentasyon gibi istatistiksel araştırma teknikleriyle ilgili deneyimlere sahip olması gerekmektedir.



Şekil 1.5: Veri Bilimci Profili
Kaynak: AALST Wil M. P. van der, *"Data Scientist: The Engineer of the Future"*, In book: Enterprise Interoperability VI, Proceeding of the I-ES Conferences 7, DIO:10.1007/978-3-319-04948-9_2, 2014, pp. 13-26.

Türkiye de ilk kez 2014 yılında Murat Mığdısoğlu, Teradata Türkiye'de "Veri Bilimci" olarak çalışmaya başladı. Mığdısoğlu'na göre: "Veri Bilimci, *istatistik bilgisi, yüksek miktarda veriyi işlemek için gerekli teknolojilere hâkimiyet, en az bir programlama dilinde uzmanlık, tecrübe ve iş bilgisi gerektiren bir meslek dalı olarak açıklanabilir. Veri bilimcileri, farklı kaynaklardan gelen farklı yapıdaki veriyi bir araya getirir; gerekli temizlemeyi yapar; iş birimlerinin ihtiyaçlarını anlar; veri üzerinde*

yapılabilecek analizleri belirler ve bunu mümkün kılacak teknolojileri seçerken, iş birimlerine yol gösterecek sonuçlar elde etmek için de bir takım çalışması yürütür." Mığdısoğlu, sadece istatistik ya da programlama dilini bilmek bu meslek için yeterli olmayacağını şu sözleriyle anlatıyor: "*Büyük hacimdeki farklı veri tiplerini bir araya getirmek, verideki kirliliği temizlemek, analitik modeli geliştirmek ve bu modeli büyük veriyle başa çıkabilecek teknolojilerde hayata geçirmek, veri bilimcinin işinin bir parçasıdır. Bu nedenle üniversite eğitiminde sadece bilgisayar mühendisliği değil, istatistik ve matematik mühendisliği gibi bölümlerde eğitim görenler de veri bilimci olabilirler*." Mığdısoğlu'na göre veri bilimci ikiye ayrılır: Birincisi dikey veri bilimci, bu kişiler doktorasını genelde istatistik üzerine yapmış kişilerdir. İkincisi yatay veri bilimci, bu kişiler ise istatistik, dağıtık programlama, veri tabanları, programlama ve yapay zekâ gibi birden fazla alanda uzmanlaşmış kişilerdir[35].

Bugün Dünya'daki gelişmiş ülkelerde ve özellikle de ABD'de veri bilimi çok popüler hâle geldi. "Veri Bilimci olmak ABD'de bir numaralı iş hâline geldi"[36]. ABD'nin 44. devlet başkanı Barack Hussein Obama ilk kez Şubat 2015'te Beyaz Saray'a Dr. DJ Patil'i Baş Veri Bilimci olarak atadı. Yine bugün dünyada çok sayıda veri bilimi üzerine master ve doktora programı açılmış ve daha da açılmaya devam edeceği tahmin ediliyor.

Türkiye de ise 2016 yılı itibariyle veri bilimi alanında kayda değer büyük bir gelişme olmamakla birlikte, 2014 yılında ilk kez Sabancı Üniversitesi IBM ile iş birliği yaparak Veri Analitiği Tezsiz Yüksek Lisans programı açmıştır. Yine aynı yıl İstanbul Şehir Üniversitesi IBM Türk Ltd. Şti işbirliği ile ortak olarak Veri Mühendisliği Tezli Yüksek Lisans Programını açmıştır. Gelecekte daha büyük önem kazanacak olan veri bilimi özellikle üniversitelerin sayısal alanlarından mezun olanlar için iyi bir alternatif olacaktır.

1.5. Büyük Veri Tanımı ve Özellikleri

1990'lardan günümüze kadar ki süre *Bilişim Çağı* veya *Bilgi Çağı* olarak kabul edilir. Bu dönemde bilgi kavramı değişen anlam ve içeriği ile ortaya çıkmıştır. Bilişim

---

[35] EticaretMag, *"E-Ticarette Veri Bilimcilerin Önemi Artıyor"*, 06 Ağustos 2014, http://eticaretmag.com/e-ticarette-veri-bilimcilerin-onemi-artiyor, (29.06.2016).

[36] GUTIERREZ Daniel, *"Evolution of the Data Scientist: How Number Crunching Became the Number One Job in America"*, March24, 2014, https://insidebigdata.com/2016/03/24/evolution-of-the-data-scientist-how-number-crunching-became-the number-one-job-in-america/ (28.12.2016).

Çağında, özellikle bilişim ve iletişim teknolojilerinde çok hızlı gelişmeler sağlandı. Bugün bilişim ve iletişim altyapısının Dünya üzerinde hemen hemen her yere ulaşmış olması ve mobil iletişim teknolojileri sayesinde veri ve enformasyon erişiminin zaman ve mekândan bağımsızlaşması, bireylerin, kurumların ve toplumların birbiri ile olan ilişkilerinin bir bölümünü iletişim ve bilgisayar ağları üzerinden yürütebilmelerine imkân sağlamıştır. Bu teknolojik gelişmeler; toplumsal, ekonomik ve bilimsel değişimin yeniden belirlenerek ağ toplumunun ortaya çıkmasını sağlamıştır[37]. Ağ toplumunun ortaya çıkmasıyla birlikte üretilen bilgi miktarında ve hızında büyük bir artış yaşanmış ve bilginin gücü ön plana çıkmıştır. Ortaya çıkan bu bilgi başlangıçta "Bilgi Çöplüğü" olarak anılıyordu. Fakat bugün gelişen teknoloji sayesinde bu çöplükten anlamlı verilerin de çıkabileceğini düşünen birçok yazılım şirketleri ve bilim adamları çalışmalarını sürdürerek "Büyük Veri" olarak adlandırılan kavramı karşımıza çıkardı. Büyük Veri, e-postalar, sosyal medyadaki tıklanma sayısı, fotoğraf, ses, video, log dosyaları… vb. farklı kaynaklardan gelen verilerin anlamlı ve işlenebilir hâle dönüştürülmesi ile ilgilenmektedir.

1.5.1. Veri Akışı

Son dönemlerde "veri akışı" kavramının ortaya çıkmasıyla birlikte veri bilimi olarak adlandırılan yeni bir paradigma ortaya çıktı. Bu paradigma değişimi doğal olarak ortaya çıkmıştır. Yüzyıllar boyunca, bilim çoğunlukla deneysel gözlemlere dayalı olarak gelişmiştir. Daha sonraki süreçte teorik modeller kullanılarak dünya hakkında gözlemler yapılmıştır. Bu modeller çok karmaşık olduğundan ancak, teknolojinin elverdiği imkânlar çerçevesinde çözülür ve analitiksel olarak yorumlanırdı. Bilim dünyasında hesaplama paradigmasının ortaya çıkmasıyla birlikte, teorik modellerin analizi ve simülasyonu için bilgisayarlar kullanılmaya başlanmıştır. Ancak bu hesaplama odaklı bilim, bilimsel veri setlerinde sürekli bir büyümeye yol açmaktadır. Bu büyüme trendi, bilim adamları için verinin toplanması ve dağıtımı için kullanılan araçların etkinliğinin ve çeşitliliğinin artışını da hızlandırmıştır. Bu gelişmeler bilim adamlarını, büyük veri setlerinde keşifler yapmaya sevk etmiştir.

Information Management'ın yakın tarihteki bir yazısında " Dünyada her gün 2.5 katrilyon (1018 bayt) veri yaratılmakta ve dünyadaki verinin %90'nı sadece son iki yılda

[37] Wikipedia, *"Bilişim Çağı",* 2016, https://tr.wikipedia.org/wiki/Bilinote-1, (10.05.2016).

yaratılmıştır. Her saat, Amerikan perakende satış mağazalar zinciri olan Wal-Mart'de 1 milyon işlem yapılmakta, veritabanında 2.5 petabayt (1015 bayt) veri kaydedilmekte olup bu veri Amerikan Kongre Kütüphanesi'ndeki verinin yaklaşık 170 katı büyüklüğündedir. ABD Posta Servisi yılda sadece 5 petabaytlık gibi bir veriyi toplarken, Google sadece bir saatte bu kadar veriyi işleyebiliyor. Bugün dünyada yaratılan bilgi miktarının toplam olarak 1 zettabayt'ın biraz üzerinde olduğu tahmin edilmektedir[38].

TechAmerica Foundation Big Data Commission tarafından yapılan bir çalışma da ABD Hükûmeti ile ilgili şu ifadeler yer almaktadır: "2000 yılından bu yana, federal hükûmetin sahip olduğu kayıtlı bilgi miktarı katlanarak arttı. 2009'da ABD hükûmeti 848 petabayt veri üretti ve ABD sağlık verileri tek başına 150 exabayt'a ulaştı. Yeryüzünde bu güne kadar insanlar tarafından konuşulan tüm sözcükler dünyadaki verinin 5 exabayt'lık (10^{18} $bayt$) kısmını oluşturmaktadır. Bu gidişle, ABD sağlık için Büyük Veri de çok yakın zamanda zettabayt (10^{21} $gigabayt$) ve kısa süre sonra da yottabayt (10^{24} $bayt$) büyüklüğünde bir veriye ulaşacaktır[39]."

Aşağıdaki Tablo 1.1'de Büyük Veri'nin ölçülmesinde en çok kullanılan ölçü birimleri verilmiştir.

Sembol	İsim	Bayt Değeri	İkili Değeri (Bayt)
KB	Kilobayt (Kilobyte)	$10^3 = 1000^1$	$10^{10} = 1024^1$
MB	Megabayt (Megabyte)	$10^6 = 1000^2$	$10^{20} = 1024^2$
GB	Gigabayt (Gigabyte)	$10^9 = 1000^3$	$10^{30} = 1024^3$
TB	Terabayt (Terabyte)	$10^{12} = 1000^4$	$10^{40} = 1024^4$
PB	Petabayt (Petabyte)	$10^{15} = 1000^5$	$10^{50} = 1024^5$
EB	Exabayt (Exabyte)	$10^{18} = 1000^6$	$10^{60} = 1024^6$
ZB	Zettabayt (Zettabyte)	$10^{21} = 1000^7$	$10^{70} = 1024^7$
YB	Yottabayt (Yottabyte)	$10^{24} = 1000^8$	$10^{80} = 1024^8$

Tablo 1.1: Veri Ölçü Birimleri

[38] BETTİNO Larry, *"Transforming big data challenges into opportunities"*, HealthData Management, *18* April 2012, http://www.healthdatamanagement.com/news/big-data-Starvest-IBM-Walmart-44338-1.html?zkPrintable=true, (10.12.2016).

[39] MILLS Steve –Steve LUCAS – Leo IRAKLIOTIS – Michael RAPPA – Teresa CARLSON – Bill PERLOWITZ, "*Demystifying Big Data–A Practical Guide to Transforming The Business of Government"*, prepared by TechAmerica Foundation's Federal Big Data Commission, 2016, https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf, (10.12.2016).

Charles Sadron Enstitüsü ve Aix-Marseille Üniversitesi'ndeki bir grup bilim adamı, insan saçına göre 60.000 kat daha ince olan sentetik polimerler üzerine birtakım verileri kaydetmeyi başardı. Jean Francois Lutz, *Nature Communications* dergisindeki makalesinde, bu gelişmenin gelecekte terabaytlarca bilginin nano boyutlarda saklanabilmesine imkân sağlayacağını söylemektedir. Bugün bir zettabayt (1 milyar terabayt) bir ton kobalt alaşımlı depolama cihazlarında depolanabiliyor. Oysaki Lutz'un sözünü ettiği bir zettabayt veri sadece, 10 gram ağırlığında olup bunlar sanal ortamda 1 ve 0'ı temsil edecek şekilde kodlanıp depolanmaktadır. Adı geçen bilim adamları bu sistemi kimyasal bileşikler üzerine uygulamışlardır. Bazı bileşikler 1'i temsil ederken bazısı da 0'ı temsil etmektedir. Böylece oluşturulan sentetik polimerler istenen verileri rahatlıkla kayıt altına alabilmektedir. Harvard Tıp Okulu'ndaki bir grup bilim adamı ise DNA'nın içerisine veri kaydetmeyi başarmıştır. DNA'da kodlar ikili değil dörtlü şeklindedir. Bilgisayarda oluşturulan 2 kod DNA'da bulunan bazlar (kodlar) eşleştiriliyor. Bu yöntem sayesinde bu grup 10 megabayt veriyi/bilgiyi DNA içerisine yerleştirmiştir. Böylece bu işlem sayesinde çok büyük miktardaki bilginin kolayca arşivlenebileceği düşünülmektedir. 10.000 yıldır DNA hiçbir değişime uğramadı ve sadece bir damlasında bile petabaytlarca bilgiyi saklayabiliyor. Bu da gösteriyor ki, DNA manyetik disklerden çok daha kullanışlı bir depolama aygıtı olabilir. Fakat burada önemli bir kısıtlama ile karşılaşılır. Sadece 10 megabayt verinin depolanması bile saatler sürmektedir. Lutz yine de umutsuz olmadığını ve tüm bunların iki ya da üç yıl içerisinde gerçekleşeceğini düşünmektedir. Lutz sentetik polimerler projesinin DNA'dan çok daha iyi bir depolama yöntemi olduğunu söylüyor*, " Biyoloji ve evrim tarafından dizaynedilmiş olan DNA biyolojik yaşam için olmazsa olmaz, fakat nano boyutlara indiğinizde işler değişiyor. Bizim fikrimiz, DNA'dan daha kullanışlı ve ucuz bir polimer meydana getirmek ve depolama işlemini onun içerisine yapmak."*[40]

Bugün insanlık hızlandırılmış değişimin sabit olduğu bir dünyaya girdi. Günümüz teknolojilerin gelişmekte olan hızı, insanlığın var olduğundan beri çok farklıdır. Son iki yüzyıla bakıldığında, toplumu bilinen şekliyle değiştiren çok sayıda icat görülmektedir. İlk olarak, kitap basım için icat edildi ve kitaplar halka açık hale getirildi. Daha sonra insanlar dünyadaki herhangi bir sanayiyi önemli ölçüde değiştiren ve insanlığı bir üst

[40] Habermag, *"Nano Boyuttta Veri Depolamak"*, 2016, http://habermag.net/nano-boyutlara-veri-depolamak/ (21/12/2016).

seviyeye taşıyan buhar makinesinin icadına tanık oldu. Son olarak, 20. yüzyıda internet ve bilgisayar keşfedildi. Bu en son "bilgi devrimi", endüstriyel devrimin buhar makinesiyle çalışmasından oldukça farklıdır.

Peki, dünya zaman içinde bu noktaya nasıl geldi, dünya nerede sürüm 1.0'dan sürüm 2.0'ye yükseldi, yükseltilen daha akıllı olan sürüm hangisi? Hiç kuşku yok ki tüm bunların ortaya çıkmasında birçok trend etkin rol oynamıştır. Dünya'nın 2.0 dijital yönü üzerine odaklanılırsa, bilgi devriminin yaşanmasına sebep olan üç büyük trend üzerinde durmak gerekmektedir. Bunlar[41]:

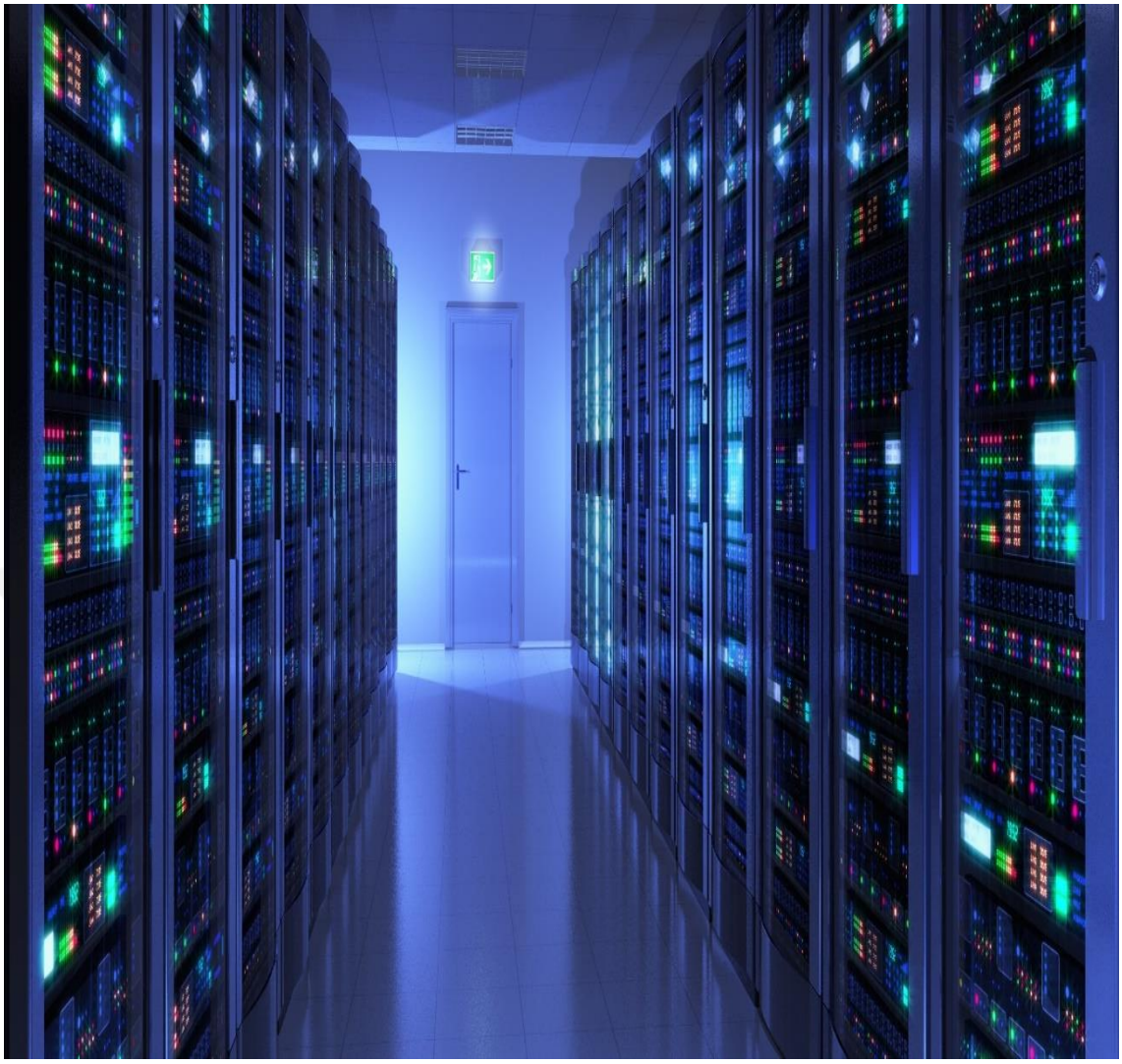
- Buluttaki sınırsız işlem gücü
- Her şeyi akıllı hâle getirecek sensörler kümesi
- Akıllı algoritmalar sayesinde Yapay Zekâ ve Makine Öğrenme

Bu üç büyük trend yaptığımız iş de dâhil olmak üzere, toplumun herhangi bir bölümünü etkileyecektir. Şimdi bunların her birini incelemeye çalışalım.

1.5.1.1. Sınırsız İşlem Gücü

Son yıllarda, veri depolama fiyatı Gigabayt başına yaklaşık olarak $0,03'a kadar düşmüştür. Moore yasasına göre bu fiyatın önümüzdeki yıllarda daha da düşmesi beklenmektedir. Bu sayede kuruluşlar için veri toplama ve depolama sorunu büyük ölçüde ortadan kalkmıştır. Bu da herhangi bir verinin toplanabilmesi ve depolanabilmesi, akıllı algoritmalar kullanılarak analiz edilebilmesinin yanında çok daha fazlasının yapılabileceği anlamına gelmektedir.

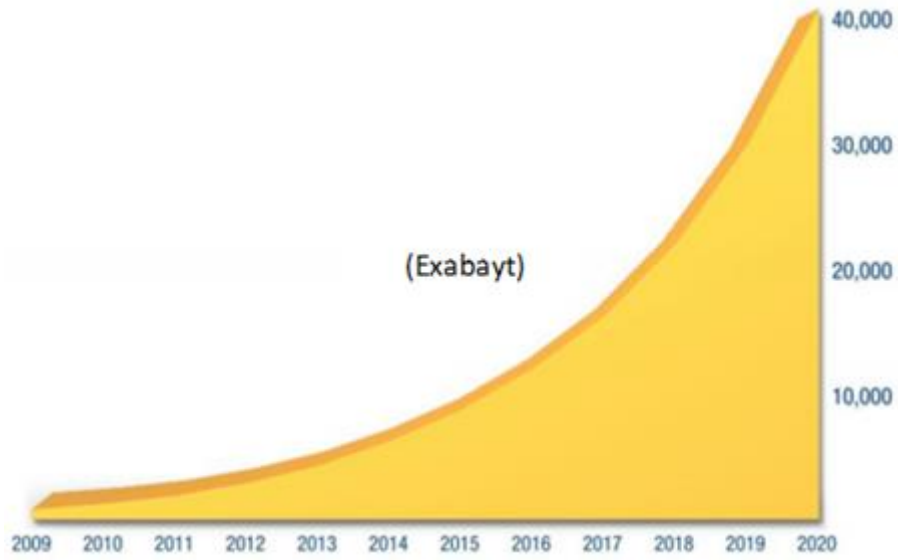
[41] DATAFLOQ, *"How Unlimited Computing Power, Swarms of Sensors and Algorithms Will Rock our World"*, 20 JUNE, 2016, https://datafloq.com/read/unlimited-computing-swarm-sensors-algorithms-world/2138, (10.01.2017).



Şekil 1.6: Veri Depolama Merkezi
Kayak: https://geturkiyeblog.com/wp-content/uploads/2017/02/generalelectric-buyuk-veri-, (10.01.2017).

Bununla birlikte, daha akıllı bir dünya için gerekli olan tek şey verinin depolanması değildir. Özellikle verilerin analiz hızı kuruluşlar için hayati öneme sahiptir. Yine, Moore yasasına göre depolanan bilgi miktarı geçmiş yıllarda üssel (exponential) olarak artış (Bkz. Şekil 1.7.) göstermiştir. Depolanan bilgi miktarındaki bu artış gelecekte de devam edecektir. Sonuç olarak, bilgisayar mikroişlemcisinin bugünün süper bilgisayarlarında kullanılmasıyla, 2020 yılına kadar hiçbir bilgi boşa akıp gitmeyecektir. Muhtemelen 2020 yılına kadar herkesin cebinde bir süper bilgisayarı olacaktır[42].

[42] DATAFLOQ, *"How Unlimited Computing Power, Swarms of Sensors and Algorithms Will Rock our World"*, 20 JUNE, 2016, https://datafloq.com/read/unlimited-computing-swarm-sensors-algorithms-world/2138, (10.01.2017).



Şekil 1.7: Dijital Dünya'daki Büyüme 2010-2020
Kaynak: IDC's Digital Universe Study, sponsored by EMC, December 2012.

Fakat donanım, sınırsız hesaplama gücünün tek bir bileşeni değildir. Son yıllarda, görülen yeni teknolojiler tüm veri analizlerini yapabilecek kapasiteye sahiptir. Burada saniyeler içerisinde analizi yapılan Terabayt ya da Petabayt büyüklüğündeki verilerden bahsedilmektedir.

Apache Spark gibi açık kaynak teknolojileri, binlerce düğümden oluşan dağıtılmış bir ağdaki verilerin bellek içi analizini yapabilme kapasitesine sahiptirler. Bu açık kaynak teknolojileri sensörlerden gelen büyük miktardaki veriyi, Nesnelerin İnterneti'nden (Internet of Things-IoT) gelen verileri ve ayrıca dünyada tüketiciler tarafından yaratılan tüm yapılandırılmamış verileri analiz edebilecek kapasiteye sahiptirler.

Bugün sınırsız hesaplama gücüne süper bir bilgisayar olmadan da artık erişilebilmektedir. Ancak bunun için AWS veya Microsoft Azure gibi bulut hizmeti işbirliği içinde olmak veya konuyla ilgili herhangi bir başka bulut hizmeti kullanmak gerekmektedir. Bulut kullanırken sadece kullanılan şey için ödeme yapılır ve bunu sınırsız olarak ölçeklemek mümkündür. Elbette, bu iş onu nasıl yaptığımıza göre bazı değişiklikler gösterecektir.

1.5.1.2. Sensörler Kümesi

Sınırsız işlem gücü bize geçmiş yıllara bağlı bir dünya yaratmaktadır. Bu dünya giderek daha ve daha fazla bağlantılı olmakla birlikte, gelecek on yıl içinde bir trilyon kadar bağlanabilir cihazın olacağı tahmin ediliyor. Tüm bu sensörler dünyamızı akıllı hâle

getirecek ve bizler bu sensörleri Nesnelerin İnterneti olarak arayacağız. Dünyaca tanınmış ağ teknoloji şirketi Cisco'ya göre, Nesnelerin İnterneti gelecek on yıl içinde 19 trilyon dolarlık bir pazara sahip olacaktır. Üretim için bu oran 2.9 trilyon dolar olarak öngörülmektedir[43].

Gelecekte, aklınıza gelebilecek herhangi bir yerde sensörleri bulmak mümkün olacaktır. Trafiği izlemek için yollara belirli aralıklarla sensörler yerleştirilmekte, sensörler makinalarda tahmine dayalı bakım ya da araçlardaki sürüş davranışlarını izlemek ve buna göre sigorta poliçesi ayarlamak için kullanılmaktadır. Sensörler gelecekte çok ucuz ve çok küçük olacak öyle ki, onları giysilerin ve ilaçların içerisinde bile bulabilmek mümkündür. Daha şimdiden insan derisinin altına yerleştirilen sensörler kullanılmaya başlanmıştır.

Sensörler sürüsü dünyayı kaplayacak ve gerçekten akıllı bir dünya yaratacaklardır. Sensörlerin kullanılışı heyecan verici olup artık bir bilim kurgu olmaktan çıkmıştır. Nitekim Harvard Üniversitesi'ndeki bilimadamları, 1.000 kişilik ordu gücü oluşturmak için "sürü" halinde birlikte hareket eden itaatkar minyatür robotlar olan Kilobot'u geliştirmişlerdir. Bu robotların her biri sadece birkaç santimetre mesafeyi ölçmekte ve bağımsız olarak kendilerini organize edebilmektedir. Ayrıca, bu robotlar küresel konumları hakkında doğrudan bilgi verilmezler, sadece koordineli bir sistem oluştururlar[44].

Kilobotlar bugün için sadece bir başlangıç düzeyindedir. Bu veya daha küçük robotlarla dolu akıllı bir dünya; kuruluşların nasıl yönetileceğini, inovasyona nasıl yaklaşılması gerektiğini ve kuruluşların müşterileri ile nasıl bağlantı kurması gerektiğini önemli ölçüde değiştirecektir. Dolayısıyla, Nesnelerin İnterneti şirketlerin kendi kültürlerini değiştirmelerini zorunlu kılacaktır. Bu da kararların, çoklu veri kaynaklarını ağ ortamında bir araya getirerek derin veri analizi ve görüşlerine dayandığı yeni bir kültürü yaratacaktır.

[43] DATAFLOQ, *"How Unlimited Computing Power, Swarms of Sensors and Algorithms Will Rock our World"*, 20 JUNE, 2016, https://datafloq.com/read/unlimited-computing-swarm-sensors-algorithms-world/2138, (10.01.2017).

1.5.1.3. Akıllı Algoritmalar

Trilyonlarca sensörün ürettiği büyük miktardaki veriyi anlamak için akıllı algoritmalara ihtiyaç duyulmaktadır. Neyse ki geçmişteki algoritmaların geliştirilmesi Yapay Zekâ, Makine Öğrenme ve Derin Öğrenme sayesinde bir üst seviyeye gelindi.

Yapay Zekâ'nın günümüzde çok büyük bir kullanım potansiyeli vardır. 2013 yılında Oxford Üniversitesi tarafından yapılan bir araştırmada, Yapay Zekâ'nın yakın gelecekte ABD'deki tüm işlerin yaklaşık yarısını elinden alabileceği tahmin ediliyor. Yapay Zekâ'nın en yaygın uygulaması, muazzam büyüklükteki veride desen bulma ve otomatik olarak işlem yapma ile ilgilidir. Yapay Zekâ şirketlerin karmaşık tanımlayıcı, öngörcü ve kuralcı analitik görevleri otomatikleştirmelerine ve iyileştirmelerine yardımcı olarak eşi benzeri görülmemiş seviyede değer yaratmalarına imkân sağlamaktadır.

Günümüzde tüm büyük teknoloji şirketleri, Google DeepMind'i ön planda tutarak Yapay Zekâ üzerinde çalışıyor. AlphaGo algoritmasının Go oyununu kazanmasıyla birlikte işin Sezgisel Yapay Zekâ'ya taşınabileceği görülmüştür. Sezgisel Yapay Zekâ, Yapay Zekâ dan daha ileri düzeydedir. Çünkü Sezgisel Yapay Zekâ, işlediği ve analiz ettiği verileri öğrenebilmekte ve kendini geliştirmek için derin öğrenmeyi kullanabilmektedir[45].

Derin öğrenme, yeni uygulamalar geliştirmek ve yeni bilgiler oluşturmak için sinir ağları kullanan çok güçlü bir teknikler dizisidir. Derin öğrenme genellikle konuşma tanıma, görsel nesne tanıma veya nesne tanımada kullanılır. Derin öğrenme; genomik, uyuşturucu tehşisi ve dolandırıcılık algılama gibi uygulamalarda veya daha yaygın olarak algoritmalara dayanan "new Siri"gibi uygulamalarda kullanılmaktadır.

Bugün akıllı algoritmalar dünyayı ele geçiriyor ve çoğu işi devralıyor. Bu nedele, akıllı algoritmaların potansiyeli ve kuruluşlar için neler yapılabileceğinin farkında olunması önemlidir. Aksi takdirde, kuruluşların büyük risklerle karşı karşıya kalabilecekleri unutulmamalıdır.

1.5.2. Büyük Veri Tanımları

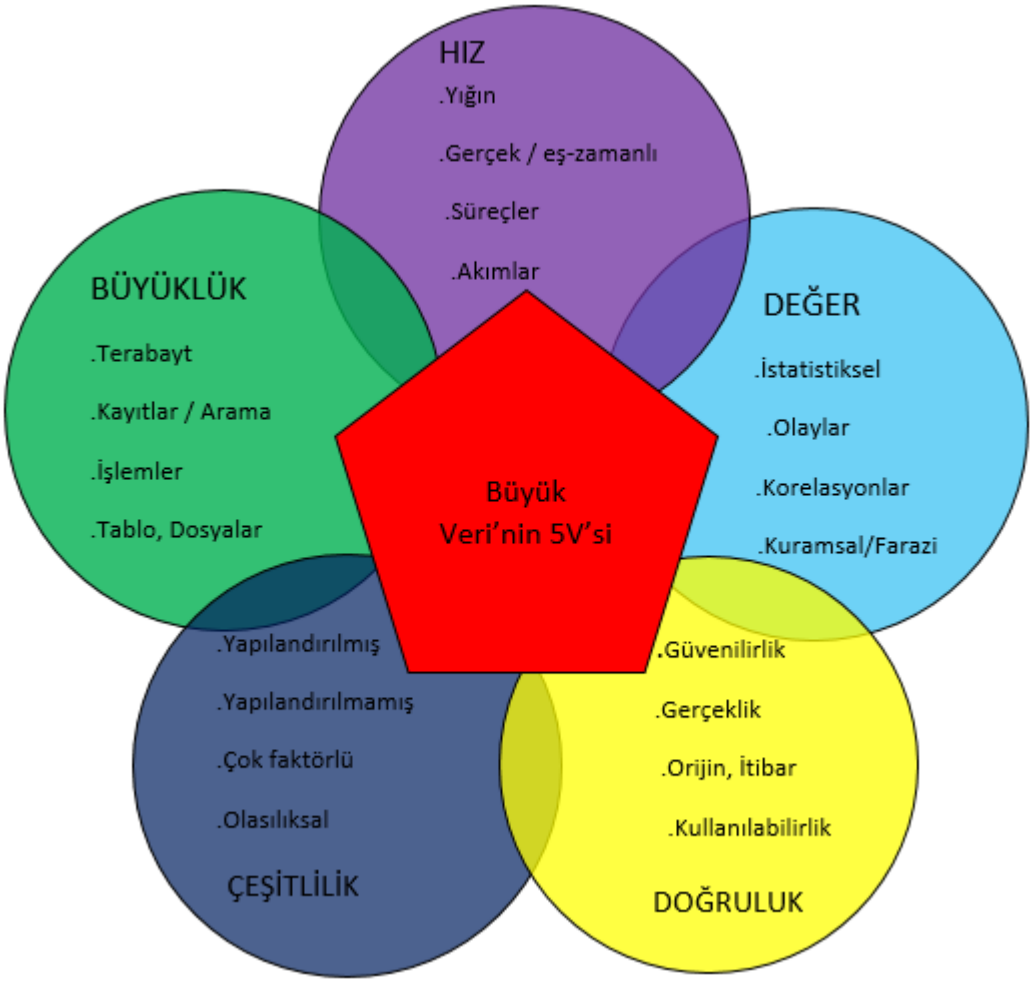
Büyük Veri, yazılım ve donanım şirketlerinin satışlarını arttırmak için sık sık kullandığı bir terimdir. Bununla birlikte etkileyici bir iş potansiyeli olan ve son derece

[45] DATAFLOQ, *"How Unlimited Computing Power, Swarms of Sensors and Algorithms Will Rock our World"*, 20 JUNE, 2016, https://datafloq.com/read/unlimited-computing-swarm-sensors-algorithms-world/2138, (10.01.2017).

önemli olan bir teknolojik trenddir. Büyük Veri kavramı internetin hayatımıza girdiği andan itibaren ortaya çıkmış olmasına rağmen insanlar tarafında yeni yeni fark edilmeye başlanmıştır.

Bugün internette yapılan her hareket veya tıklama veri oluşumuna neden olmaktadır. Bu hareketi gün içerisinde milyonlarca insanın yaptığını düşünülürse her saniyede devasa büyüklüğünde bir veri yığınının oluştuğu anlamına gelmektedir. Fakat bunu sadece sosyal medya ile sınırlandırmak yanlış olur. Öyle ki, neredeyse her gün hayatımıza yeni bir teknoloji girmektedir. Buna örnek olarak Nesnelerin İnterneti gösterilebilir.

Büyük Veri aslında "büyük" değil, çeşitlidir. Buradaki "büyük" kavramı yanıltıcıdır. "Büyük" sözcüğü ile kastedilen, çok sayıda kaynaktan, eş zamanlı olarak gelen bol miktardaki veridir. "Büyük Veri" kavramını daha iyi anlamak için onu oluşturan temel bileşenleri incelemek gerekir. Büyük Veri'nin 5V'si (Bkz. Şekil 1.8) olarak adlandırılan bu bileşenler: büyüklük (volume), hız (velocity), çeşitlilik (variety), değer (value) ve doğruluk (veracity) dır.



Şekil 1.8: Büyük Veri'nin 5V'si
Kaynak: ZORLU Emre, *"Big Data"*, 12 TEMMUZ 2012.

1. Veri Büyüklüğü: Şirketlerin karar alma süreçlerinde kullandıkları çok büyük miktardaki veriyi ifade etmektedir. Verinin büyüklüğü büyük bir hızda artmaya devam etmektedir. Bazı uzmanlar Büyük Veri'nin başlangıç noktasının petabayt olduğunu iddia etmiştir. Birçok şirket ise bir terabayt ve bir petabayt arasındaki veri setlerini Büyük Veri olarak kabul etmektedir. Öyle ki üretilen veri miktarı arttıkça, başlangıç noktası da çok hızlı bir şekilde büyümektedir. Yani, bugün büyük olan bir veri, yarın küçük olabilecektir.
2. Hız: Hız; verinin üretilmesi, işlenmesi ve analiz edilmesi için devamlı artmaktadır. Daha yüksek hız ise üretilen verinin doğal gerçek-zamanında ve akan verinin iş süreçleri ile birleştirilmesi ihtiyacından kaynaklanmaktadır. Bugün, veri sürekli artan bir oranda üretilmektedir. Ancak bu veriyi geleneksel yöntemler kullanarak yakalamak, depolamak ve analiz etmek mümkün değildir. Özellikle çok kanallı anlık pazarlama gibi zamana

duyarlı süreçler için ve verinin iş değeri için eş zamanlı olarak analiz edilmesi büyük önem teşkil etmektedir.

3. Çeşitlilik: Büyük Veri yapısında fotoğraflar, tıklanma sayıları, e-postalar, sesler, videolar, HTLM, PDF ve ekg verileri gibi çok çeşitli veri türlerini barındırmaktadır. Bu veriler; yapılandırılmış (düzenlenmiş), yarı yapılandırılmış ve yapılandırılmamış türdeki veriler. Büyük Veri'nin büyük kısmını, yapılandırılmamış yani klasik formatta satır ve sütuna yerleştirilemeyen veriler oluşturmaktadır[46].

4. Değer: Büyük Veri'yi açıklayan 3V **(**volume, velocity ve variety) tanımları yapıldıktan sonra verinin toplanıp işlenmesinden sonra işe yarar sonuçlar elde etmek için bu analizlerin bir " değer" üretmesi gerekir. Dolayısıyla bu V diğer 3V'nin **(**volume, velocity ve variety) birleşim noktasıdır.

5. Doğruluk: Doğruluk belirli veri türleri ile ilgili güvenilirlik düzeyi anlamına gelmektedir. Büyük Veri için yüksek veri kalitesi önemli bir gereklilik ve mücadele arayışıdır. Fakat en önemli veri temizleme yöntemleriyle bile bazı verilerin (hava durumu, ekonomi ya da bir müşterinin satın alma kararları gibi) doğasında var olan tahmin edilemezliği kaldırılamaz.

TechAmerica Foundation Big Data Commission Büyük Veri ile ilgili şu tanımı vermiştir: "Büyük Veri; yüksek hızlı, kompleks ve değişken verinin büyük miktarının (hacminin), bilginin yakalanması, depolanması, dağıtımı, yönetimi ve analizi için gelişmiş teknikler ve teknolojiler gerektiren bir terimdir."

McKinsey Global Institute 2011 raporuna göre "Büyük Veri: veri setleri anlamına gelir ki bu veri setlerinin hacmi; yakalama, depolama, yönetme ve analiz etmek için tipik (geleneksel) veritabanı yazılım araçlarının yeteneğinin (özelliklerinin) çok daha ötesindedir. "Büyük Veri" terimini belli bir terabayt (binlerce gigabyt) sayısından daha büyük şeklinde tanımlayamayız."

Teknoloji zamanla geliştikçe Büyük Veri'yi ifade eden veri setlerinin boyutunun da artmasına neden olmuştur. Ayrıca Büyük Veri tanımı sektörün yaygın olarak bağlı bulunduğu yazılım araçlarına ve belli bir endüstride yaygın olan veri setlerinin boyutlarına bağlı olarak değişebilmektedir. Bunlara bağlı olarak bugün pek çok sektörde Büyük Veri birden fazla petabayt ile birkaç düzine terabayt (binlerce terabayt)

[46] BAYRAKCI Serkan, *"Büyük Veri Nedir?"*, 19 NİSAN 2015, https://serkanbayrakci.wordpress.com/tag/big-data/, (10.11.2016).

aralığındadır[47]. Örnek olarak Intel, Büyük Veri'yi, "Bir haftada ortalama olarak 300 terabayt veri üreten firmalar için Büyük Veri fırsatlarının ortaya çıktığından söz ediyor" şeklinde tanımlamaktadır (Bkz. Tablo 1.2).

Sınıf	Büyüklük	Neyle yönetilir?	Nerede saklanır?	Örnekler
Küçük	<10 GB	Excel, R	Bir makinenin belleği	Binlerce satış sayısı
Orta	10 GB-1 TB	Endekslenmiş dosyalar, monolitik veri tabanları	Bir makinenin diski	Milyonlarca web sayfası
Büyük	>1 TB	Hadoop, Spark, Dağıtık veri tabanları	Çok sayıda makinede saklanır	Milyarlarca web tıklanması

Tablo 1.2: Büyük Veri Nedir?
Kaynak: NARİN Bilge, *"Big Data"*, 24 Mart 2015.

Tim O'Reilly, belki de diğer tüm tanımları içeren çok kısa bir tanım vermiştir. "Büyük Veri: Bilgi depolama maliyetinin karar verme maliyetinden daha az olmasıdır."

IDC, Büyük Veri teknolojilerini, yüksek hızda yakalama, keşif ve/veya analiz yaparak, çok geniş bir veri çeşidinden ekonomik olarak değer ayıklamak üzere tasarlanmış yeni nesil teknolojiler ve mimariler olarak tanımlamaktadır. Büyük Verilerin üç temel özelliği vardır: verilerin kendisi, verilerin analitiği ve analiz sonuçlarının sunumudur[48].

Büyük Veri ile ilgili daha pek çok tanım mevcut olup bu tanımlar mevcut farklı türdeki veri kaynaklarını sınıflandırma yoluna gitmişlerdir.

1.5.3. Veri Kaynakları

Latince de olgu anlamına gelen "veri" kelimesi "bilinen" anlamına gelmektedir. *İstatistikte veri*, işlenmemiş ham bilgi yığınına veya bir araştırma sonucunda ortaya çıkan bilgilere denir. Başka bir ifadeyle yorumlanmak ve sunulmak için bir araya getirilmiş,

[47] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.
[48] GANTZ John – David REINSEL, *"The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.",* Study report, IDC, December 2012.

çözümlenmiş ve özetlenmiş bilgilere veri denir. *Bilgisayar dilinde* ise, bilgisayarın alabildiği, işleyebildiği, sonuç üretebildiği ve hesaplayabildiği her şeye veri denir. Veriler bilgisayara ikili (binary), harfler veya rakamlar şeklinde kodlanarak girilir. Bilgisayara veri girildiğinde eğer bilgisayar bu veriyi tanımaz ise bu veri hiçbir anlam ifade etmez. Fakat veri bilgisayarın onu tanıyacağı formata getirilip girişi yapılırsa o zaman anlamlı ve hayatımızı kolaylaştıracak bilgiler sunabilmektedir. Günümüzde yarı yapılandırılmış ve yapılandırılmamış veri miktarının çok hızlı artması potansiyel veri kaynaklarının miktarının ve çeşitliliğinin artmasına sebep olmuştur. Aşağıda bu veri çeşitliliğini sağlayan kaynaklardan bahsedilmiştir.

Soares Büyük Veri kaynaklarını; web ve sosyal medya, makineden makineye veri, büyük işlem verileri, biyometrik ve insan kaynaklı veri olmak üzere beş kategoride sınıflandırmıştır[49] .

1.5.3.1. Web Verisi & Sosyal Medya

İnternet (Web) veri bakımından çok zengin olmakla birlikte, analiz için çok çeşitli veri kaynaklarına sahiptir. Bu veri kaynaklarından birisi de içerisinde bilgi ya da kamuoyu görüşünü doğrudan seçerek veren internet kaynakları olup, bunlar başlangıçta bir kitleye yönelik oluşmuştur. Bu kaynaklar insanlar tarafından okunabilir olan web sayfaları, çevrimiçi makaleler ve bloglardır[50]. Bu tür internet kaynaklarının genel ortak özelliği metin, video ve görüntü gibi yapılandırılmamış türdeki verileri içermeleridir. Ancak bu kaynakların çoğu belli bir yapıya sahiptir. Örneğin bu veri kaynakları köprüler aracılığı ile birbiriyle ilişkili veya bulut boyunca etiketlenip sınıflandırılmıştır. Daha sonra, makine-okunabilirliği sağlamak için yapılandırılmış bir web içeriği ve bilgisi vardır. Burada verilere erişmek için uygulamaları etkinleştirmek, veriyi sınıflandırarak anlamak, farklı kaynaklardan gelen verileri entegre etmeye izin vermek, yeni bilginin ilişkilendirilmesi ve anlamlandırılması amaçlanmaktadır. Böyle kaynaklar web sayfalarına entegre olmuş makineler tarafından okunabilen büyük veri, bağlantılı açık veri projesi olan girişimlerdir. Ayrıca, bu kaynaklar web standartlarına uygun sınıflanmış

[49] SOARES Sunil, "*Big Data Governance - An Emerging Imperative*", MC Press Online, LLC, 1st edition edition, 2012.

[50] CHEN Jinchuan – Yueguo CHEN – Xiaoyong DU – Cuiping LI – Jiaheng LU – Suyun ZHAO – Xuan ZHOU, *"Big data challenge: a data management perspective",* Frontiers of Computer Science, 7 (2):157–164, April 2013. doi: 10.1007/s11704-013-3903-7.

veri formatları olup aynı zamanda kamuya açık web hizmetleridir. Bu veri tipi genellikle grafik-şekilli olup yarı yapılandırılmış formattadır.

Diğer web kaynakları navigasyon verilerinin iletilmesini sağlar ki, bunlar web kullanıcıları arasındaki etkileşimden ve web'deki hareketlerden bilgi elde edilmesini sağlamaktadır. Bu veriler web uygulamalarının yanı sıra arama sorgularında toplanan günlük tıklanma sayılarını da içermektedir. Şirketler özellikle müşterilerin web üzerinde yaptıkları gezintiden yola çıkarak, müşterilerin satın alma davranışları hakkında fikir sahibi olurlar. İşte burada kullanılan veriler yarı yapılandırılmış formattadır.

Son olarak Büyük Veri, sosyal etkileşimlerden kaynaklanan web verisidir. Bu tür verilere örnek olarak anlık mesaj hizmetleri ya da sosyal medya sitelerindeki durum güncellemelerindeki iletişim bilgileri verilebilir. Bu tür mesajların verileri genellikle yapılandırılmamış metin veya görüntü verileridir. Fakat bu veriler yarı yapılandırılmış seviyesine yükseltilebilir. Örneğin; bu veriler kimin kiminle iletişim halinde olduğunu göstermektedir. Ayrıca sosyal etkileşim, sosyal bağlantıların daha iyi yapılandırılmış bir nosyonunu açıklayan veriler, genellikle sosyal grafik ya da sosyal ağ olarak adlandırılır. Sosyal ağlara örnek olarak facebook'taki "dostluk" ilişkileri verilebilir. Bu tür veriler genellikle yarı yapılandırılmış ve grafik şeklindedir. Unutulmaması gereken bir şey de bu tür iletişim verilerine erişim bugün artık çok kolay hâle gelmiştir. Bu da insanların sosyal medya üzerinden kendileri hakkında yayınladıkları bilgi iletişimi ve kendilerini sunmalarına yol açmaktadır. Fakat bu bilgiler bazen prestij amaçlı yayımlandığı için yanlı, eksik ya da yanlış olabilmektedir[51].

Ayrıca, bu tür farklı web verilerinin mutlaka özel olması gerekmez. Çoğu birleşmiş olabilir. Sosyal medya iletileri bilginin ve iletişimin insanlar tarafından okunabilir yayını olabilmektedir.

1.5.3.2. Makineden Makineye Veri

Makineden makineye iletişim bazı ağlar üzerinden bağlı olan teknik cihazlar ile ilgili iletişim sistemlerini açıklamaktadır. Bu cihazlar hareket veya sıcaklık gibi fiziksel bir fenomeni (olguyu) ölçmek ve bu fenomenin içindeki olayları yakalamak için kullanılırlar. Ağ üzerindeki cihazlar bir uygulama ile iletilmektedir ki, bunlar ölçülen ve

[51] MAIER Markus, *"Towards a Big Data Reference Architecture"*, Eindhoven University of Technology, Department of Mathematics and Computer Science, Master's Thesis, 13th October 2013.

yakalanan olayların mantıklı olanlarından bilgileri ayıklarlar. Bu verilere örnek olarak makineden makineye iletişim olan 'Nesnelerin İnterneti' verilebilir[52].

Makineden makineye veride ölçümler için kullanılan cihazlar; tipik sensörler, RFID çipleri veya GPS alıcılarıdır. Bunlar genellikle bazı sistemler içine yerleştirilirler. Örnek olarak araçlar içine yerleştirilmiş olan teknik tanı yapan sensörler ya da evlerde ortam istihbaratında kullanılan akıllı sayaçlar verilebilir. Bu sistemler tarafından oluşturulan veriyi işlemek oldukça zordur. Örneğin; BMW grubu, bu cihazlara Bağlı Sürücü araçların 2017'de günlük bir petabaytlık veri üreteceğini öngörmektedir[53]. Diğer bir örnek de GPS alıcılarıdır. Bunlar genellikle cep telefonu içine yerleştirilmiş olup aynı zamanda diğer mobil cihazlarda da bulunmaktadır. Yine bu cihazların bir örneği olarak konumsal veri (locational data) üreten, aynı zamanda mekânsal veri (spatial data) olarak adlandırılan cihazlar da verilebilir[54].

1.5.3.3. Büyük İşlem Verileri

İşlem verileri kayıt sistemleri boyutlarının büyümesi ve büyük miktarlarda işlemlerin gerçekleşmesiyle birlikte artmıştır[55]. İşlem verilerine örnek olarak büyük web mağazalarından satın alınabilen mallar, telekomünikasyon şirketlerindeki detaylı kayıtlar ya da kredi kartı şirketlerinin ödeme işlemleri verilebilir. Bu veriler genellikle yapılandırılmış veya yarı yapılandırılmış veri türlerinden oluşmaktadır. Ayrıca büyük işlem verileri; birleştirilmiş ya da insanlar tarafından yaratılan, yapılandırılmamış, çoğunlukla metin verilerinden oluşmaktadır. Büyük işlem verilerine örnek olarak çağrı merkezinde birleştirilmiş kayıtlar ile servis acentesindeki personel notları, trafik kazalarının açıklaması ile birlikte yapılmış olan sigorta talepleri veya doktor tarafından yazılmış sağlık işlemleri olan tanı ve tedavi notları verilebilir.

---

[52] SORES Sunil, "*Big Data Governance - An Emerging Imperative*", MC Press Online, LLC, 1st edition edition, 2012.

[53] Camille Mendler. M2M and big data. Website, *"A report the Economist: Intelligence Unit"*, 2013.

[54] SHEKHAR Shashi – Viswanath GUNTURI – Michael R. EVANS – KwangSoo YANG, *"Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing"*, In Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, MobiDE '12, pages 1–6, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1442-8.

[55] CHEN Jinchuan – Yueguo CHEN – Xiaoyong DU – Cuiping LI – Jiaheng LU – Suyun ZHAO – Xuan ZHOU, *"Big data challenge: a data management perspective"*, Frontiers of Computer Science, 7 (2):157–164, April 2013.

1.5.3.4. Biometrik Veri

Biometrik veri, genel olarak biyolojik organizma verisi olarak tanımlanmakta ve çoğunlukla insanların ayırt edici anatomik ve davranışsal karakterleri ve özellikleri ile bireyleri tanımlamak için kullanılmaktadır. Anatomik özelliklere örnek olarak parmak izi, DNA ya da gözdeki retina taraması verilirken, davranışa örnek olarak el yazısı veya tuşa vuruş analizi (klavye vuruş analizi) verilebilir[56]. Yine biyometrik veriye örnek olarak gen analizi için bilimsel uygulamalarda büyük miktarda biometrik verinin kullanılması verilebilmektedir.

1.5.3.5. İnsanların Ürettiği Veri

Nihai olarak tüm bilgilerin kaynağı insanlardır. Bu bilgiler, insan deneyimlerinin önceleri kitap ve sanat eserleri olarak ve daha sonra da fotoğraf, ses ve video olarak büyük ölçüde kişiye özgü kayıtlarıdır. İnsan kaynaklı bilgi, film ve tweetler bugün artık neredeyse her yerde dijital ve elektronik olarak kayıt altına alınmaktadır. Örneğin insan kaynaklı bilgi için süreç-aracılı veri daha güvenilir iş dönüşümüne izin vererek gerçeğin ortak bir versiyonunu modelleyerek tanımlar. Bu, işletim sistemlerine veri girişi ve doğrulama ile başlanılır ve İş Zekâsı (Business Intelligence-(BI)) için veri akışı temizleme ve toplama süreçleriyle devam eder (Sosyal Ağların Büyümesi**)**[57].

[56] SOARES Sunil, "*Big Data Governance - An Emerging Imperative"*, MC Press Online, LLC, 1st edition edition, 2012.
[57] DEVLIN Barry – Shawn ROGERS – John MYERS, "*Big DataComes of Age"*, EMA and 9sight Consulting Research Report, 11.01.2012, http://www.enterprisemanagement.com/research/asset.php/2409/Big-Data-Comes-of-Age, (17.03.2017).

İKİNCİ BÖLÜM
BÜYÜK VERİ TEKNOLOJİSİ VE TEKNİKLERİ

Veri üretimi her geçen gün katlanarak artmaktadır. Artan bu verinin hacmi, çeşitliliği ve hızı mevcut verilerden daha fazla olduğundan bu veriden, bilgi elde etmek için analiz etmek, yorumlamak ve görselleştirmek oldukça zordur. Milyarlarca ağa bağlı sensör; akıllı telefonlar, otomobiller, sosyal medya siteleri, dizüstü bilgisayarlar, PC'ler ve endüstriyel makineler gibi verileri çalıştıran, üreten ve ileten cihazlara yerleştirilmiştir. Bu nedenle, çeşitli kaynaklardan elde edilen veriler yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış biçimde bulunulmaktadır. Geleneksel veritabanı sistemleri bu veri formatlarını işlemekte yetersiz kalmaktadır. Bu nedenle, bu verilerle çalışmak için yeni araç ve tekniklere ihtiyaç duyulmaktadır. Bugün geliştirilen bazı teknik ve teknolojiler sayesinde çok büyük yapılandırılmış ve yapılandırılmamış veri setleri kolayca analiz edilebilmektedir.

Bu bölümde büyük veri setlerinin analizinde en çok kullanılan teknik ve teknolojilerden bazıları verilmiştir.

2.1. Büyük Veri Teknolojileri

Büyük veri setlerini depolamak, işlemek, yönetmek ve analiz etmek için kullanılan teknolojilerin sayısı giderek artmaktadır. Büyük Veri Teknolojileri her türlü veriyi işleme (esnek), ihtiyaca göre genişleme (ölçeklenebilir), verilerin yedeklenir ve erişilebilir olması (veri garantili) ve açık kaynaklı projeler (düşük maliyetli) olma gibi özelliklere sahiptirler. Burada daha çok büyük veri setlerinin yönetilmesinde ve analizinde yaygın olarak kullanılan teknolojilerin bir listesi verilmiştir. Fakat bu liste çok ayrıntılı olmayıp, özellikle Büyük Veri teknolojisinin sürekli gelişmesini destekleyen teknolojilere yer verilmiştir.

2.1.1. Büyük Tablo

Büyük Tablo, Google Dosya Sistemi (Google File System-(GFS)) üzerine kurulmuş tescilli dağıtık veritabanı sistemidir. Büyük Tablo'nun temel amacı, web sayfalarının daha hızlı ve başarılı bir şekilde bulunması, depolanması ve güncellenmesidir.

2.1.2. İş Zekâsı

İş Zekâsı (BI), yöneticilerin ve diğer kurumsal kullanıcıların bilgiye dayalı iş kararları almalarını sağlamak için verileri analiz etmek ve uygulanabilir bilgileri sunmak için teknoloji odaklı bir süreçtir. BI, kuruluşların dâhili sistemlerden ve harici kaynaklardan veri toplamasına olanak tanıyan çok çeşitli araçlar, uygulamalar ve metodolojileri kapsamaktadır. BI araçları; verilerin analize hazırlanması, sorgulanması, analitik sonuçların karar mercileri ve operasyonel çalışanlar tarafından kullanılmasını sağlamak için raporlar, gösterge tabloları ve veri görselleştirmeleri yapılmasını sağlamaktadır. BI araçlarının potansiyel faydaları, karar vermeyi hızlandırmak ve geliştirmek, iç iş süreçlerini optimize etmek, operasyonel verimliliği arttırmak, yeni gelir elde etmek ve iş rakiplerine kıyasla rekabet avantajı sağlamaktır[58].

2.1.3. Bulut Bilişim

Bulut Bilişim, Genellikle dağıtılmış bir sistem olarak yapılandırılmış yüksek ölçeklenebilir bilgi işlem kaynakların bir ağ üzerinden bir hizmet olarak temin edildiği bir işlem paradigmasıdır. Bulut, büyük verilerin hem temel altyapısında hem de analitik altyapısında kolaylaştırıcı olarak ortaya çıkmıştır. Bulut, hem genel hem de özel bulut ayarlarında büyük veri analizi için bir dizi seçenek sunmaktadır. Altyapı tarafında, Bulut, çok büyük veri setlerini yönetmek ve bunlara erişmek için seçenekler sunarken aynı zamanda güçlü altyapı unsurlarını nispeten daha düşük maliyetle desteklemektedir.

Bugün Bulut Bilişim sayesinde hard disklerde depolanan veriler internet ortamında sanal sunucularda saklanılabilmektedir. Bulut Bilişim, daha hızlı veri transferi, kıt Bilgi Teknolojisi (BT) kaynaklarının daha etkin kullanılması ve daha hızlı yenilik (inovasyon) kabiliyetine izin vermektedir. İnovasyon düşük maliyetli sanal ortamların dinamik kullanımı ile etkin olup bu talep üzerine şirketleşme (birleşme) olabilmektedir. Özellikle büyük şirketler için iş gücü tasarrufu büyük önem arz etmektedir. Bugün sosyal ağlarda yüklenen video, müzik ve fotoğraf gibi birçok veri o sitelerin bulutlarında depolanmaktadır. Bulut depolama hizmetlerine örnek olarak Dropbox, Google Drive,

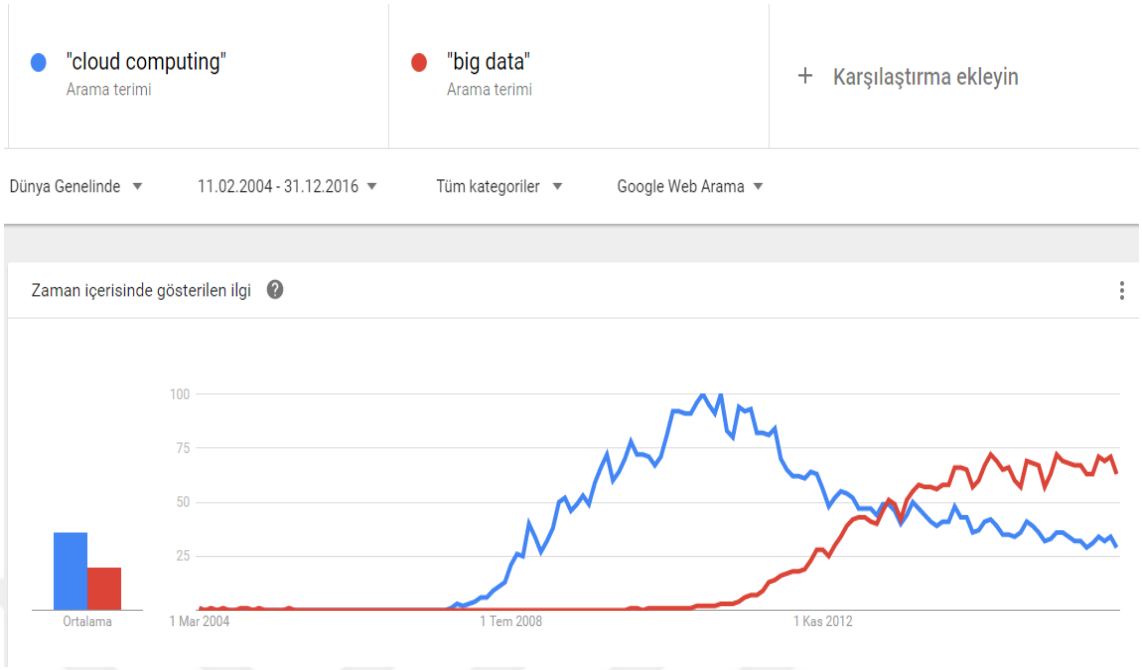
[58] ROUSE Margaret, "business intelligence (BI)", August 2017, http://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI, (20.11.2017)

SkyDrive, iCloud, Yandex. Disk, Turkcell Akıllı Bulut, TTNET Bulut ve Ubuntu One verilebilir[59].

Şekil 2.1: Bulut Bilişim ve Bileşenleri

Bulut özellikle büyük verilerin analizinde çok büyük kolaylıklar sağlamaktadır. Bulut; sanal, uyarlanabilir, esnek ve güçlü yapısı sayesinde büyük verilerin değişen çevreye uygun hâle gelmesini sağlamaktadır (Bkz. Şekil 2.2). Bulut mimarileri, çok büyük veri kümelerinin işlenmesi için ideal olan sanal makine dizilerinden oluşmakta ve bu işlemler sayısız paralel süreçlere bölünebileceği ölçüde gerçekleştirilir. "Küme işlem" adı verilen bu paralel işlem mimarilerinde işlem düğümleri olan sunucular raflarda (racks) depolanmaktadır[60]. Bu da genellikle doğrudan analiz için kullanılabilecek Hadoop kümelerinin geliştirilmesine yol açmıştır.

[59] DEMİR Timur, *"Bulut Bilişim (Cloud Computing) Nedir?"*, 2016, http://www.timurdemir.com.tr/bulut-bilisim-cloud-computing-nedir, (07.01.2016).
[60] GÜRSAKAL Necmi, "Büyük Veri", Genişletilmiş 2. Baskı, Dora, Bursa, ISBN:978 605-4798-803, 2014, syf. 157.



Şekil 2.2: "Cloud Computing" ve "Big Data" Konulu Aramalar

Bulut bilgi işlemi, internet üzerinden paylaşılan bilgi işlem kaynaklarına dayanan bir hesaplama tekniğidir. Bulut ortamında bilgi işlem kaynakları uzaktan barındırılmakta ve bulut bilgi işlem, sunucuları kısıtladığı kadar çok kişi tarafından erişilebilmektedir. Bulut bilgi işlem de her bir bireye veya kuruluşa bir "çalışma alanı" tahsis edilir veya işlevlerine uygun uygulama ve yazılımlara erişim izni verilmektedir[61]. Bulut uzak servisleri kullanıcı verilerinin, yazılımını ve hesaplamalarını yapar. Bununla birlikte, bulut sayesinde sanal makinelerin ve uygun fiyatlı işlemcilerin çok sayıda birleşimi, internet tabanlı şirketlerin büyük ölçekli hesaplama kümelerine ve gelişmiş veri depolama sistemlerine yatırım yapmalarını mümkün hâle getirmiştir[62].

2.1.4. Veri Ambarı

Veri ambarı, verileri raporlamak için optimize edilmiş özelleştirilmiş veritabanıdır. Bu veritabanı genellikle yapılandırılmış büyük miktardaki veriyi depolamak için kullanır. Veriler işlemsel veri depolarındaki ETL (extract, transform, and load) araçları kullanılarak yüklenir ve sonuçlar genellikle iş zekâsı araçları kullanılarak üretilmektedir.

[61] DATAFLOQ, *"How Cloud Computing Affects Individuals and Organizations"*, 30 DECEMBER 2016. https://datafloq.com/read/cloud-computing-affects-individuals-organizations/2559, (10.05.2017).
[62] CHEN C.L. Philip – Chun-Yang ZHANG, ''Data-intensive applications, challenges, techniques and technologies: A survey on Big Data'', Information Sciences Volume 275, 10 August 2014, Pages 314-347

2.1.5. Data Mart

Bir veri ambarı, bir kuruluşun tüm verileri için merkezi bir depodur. Bununla birlikite, bir data mart'ın amacı, insan kaynakları yönetimi gibi organizasyon içindeki belirli bir kullanıcı grubunun belirli taleplerini karşılamaktır. Genel olarak, bir kuruluşun data martları kuruluşun veri ambarının alt kümeleridir[63].

2.1.6. Dağıtık Sistem

Dağıtık sistem, birden fazla bilgisayar, bir ağ üzerinden iletişim kurarak, ortak bir hesaplama problemini çözmek için kullanılır. Problem paralel çalışan bir ya da daha fazla bilgisayar tarafından çözülmekte ve bu bilgisayarların her biri birden fazla görevi gerçekleştirmektedir. Dağıtık sistemlerin avantajları düşük bir maliyetle yüksek performans, yüksek güvenilirlik ve daha fazla ölçeklenebilirliği içermesidir.

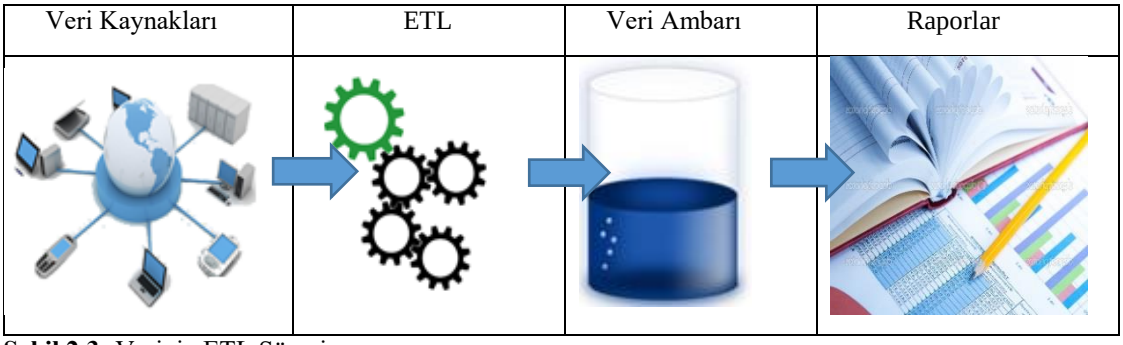
2.1.7. Dinamo (Dynamo)

Dinamo, Amazon tarafından geliştirilen tescilli dağıtık veri depolama sistemidir. Amazon DynamoDB, herhangi bir ölçekte tutarlı, tek basamaklı milisaniyelik gecikmelere ihtiyaç duyan uygulamalar için hızlı ve esnek bir NoSQL veritabanı hizmetidir. Dinamo esnek veri modeli ve güvenilir performansı sayesinde mobil, web, oyun, reklem teknolojisi, Nesnelerin İnterneti ve gerçek zamanlı veri işleme türü de dâhil olmak üzere diğer birçok uygulama için mükemmel bir uyum sağlar.

2.1.8. Ayıklama, Dönüştürme ve Yükleme (ETL)

ETL, dış kaynaklardan gelen verileri ayıklamak, işlemsel ihtiyaçlara uygun olarak onları dönüştürmek ve bir veritabanı veya veri ambarı içine yüklemek için kullanılan yazılım araçlarıdır. Burada ham veriler (normalleştirilmemiş) veri ambarına girmeden önce ETL sürecinden geçerek kullanılacağı şekle göre istenilen formata dönüştürülür. ETL sayesinde veri ambarında depolanan verileri raporlama ve analiz için kullanmak oldukça kolaydır. Şekil 2.3 de verinin ETL sürecindeki aşaması gösterilmiştir.

[63] ROUSE Margaret, *"data mart (datamart)"*, May 2014. http://searchsqlserver.techtarget.com/definition/data-mart, (12.12.2017).



Şekil 2.3: Verinin ETL Süreci

Hadoop Ekosistemi ile ETL transferin de ise Apache Hadoop, basit bir programlama modeli kullanarak ticari bilgisayarların (Ticari bilgisayar, x-86 temelli bir Windows PC gibi kolay satın alınabilen bir dizüstü veya masaüstü bilgisayarlardır.) kümeleri arasında büyük veri kümelerinin dağıtılan işlem için izin veren bir çerçevedir.

2.1.9. Google Dosya Sistemi

Google Dosya Sistemi (GFS), Google tarafından geliştirilen tescilli dağıtık dosya sistemi olup; Hadoop geliştirilirken GFS den esinlenmiştir[64]. GFS'nin amacı, büyük dosyaları depolamak ve bunlara erişimi sağlamaktır. Burada ki büyük dosyalar, sabit sürücüye depolanamayan dosyalardır.

2.1.10. Hadoop Bileşenleri ve Mimarisi

Hadoop, bir makineden başlayarak, yüzlerce makine üzerine dağılabilen büyük veri kümelerini işlemek için kullanılan, Java ile geliştirilmiş (ücretsiz) yazılım çatısıdır. Bu uygulamalarda genellikle Web üzerinde kullanılabilen ve çoğunlukla kullanılan açık uygulama programlama arayüzleri aracılığıyla açık veri kaynaklarından erişilen veriler kullanılır. Hadoop ve Geleneksel RDBMS arasındaki farklar aşağıdaki Tablo 2.1'de gösterilmiştir.

[64] GHEMAWAT Sanjay – Howard GOBIOFF – Shun-Tak LEUNG, *"The Google file system"*, 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003.

Kriterler	Hadoop	RDBMS
Veri Tipleri	Yarı yapılandırılmış ve yapılandırılmamış verileri işler.	Yapısal verileri işler.
Şema	Şema üzerinde okunur.	Şema üzerinde yazılır.
Uygulamalar için En Uygun	Veri keşfi ve Yapılandırılmamış verilerin Büyük Depolama/İşleme.	OLTP ve karmaşık ACID işlemleri için en uygunudur.
Hız	Yazılar Hızlı	Okumalar Hızlı

Tablo 2.1. Hadoop ve Geleneksel RDBMS Arasındaki Farklar

2.1.10. 1. Apache Vakfı Tarafından Tanımlanan Hadoop

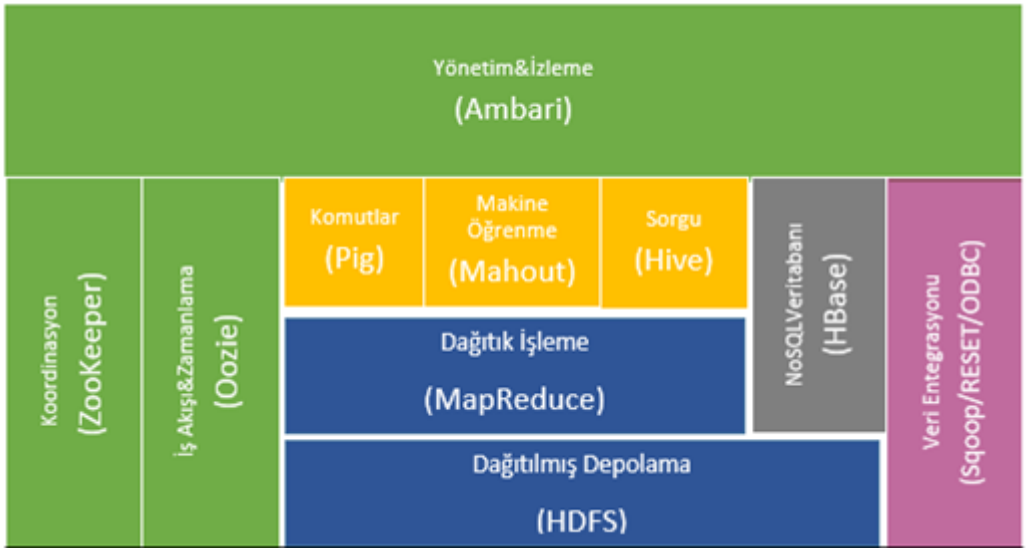
Hadoop, Google'ın EşleİndirgeveGoogle File System'inden esinlenerek geliştirilmiştir. Başlangıçta Yahoo geliştirmiş ve şu an Apache Yazılım Vakfı (Apache Software Foundation) bu sistemi bir proje olarak yönetmektedir. Apache Hadoop yazılım kütüphanesi, basit programlama modelleri kullanarak büyük veri kümelerinin bilgisayar kümeleri arasında dağıtılmasını sağlayan bir çerçevedir. Tekli sunuculardan binlerce makineye ölçeklenmek üzere tasarlanmış olup her biri yerel hesaplama ve depolama imkânı sunmaktadır. Yüksek erişilebilirlik sağlamak için donanıma güvenmek yerine, kütüphane kendisi, başarısızlıkları uygulama katmanında algılamak ve ele almak üzere tasarlanmıştır; bu nedenle, her biri başarısızlıklara eğilimli olabilen bir bilgisayar kümesinin üstünde yüksek oranda mevcut bir hizmet sunmaktadır. Apache Hadoop, anlamlı bilgiler elde etmek için analitikten yararlanmak için büyük miktarda veri kullanıldığında, büyük verileri işlemek için bir çözümdür. Apache Hadoop mimarisi, çeşitli hadoop bileşenleri ve karmaşık iş problemlerini çözmek için muazzam yetenekleri olan farklı teknolojilerin birleşmesinden oluşur.

Hadoop ekosistemindeki tüm bileşenler açık bir şekilde belirginleştirilmiştir**.** Hadoop mimarisinin bütünsel yapısını Hadoop Ekosistemi'ndeki; Hadoop Ortak (Hadoop Common), Hadoop YARN (Yet Another Resource Negotiator), Hadoop Dağıtılmış Dosya Sistemi (Hadoop Distributed File System-(HDFS)) ve Eşleİndirge (MapReduce) elemanları oluşturmaktadır. Bu ana bileşenlerin altında ise başka araçlar bulunmaktadır. Hadoop Ortak, tüm Java kitaplıkları, yardımcı programlar, OS (Operating System) seviyesinde soyutlama, gerekli Java dosyalarını ve Hadoop'u çalıştırmak için komut dosyası sağlarken; Hadoop YARN, iş planlaması ve küme kaynak

yönetimini yapan bir çerçevedir. Hadoop mimarisindeki HDFS, uygulama verisine yüksek verimlilikte erişim sağlar ve Hadoop EşleİndirgeBüyük veri kümelerinin YARN tabanlı paralel işlenmesini sağlar.

2.1.10.2. Apache Spark, Hadoop ekosistemini nasıl geliştirdi?

Verilen iş sorunlarına doğru çözümler üretmek için Hadoop mimarisine ve bileşenlerine derinlemesine girmek gerekir. Şekil 2.4'te Apache Hadoop Ekosistemi gösterilmiştir.



Şekil 2.4: Apache Hadoop Ekosistemi
Kaynak: mssqpltips.com (10.12.2016).

Apache Hadoop Ekosisteminin geliştirdiği, Büyük Veri Ekosisteminin Mimari Bileşenleri aşağıdaki gibi tanımlanmıştır.

2.1.10.3. Hadoop Çekirdek Bileşenleri

Hadoop Ekosistemi; Hadoop Ortak, Hadoop Dağıtılmış Dosya Sistemi (HDFS), EşleİndirgeApache Hadoop'un Dağıtık Veri İşleme Çerçevesi) ve YARN olmak üzere dört temel bileşenden oluşmaktadır.

2.1.10.3.1. Hadoop Ortak

Apache Vakfı, Hadoop ekosistemindeki diğer modüller tarafından kullanılabilen önceden tanımlanmış bir dizi yardımcı program ve kütüphaneye sahiptir. Örneğin, HBase

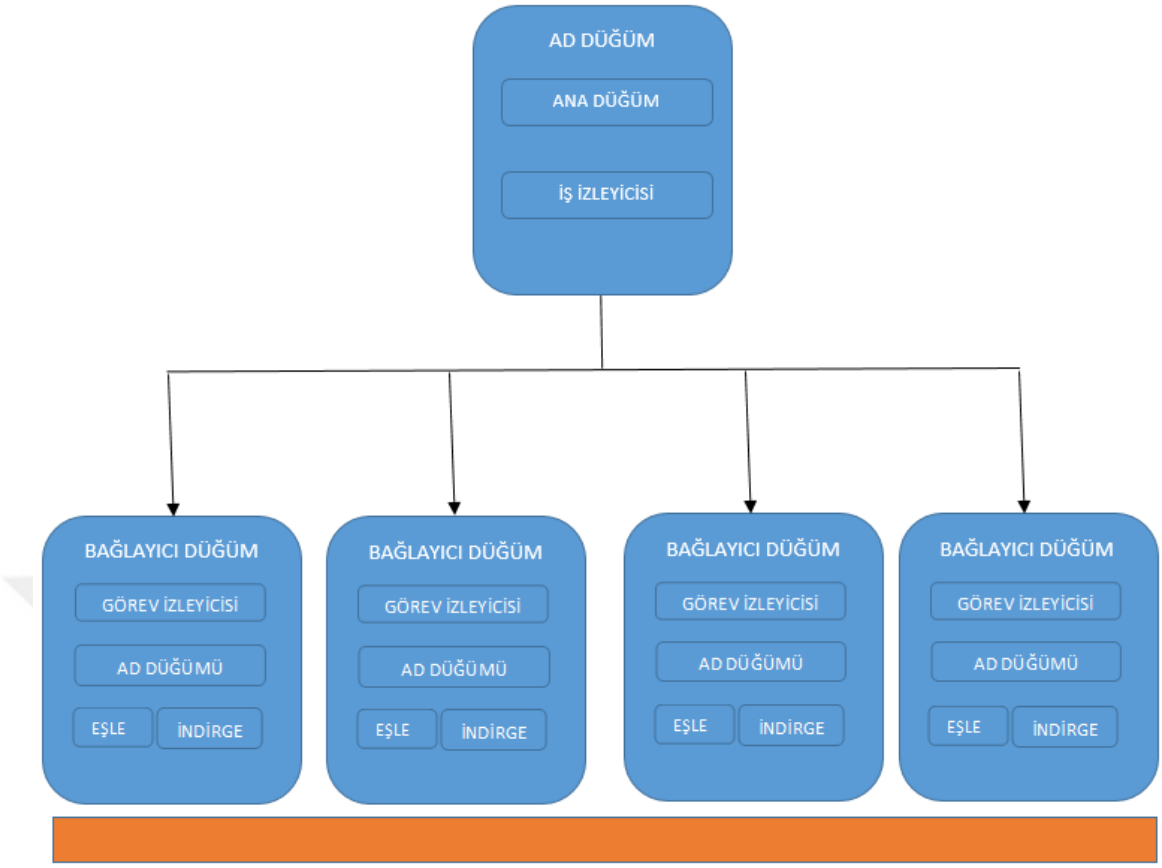
ve Hive, HDFS'ye erişmek istiyorsa, Hadoop Ortak'ta saklanan Java arşivlerini (JAR dosyaları) oluşturmaları gerekmektedir.

2.1.10.3.2. HDFS

HDFS, Google Dosya Sistemi'ne dayanmakta ve güvenilir, hataya dayanıklı küçük bilgisayar makinelerinin büyük kümeleri (binlerce bilgisayar) çalıştırılacak şekilde tasarlanmış bir dağıtılmış dosya sistemidir.

Apache Hadoop için varsayılan büyük veri depolama katmanı HDFS'dir. Kullanıcılar, büyük veri kümelerini HDFS'ye dökebilecekleri için HDFS, Apache Hadoop bileşenlerinin "Gizli Sosu" olarak adlandırılır ve veriler analiz için burada hazır hâle getirilir. HDFS bileşeni, güvenilir ve hızlı veri erişimi için farklı kümeler arasında dağıtılacak veri bloğunun birkaç kopyasını oluşturur. HDFS, Ad Düğümü (NameNode), Veri Düğümü (DataNode) ve İkincil Ad Düğümü (Secondary NameNode) olmak üzere 3 önemli bileşenden oluşmaktadır. HDFS, Ad Düğüm'ün depolama kümesinin kaydını tutmak için ana düğüm görevi gören ve Ad Düğüm'ün bir Hadoop kümesindeki çeşitli sistemlere toplanan bir bağımlı düğüm görevi gören bir Ana/Bağlayıcı Düğümü (Master/Slave Node) (Bkz. Şekil 2.5) mimarisi modelinde çalışır[65].

[65] DeZyre, *"Hadoop Components and Architecture: Big Data and Hadoop Training"*, November 24, 2016, https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114, (20.01.2017).



Şekil 2.5: Hadoop Ana/Bağlayıcı Düğüm Mimarisi
Kaynak: slidehshare.net, (20.10.2016).

Tipik Bir Hadoop dağıtık dosya sisteminde makine rolünü üstlenen İstemci makineleri, Ana düğümleri ve Bağlayıcı düğümleri bulunur. Ana düğümler, Hadoop'u oluşturan iki önemli fonksiyonel parçayı denetlemektedir: çok miktarda veri depolamak (HDFS) ve bu verilerin hepsine paralel hesaplamalar yürütmek (EşleİndirgeAd Düğümü, veri saklama işlevini (HDFS) denetler ve koordine ederken, İş İzleyici (Job Tracker), EşleİndirgeAd Düğümü, veri saklama işlevini (HDFS) denetler ve koordine ederken, İş İzleyici (Job Tracker), EşleİndirgeAd Düğümü, veri saklama işlevini (HDFS) denetler ve koordine ederken, İş İzleyici (Job Tracker), Eşleİndirge'yi kullanarak paralel işlemeyi denetler ve koordine eder. Bağlayıcı Düğümler, makinelerin büyük çoğunluğunu oluşturur ve verileri depolamak ve hesaplamaları çalıştırmak için tüm pis işi yapar. Her bir bağlayıcı düğüm hem ana düğümler ile iletişim kuran ve komutlarını alan bir Veri Düğümü (Data Node) ve Görev İzleyicisi arka plan programını (Task Tracker daemon) çalıştırır. Görev İzleyicisi arka plan programı, İş İzleyicisine, Veri Düğümü arka plan programına bir bağlayıcı olarak Ad Düğümü'ne bağlı olur.

İstemci makinelerde tüm küme ayarlarıyla birlikte Hadoop kuruludur. Ancak, Ana veya Bağlayıcı düğümler kurulu değildir. Bunun yerine, İstemci makinesinin rolü,

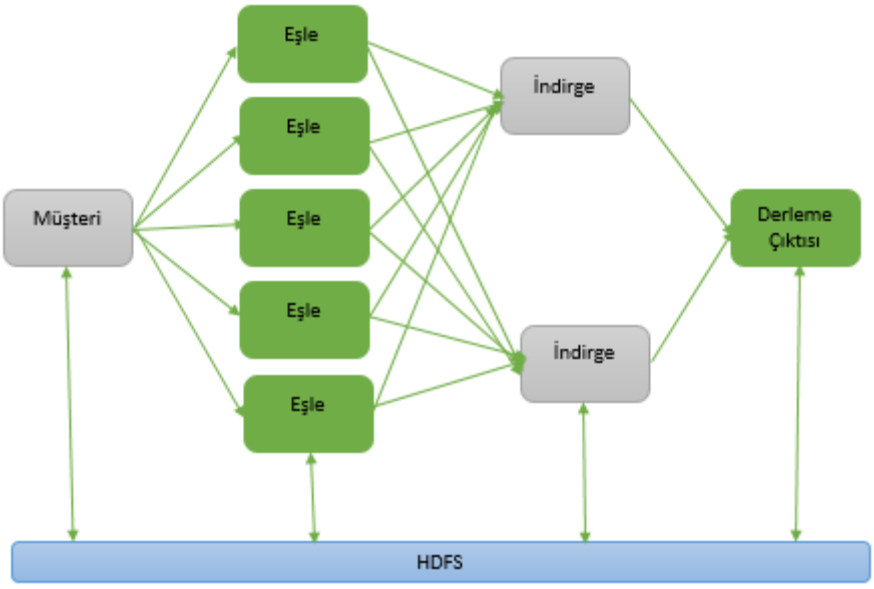
kümeye veri yüklemek, bu verilerin nasıl işleneceğini açıklayan EşleİndirgeEşleİndirge işlerini göndermek ve iş bittiğinde işin sonuçlarını almak ya da görüntülemektir. Yaklaşık 40 düğümden oluşan küçük kümelerde, hem İş İzleyicisi hem de Ad Düğümü gibi birden fazla rol oynayan tek bir fiziksel sunucu olabilmektedir. Bununla birlikte, orta ve büyük kümelerde her rol için tek bir sunucu makinesi gerekmektedir[66].

2.1.10.3.3. EşleİndirgeEşleİndirge (MapReduce)-Apache Hadoop'un Dağıtık Veri İşleme Çerçevesi

EşleİndirgeEşleİndirge, Google tarafından oluşturulan ve HDFS içerisindeki gerçek verilerin verimli bir şekilde işlenmesini sağlayan Java tabanlı bir sistemdir. EşleİndirgeEşleİndirge, büyük bir veri işleme işini küçük görevlere bölerek yapar. EşleİndirgeEşleİndirge, sonuçları bulmak için veriyi küçültmeden önce büyük veri kümelerini paralel olarak analiz eder. Hadoop ekosisteminde, Hadoop EşleİndirgeEşleİndirge, YARN mimarisine dayanan bir çerçevedir. YARN tabanlı Hadoop mimarisi, büyük veri kümelerinin paralel işlenmesini destekler ve EşleİndirgeEşleİndirge, arıza ve hata yönetimini göz önüne alarak, binlerce düğümde kolayca uygulamalar yazmada bir çerçeve sağlar.

EşleİndirgeEşleİndirge'nin arkasındaki temel çalışma prensibi şöyledir: Eşle (Map) ve İndirge (Reduce) birer fonksiyon olup bu fonksiyonlar sayesinde işlenecek veriler birbirinden bağımsız parçalara ayrılır. Ayrılan bu parçaların her biri Eşle'ye anahtar-değer çiftleri şeklinde eşlenerek iletilir. Daha sonra Eşle'den çıkan veriler gruplanıp sıralanarak tekrar İndirge'ye iletilir. İndirge kısmında ise bütün çiftler aynı anahtar değeri ile indirgenir (Bkz. Şekil 2.6). Bu arada, görev giriş ve çıkışları bir dosya sisteminde saklanır. EşleİndirgeEşleİndirge, bu işlerin zamanlamasını yapar, işleri izler ve başarısız olan görevi yeniden gerçekleştirir.

[66] HEDLUNd Brad, *"Understanding Hadoop Cluster and the Network"*, September 10, 2011, http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/, (10.03.2017).



Şekil 2.6: EşleİndirgeÇalışma Prensibi
Kaynak: https://s3.amazonaws.com/files.dezyre.com/Tutorials/HDFS, (20.10.2016).

HDFS dosya sistemi, veri düğümünü oluştururken EşleİndirgeÇerçeve hesaplama düğümünü oluşturmaktadır. Bununla birlikte, tipik Hadoop ekosistem mimarisinde veri düğümü ve hesaplama düğümü aynı kabul edilir.

Skybox, bugün dünyanın herhangi bir yerindeki videoları ve görüntüleri yakalamak için ekonomik bir görüntü uydu sistemi geliştirdi. Skybox, bugün uydulardan indirilen büyük miktardaki görüntü verilerini analiz etmek için Hadoop'u kullanıyor. Skybox'ın görüntü işleme algoritmaları C ++ ile yazılmıştır. Skybox'un tescilli bir çerçevesi olan Busboy, Java tabanlı EşleİndirgeÇerçevesinden yerleşik kod kullanır[67].

2.1.10.3.4. YARN

YARN olarak bilinen Hadoop 2.0, günümüzde dağıtılan büyük verilerin işlenmesi ve yönetilmesi için yaygın olarak kullanılmakta olan, Ekim 2013'te piyasaya sürülen en son teknolojidir. Hadoop YARN, Hadoop veritabanı ve HBase ile birlikte Hadoop Ekosistemi ile bağlantılı tüm teknolojilere fayda sağlayacak performans geliştirmeleri sağlamak üzere Hadoop 1.0'a bir yeniliktir. Hadoop YARN, Hadoop distribitörleri tarafından gönderilen Hadoop 2.x dağıtımlarıyla birlikte gelir. YARN, Hadoop EşleİndirgeÇerçevesini Hadoop Sistemlerinde kullanmak zorunda kalmayan iş planlaması ve

[67] DeZyre, *"Hadoop Components and Architecture: Big Data and Hadoop Training"*, November 24, 2016, https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114, (20.01.2017).

kaynak yönetimi görevlerini yerine getirir. Hadoop YARN, Hadoop 1.0'ın özgün özelliklerinden farklı olarak geliştirilmiş bir mimariye sahiptir. Bu sayede sistemler yeni seviyelere kadar ölçeklenebilir ve Hadoop HDFS'deki çeşitli bileşenlere sorumluluklar açıkça atanabilmektedir[68].

Apache Hadoop'un yukarıda listelenen temel bileşenleri, temel dağıtık Hadoop çerçevesini oluşturmaktadır. Hadoop ekosisteminin ayrılmaz bir parçasını oluşturan birkaç Hadoop bileşeni daha vardır; bunlar, Apache Hadoop'un gücünü bir şekilde arttırmak veya veritabanları ile daha iyi entegrasyon sağlamak, Hadoop'u daha hızlı hâle getirmek veya yeni özellikler ve işlevler geliştirmek için kullanılmaktadır. Şirketler tarafından büyük ölçüde kullanılan önde gelen Hadoop bileşenleri şunlardır: Pig, Hive, Sqoop, Flume, HBase, Oozie ve Zookeeper'dır[69].

2.1.10.4. Hadoop Ekosisteminin Veri Erişim Bileşenleri-Pig ve Hive

Pig: Apache Pig, büyük veri kümelerini etkili ve kolay bir şekilde analiz etmek için Yahoo tarafından geliştirilen kullanışlı bir araçtır. Pig Latin dilini kullanarak optimize edilmiş, genişletilebilir ve kullanımı kolay yüksek düzeyde bir veri akışı sağlar. Pig programlarının en göze çarpan özelliği, yapılarının büyük veri setlerinin işlerliğini kolaylaştıran paralelleştirmeye açık olmasıdır. Örneğin; Bir kişinin kişisel sağlık verileri gizlik açısından büyük önem arz etmektedir ve başkalarıyla paylaşılmamalıdır. Bu bilgiler gizliliğin korunması için saklanmalıdır, ancak sağlık verileri çok büyük olduğundan kişisel sağlık verilerinin tanımlanarak çıkarılması önemlidir. Bugün Apache Pig, sağlık bilgilerini tanımak için kullanılmaktadır.

Hive: Hive, Facebook tarafından geliştirilmiş, Hadoop'un üzerine kurulmuş bir veri ambarıdır. Hive, yapılandırılmamış büyük veri setlerinin yönetilmesi, sorgulanması, özetlenmesi ve analizi için SQL'e benzer HiveQL olarak bilinen basit bir dil kullanır. Hive, sorgulamayı dizine ekleme yoluyla daha hızlı hâle getirir. Örneğin; Hive, Ad-hoc analiz, raporlama ve makine öğrenimi için Facebook'ta Hadoop'u günlük olarak 7500'den fazla Hive işlerinin yürüterek basitleştirir.

---

[68] DeZyre, *"Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for Hadoop Users"*, 25 November 2014, https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84, (10.02.2017).

[69] DeZyre, *"Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for Hadoop Users"*, 25 November 2014, https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84, (10.02.2017).

2.1.10.5. Hadoop Ekosisteminin Veri Bütünleştirme Bileşenleri - Sqoop ve Flume

Sqoop: Sqoop bileşeni, harici kaynaklardan HDFS, HBase veya Hive gibi ilgili Hadoop bileşenlerine veri aktarmak için kullanılır. Sqoop ayrıca, diğer harici yapılandırılmış veri kaynaklarında Hadoop'tan veri dışa aktarmak için de kullanılabilir. Sqoop veri aktarımını paralel hâle getirir, aşırı yükleri hafifletir, veriyi içe aktarmaya izin verir, etkin veri analizi yaparak verileri hızlı bir şekilde kopyalar. Bugün Çevrimiçi Pazarlamacı Coupons.com, Hadoop ve IBM Netezza veri ambarı arasındaki verilerin iletilmesini sağlamak için Hadoop ekosisteminin Sqoop bileşenini kullanır ve sonuçları Sqoop kullanarak Hadoop'a geri gönderir[70].

Flume: Flume bileşeni, büyük miktarda veriyi toplamak ve entegre etmek için kullanılır. Apache Flume, kaynaktan veri toplamak ve onu depolama yerine (HDFS) geri göndermek için kullanılır. Flume bunu, üç ana yapı kanalından, kaynaklardan ve lavabo maddelerinden oluşan veri akışlarını özetleyerek gerçekleştirir. Veri akışını flume ile çalıştıran süreçlere aracılar denir ve flume yoluyla akan veri parçacıkları olaylar olarak adlandırılmaktadır. Örneğin; Twitter kaynağı akış API'si aracılığıyla bağlanır ve sürekli olarak tweet'leri indirir (olaylar olarak adlandırılır). Bu tweet'ler JSON formatına dönüştürülür ve Twitter'da kullanıcıların ilgisini çekmek için tweet'lerin ve retweet'lerin daha ayrıntılı analizi için aşağı doğru Flume lavabolarına gönderilir.

2.1.10.6. Hadoop Ekosistemi Veri Depolama Bileşeni –Hbase

HBase: HBase, başlangıçta Powerset tarafından geliştirilmiş ve şu an da Hadoop'un bir parçası olarak Apache Yazılım temelli bir proje olarak yönetilmektedir. HBase, temel verilerin depolanması için HDFS kullanan sütuna yönelik bir veritabanıdır. HBase, EşleİndirgeEşleİndirge kullanarak rasgele okumaları ve toplu hesaplamaları desteklemektedir. HBase ile NoSQL veritabanı kuruluşu, donanım makinesinde milyonlarca satır ve sütun içeren geniş matrisler oluşturulabilir. HBase'i kullanmak için en iyi uygulama, büyük veri kümelerini rastgele "okuma veya yazma" erişimi için gereksinim olduğunda ortaya çıkmaktadır. Facebook, 2010 yılında HBase'in üzerine kurulmuş mesajlaşma platformuyla HBase'in en büyük kullanıcılarından biridir. Facebook HBase'i

[70] DeZyre, *"Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for Hadoop Users"*, 25 November 2014, https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84, (10.02.2017).

veritabanında veri analizi, dâhili izleme sistemi, Yakın Kişiler Özelliği, Arama Dizinleme ve dâhili veri ambarları için veri kazımasında (silinmesinde) kullanmaktadır.

2.1.10.7. Hadoop Ekosisteminin İzleme, Yönetim ve Orkestrasyon Bileşenleri-Oozie ve Zookeeper

Oozie: Oozie, iş akışlarının Yönlendirilmiş Asalik Grafikler (Directed Acyclic Graphs) olarak ifade edildiği bir iş akışı zamanlayıcısıdır. Oozie, bir Java servlet içerisindeki Tomcat'de çalışır ve Hadoop işlerini (MapReduce, Sqoop, Pig ve Hive) yönetmek için çalışan tüm iş akış örneklerini, devlet reklam değişkenlerini ve iş akış tanımlarını depolamak için bir veritabanı kullanır. Oozie'deki iş akışları veri ve zamana dayalı olarak yürütülür. Örneğin; Amerikan video oyun yayıncısı Riot Games, oyuncu deneyimini anlamak için Hadoop'u ve açık kaynak kodlu Oozie'yi kullanmaktadır[71].

Zookeeper: Zookeeper koordinasyonun kralıdır ve bir Hadoop kümesi için basit, hızlı, güvenilir ve düzenli operasyonel servisler sunmaktadır. Zookeeper senkronizasyon servisi, dağıtılan yapılandırma servisinden ve dağıtılmış sistemler için bir ad kayıt defteri sağlamaktan sorumludur. Elastic tarafından bulunan Zookeeper, kaynak tahsisi, lider seçimi, yüksek öncelikli bildirimler ve keşif için kullanılır.

Diğer yaygın Hadoop ekosistem bileşenleri arasında Avro, Cassandra, Chukwa, Mahout, HCatalog, Ambari ve Hama bulunur. Kullanıcılar, bir veya daha fazla Hadoop ekosistem bileşenini kullanarak Hadoop'a uygulayarak, değişen iş gereksinimlerini karşılamak için büyük veri deneyimlerini kişiselleştirebilirler.

Cassandra: Cassandra, bir açık kaynak olup dağıtılmış bir sistem üzerinde büyük miktarlardaki veriyi işlemek için tasarlanmış (ücretsiz) veritabanı yönetim sistemidir. Bu sistemi ilk olarak Facebook geliştirmiştir ve şu anda Apache Yazılım temelli bir proje olarak yönetilmektedir.

Ambari: Bir Hadoop bileşeni olan Ambari, Hadoop yönetimi için web kullanıcı arayüzü kullanımı kolay bir RESTful API'dir. Ambari, Hadoop ekosistem hizmetlerini yüklemek için adım adım bir sihirbaz sağlar. Ambari, Hadoop hizmetlerini başlatmak, durdurmak ve yeniden yapılandırmak için merkezi yönetimle donatılmıştır ve Hadoop kümesinin sağlık durumunu izleyebilen metrik toplama ve uyarı çerçevesini kolaylaştırmaktadır.

[71] DeZyre, *"Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for Hadoop Users"*, 25 November 2014, https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84, (10.02.2017).

Ambari'nin son sürümünde Apache Spark Servisleri için hizmet kontrolü eklenmiş ve Spark 1.6'yı desteklemektedir.

Mahout: Mahout, makine öğrenimi için önemli bir Hadoop bileşenidir ve çeşitli makine öğrenme algoritmalarının uygulanmasını sağlar. Bu Hadoop bileşeni, kullanıcı davranışlarına öneriler sunmada, öğeleri ilgili gruba kategorize ederek sınıflandırmaya dayalı olarak kümelere ayırır. Mahout'un algoritmaları Hadoop'un üzerine yazılmıştır, bu yüzden dağıtılmış ortamda iyi çalışmaktadır. Mahout bulutta etkili bir şekilde ölçeklemek için Apache Hadoop kitaplığını kullanır. Mahout, kodlayıcıya büyük miktarda veri üzerinde veri madenciliği görevleri yapmak için hazır bir çerçeve sunarak büyük veri kümelerinin etkili ve hızlı bir şekilde analiz edilmesini sağlar. Ayrıca Mahout; k-means, fuzy (bulanık) k-means, Canopy, Dirichlet ve Mean-Shift gibi birkaç EşleİndirgeEşleİndirge etkin kümeleme uygulaması ile matris ve vektör kütüphanelerini içermektedir. Bugün; Adobe, Facebook, LinkedIn, Foursquare, Twitter ve Yahoo gibi şirketler Mahout'u yoğun olarak kullanmaktadır. Örneğin; Foursquare belirli bir alanda mevcut yerleri, yiyecek ve eğlence bulma konusunda Mahout'tan yararlanmaktadır.

Kafka: LinkedIn tarafından geliştirilmiş ve daha sonra 2011'de açık kaynaklı bir Apache projesi hâline gelmiştir. Apache Kafa, 2012'de ise birinci sınıf bir Apache projesi olmuştur. Apache Kafka, Scala ve Java dilllerinde yazılmış ve yayın abone temelli hataya dayanıklı mesajlaşma sistemidir. Buradaki mesajlar, hızlı, ölçeklenebilir ve tasarım yoluyla dağıtılır[72]. Bugün FourSquare Kafka'yı çevrimiçi ve çevrimdışı mesajlaşmayı gerçekleştirmek için kullanıyor.

Hama: EşleİndirgeEşleİndirge'nin ötesinde gelişmiş analizler yapmaya izin veren Apache üst düzey açık kaynak projesidir. Apache Hama, Hadoop'un üstünde genel amaçlı Toplu Eşzamanlı Paralel (Bulk Synchronous Parallel-(BSP)) bilgi işlem motorudur. Bununla birlikte, Hama, kapsamlı bilimsel ve yinelemeli algoritmalar için paralel bir işleme çerçevesi sağlamaktadır[73].

---

[72] Tuturialspoint, *"Apache Kafka Tutorials"*, 2017, https://www.tutorialspoint.com/apache_kafka/index.htm, (12.01.2017).

[73] AKHTAR Nihat–Firoj PARWEJ–Yusuf PERWEJ, *"A Perusal of Big Data Classification and Hadoop Technology"*, International Transaction of Electrical and Computer Engineers System, Vol. 4, No. 1, 2017, 26-38.

2.1.11. Spark

Spark; hız, kullanım kolaylığı ve sofistike analitik üzerine kurulmuş açık kaynaklı bir büyük veri işleme çerçevesidir. Başlangıçta 2009 yılında UC Berkeley'nin AMPLab'da geliştirilmiş ve 2010 yılında açık kaynaklı bir Apache projesi olarak hazırlanmıştır. Apache Spark, piyasaya sürülmesinden bu yana geniş çaplı endüstrilerdeki işletmeler tarafından hızla benimsenmiştir. Netflix, Yahoo ve eBay gibi İnternet santralleri, toplu olarak 8000'den fazla düğüm kümeleri üzerinde birden fazla petabayt veri işleyen Spark'ı büyük çapta kullanıma açmıştır. Spark 250'den fazla şirketin 1000'in üzerinde katkıda bulunanların, büyük veri alanındaki en büyük açık kaynak topluluğu hâline gelmiştir[74].

Spark, hızlı hesaplama için tasarlanmış yıldırım hızlı küme bilgi işlem teknolojisidir. Spark, Hadoop ve Storm gibi diğer büyük verilere ve EşleİndirgeTeknolojilerine kıyasla birçok avantaja sahiptir. Her şeyden önce Spark, doğada çok çeşitli veri setleri (metin verileri, grafik verileri vb.) ve veri kaynağına ulaşıp kullanmayı sağlar[75]. Spark'ın temel özelliği, bir uygulamanın işlem hızını arttıran bellek içi küme işlemidir. Spark, toplu iş uygulamaları, yinelemeli algoritmalar, etkileşimli sorgular ve akış gibi çok çeşitli iş yüklerini kapsayacak şekilde tasarlanmıştır. Spark tüm bu iş yükünü ilgili bir sistemde desteklemenin yanı sıra, ayrı araçları muhaza ederek yönetim yükünü de azaltmaktadır. Spark aşağıdaki özelliklere sahiptir[76].

1) **Hız:** Spark Hadoop kümesinde bir uygulamayı çalıştırmaya yardımcı olmaktadır. Spark, Hadoop kümelerindeki uygulamaları bellekte 100 kat daha hızlı ve disk üzerinde çalışırken bile 10 kat daha hızlı çalıştırmayı sağlar[77]. Bu sayede, diske okuma/yazma işlemlerinin sayısı azalmaktadır[78].
2) **Birden çok dili destekler:** Spark; Java, Scala veya Python'da hızlı bir şekilde uygulamalar yazmayı sağlamaktadır. Spark 80'den fazla üst düzey operatörden

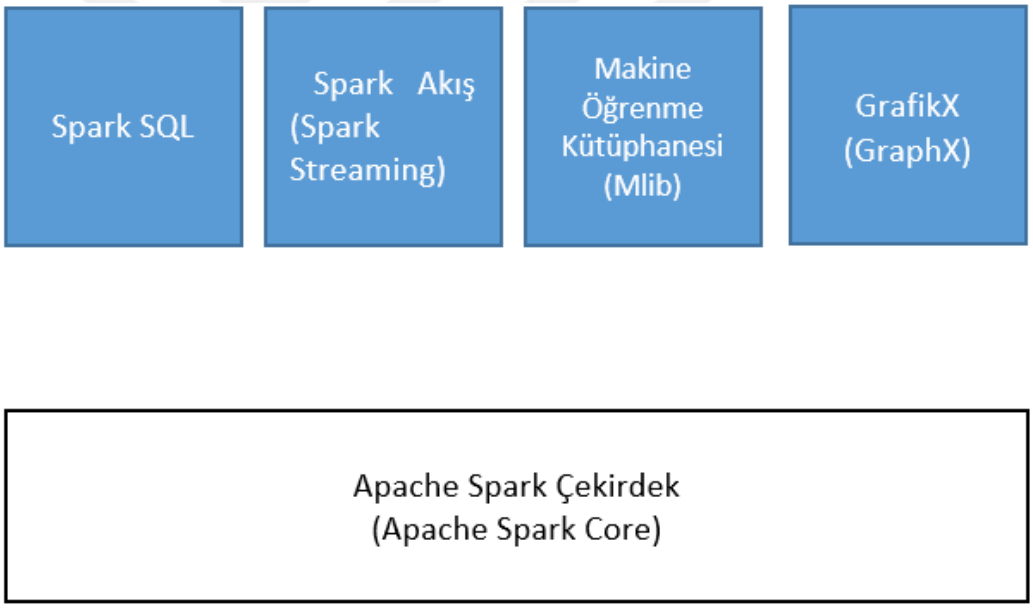
[74] Databricks, "What is Apache Spark™?", 2016, https://databricks.com/spark/about/, (01.02.2016).
[75] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).
[76] Apache Spark – Tutorial, "Apache Spark – Introduction", 2016, https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm, (01.02.2016).
[77] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).
[78] Databricks, "What is Apache Spark™?", 2016, https://databricks.com/spark/about/, (01.02.2016).

oluşan dâhili bir küme ile birlikte gelir. Bu nedenle Spark, kabuk (shell) içindeki verileri sorgulamak için etkileşimli olarak kullanılabilir[79].

3) **Gelişmiş Analitik:** Spark sadece 'Eşle' ve 'İndirge'u desteklemekle kalmaz. Aynı zamanda SQL sorguları, akış verileri, makine öğrenme ve grafik algoritmalarını da desteklemektedir. Spark geliştiricileri, bu özellikleri tek başlarına kullanabilir veya tek bir veri hattı kullanım örneğinde çalıştırmak için birleştirebilirler[80]

2.1.11.1. Spark Bileşenleri

Spark Çekirdek, API dışında Spark ekosisteminin bir parçası olan ve Büyük Veri analitiği ve Makine Öğrenimi alanlarında ek özellikler sağlayan ek kütüphaneler bulunmaktadır[81]. Aşağıdaki şekil Apache Spark ekosistemindeki bu farklı kütüphanelerin birbiri ile nasıl ilişkili olduğunu göstermektedir.



Şekil 2.7: Spark Bileşenleri
Kaynak: https://www.tutorialspoint.com/apache_spark/images/components_of_spark.jpg, (01.02.2016).

[79] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).
[80] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).
[81] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

Apache Spark Çekirdek, tüm diğer işlevlerin üzerine inşa edilmiş olup Spark platformunun altında yatan genel yürütme altyapısıdır. Apache Spark Çekirdek, harici depolama sistemlerinde bellek içi hesaplama ve referans veri kümeleri sağlar.

Spark SQL, Spark veri kümelerini JDBC API'si üzerinden sunma ve geleneksel İş Zekâsı ve görselleştirme araçları kullanarak Spark verilerinde SQL benzeri sorgular yapma imkânı sağlamaktadır. Ayrıca Spark SQL, kullanıcıların verilerini şu an içinde bulunduğu farklı biçimlerden (JSON, Parquet, bir Veritabanı gibi) ETL'lerine dönüştürmesine ve geçici sorgulama için sunmalarına imkân sağlamaktadır[82].

Spark Akış, gerçek zamanlı akan verileri işleyerek bu verileri kullanılabilir hale getirmektedir. Bu gerçek zamanlı veri işleme süreci mikro toplu işlem ve işleme biçimine dayanmaktadır. Spark Akış gerçek zamanlı verileri işlemek için temel olarak bir dizi Esnek Dağıtılmış Veri Kümeleri (Resilient Distributed Datasets-RDD) olan DStream'i kullanır[83].

Makine Öğrenme Kütüphanesi (MLlib), sınıflandırma, regresyon, kümeleme, ortak filtreleme, boyut indirgeme ve temel optimizasyon ilkelerini içeren ortak öğrenme algoritmaları ve araçlarından oluşan Spark'ın ölçeklenebilir makine öğrenme kitaplığıdır[84]. Spark MLlib, Apache Mahout'un Hadoop disk tabanlı sürümünden dokuz kat daha hızlıdır (Mahout Spark arayüzünü kazanmadan önce).

GrafikX, grafikler ve grafik paralel hesaplama için yeni (alfa) Spark API'sıdır. GrafikX, yüksek düzeyde, Esnek Dağıtılmış Özellik Grafiğini (Resilient Distributed Property Graph) sunarak Spark RDD'yi genişletmektedir. Esnek Dağtılmış Özellik Grafiği her köşe ve kenara ilişkin özellikler içeren yönlendirilmiş bir çoklu grafiktir. Grafik hesaplamayı desteklemek için GrafikX, Pregel API'sinin optimize edilmiş bir varyantının yanı sıra bir dizi temel işleç (örneğin, subgraph, joinVertices ve aggregateMessages) sunmaktadır. Bununla birlikte, GrafikX, grafik analitiği görevlerini basitleştirmek için artan bir grafik algoritması ve geliştiricisi koleksiyonuna sahiptir[85].

---

[82] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

[83] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

[84]PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

[85] PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

2.1.12. Storm

Storm (Gerçek Zamanlı Akış İşlemci), büyük verilerin gerçek zamanlı akışını işlemek için tasarlanmış teknolojilerdir. Apache Storm Hadoop ile gerçek zamanlı olarak verileri işleme imkânı sağlayan dağıtılmış, hataya dayanıklı ve açık kaynaklı bir sistemdir. Akış işlemcisi; finansal hizmetlerdeki algoritmik işlem (alım satım), RFID (Radyo Frekanslı Tanımlama) durum işleme uygulamaları, dolandırıcılık tespiti, süreç izleme ve telekomünikasyonda ki konuma dayalı hizmetler gibi uygulamalar sağlar.

2.1.13. Mashup

Mashup, yeni hizmetler oluşturmak için iki veya daha fazla kaynaktan veri sunumu ya da veri işlevselliğini kullanan ve birleştiren bir uygulamadır. Bu uygulamalar genellikle Web üzerinden kullanılabilir ve çoğunlukla açık uygulama programlama arayüzleri aracılığıyla veya veri kaynaklarından erişilebilen verileri kullanır.

2.1.14. Metaveri

Metaveri, veri dosyalarının içeriğini ve bağlamını (kaynağını) tanımlayan verilerdir. Örneğin dijital fotoğraf makinesi ile çekilen fotoğraflarda, fotoğraf dosyası içerisine kaydedilen; fotoğrafın çekildiği tarih, fotoğrafın yatay ve düşey piksel boyut, fotoğrafın yatay-düşey çözünürlüğü, fotoğrafın odak uzaklığı, fotoğraf makinesinin markası ve modeli ile fotoğrafın çekildiği yerin GPS koordinatları gibi bilgiler birer metaveridir[86].

2.1.15. İlişkisel Veritabanı

İlişkisel Veritabanı (Relational Database-RDB); tablolar, kayıtlar ve sütunlar tarafından düzenlenen çoklu veri setlerinin bir kollektif kümesidir. RDB'ler veritabanı tabloları arasında iyi tanımlanmış bir ilişki kurmaktadırlar. İlişkisel veritabanı yönetim sistemlerin de (Relational Database Management Systems-(RDBMS)) yapılandırılmış veri türü depolanır. SQL, ilişkisel veritabanlarını yönetmek için en yaygın kullanılan dildir.

[86] Wikipedia, *"Extensible Metadata Platform"*, 2016, https://en.wikipedia.org/wiki/Exif (20.07.2016).

2.1.16. İlişkisel Olmayan Veritabanı

İlişkisel olmayan bir veritabanı, ilişkisel veritabanı yönetim sistemleri (RDBMS) tarafından desteklenen tablo/anahtar modelini içermeyen bir veritabanıdır. Bu tür veritabanları, büyük şirketlerin karşılaştığı büyük veri sorunlarına çözüm üretmek için tasarlanmış veri işleme tekniklerini ve süreçlerini gerektirir. En popüler gelişmekte olan ilişkisel olmayan veritabanına NoSQL (Sadece SQL Değil) adı verilir.

2.1.17. Yapılandırılmış Veri

Yapılandırılmış veriler, satırlar, sütunlar veya çok boyutlu matrislerden oluşan grafikler ve tablolar şeklinde düzenlenebilen verilerdir[87]. Bunlar ilişkisel veritabanları ve elektronik tabloların bulunduğu verileri içermektedir. Yapılandırılmış veriler, kolayca saklanılabilecek, sorgulanabilecek ve analiz edilebilecek şekilde işlenir. Bu veriler, genellikle Yapılandırılmış Sorgu Dili (Structured Query Language-(SQL)) kullanılarak yönetilir. SQL sayesinde programlama dili yönetilmekte ve ilişkisel veritabanı yönetim sistemlerindeki veriler sorgulanmaktadır. Yapılandırılmış veri türü; sayısal, para birimi, alfabetik, isim, fatura, sınav sonuçları, e-devlet verileri, tarih veya adres vb. olabilmektedir. Yapılandırılmış veriler yarı yapılandırılmış veri ve yapılandırılmamış verilerin tam tersidir.

2.1.18. Yarı Yapılandırılmış Veri

Yarı yapılandırılmış veriler, anlamsal etiketler içeren, ancak tipik ilişkisel veritabanlarıyla ilişkili yapıya uymayan bir veri türüdür[88]. Bu veriler yapılandırılmış ancak, bir tablo veya nesne tabanlı grafik gibi rasyonel modelde organizeli (düzenlenmiş) değildir. Veri entegrasyonun da özellikle yarı yapılandırılmış veri kullanılmaktadır. Yarı yapılandırılmış veri modeli yapısı, bazı esneklikler sağlayan heterojen verileri temsil etmek için ilişkisel veri modeline alternatif olarak önerilmiştir. Eğer veriler belli bir yapıya sahip ve bütün bilgiler aynı yapıya sahip değilse, bu tür veriler yarı yapılandırılmıştır. Son zamanlarda, yarı yapılandırılmış veri olarak XML daha yaygın

[87] MARR Bernard, *"Big Data Terminology: 16 Key Concepts Everyone Should Understand (Part II)"*, 17 May 2017, http://data-informed.com/big-data-terminology-16-key-concepts-everyone-should-understand-part-ii/, (10.06.2017).

[88] ROBB Drew, *"Semi-Structured Data"*, July 3, 2017, https://www.datamation.com/big-data/semi-structured-data.html, (20.11.2017).

olmuştur. XML (eXtended Markup Language) yarı yapılandırılmış verileri temsil için standart bir dil olarak ortaya çıkmıştır. XML web sayfalarındaki veri yapısı ve anlamına yönelik daha fazla veri sağlayabilmektedir. Yarı yapılandırılmış veri örnekleri olarak XML, HTML-etiketli metin, mp3, video, doktor reçetesi verilebilir. Çünkü bu veriler her zaman yazar, konu, özet vb. yapılandırılmış veriler ile ilişkilendirilirler.

2.1.19. Yapılandırılmamış Veri

Yapılandırılmamış veriler, geleneksel grafik veya tablolara kolaylıkla yerleştirilemeyen verilerdir[89]. Klasik istatistikte satır ve sütunlara yerleştirilmiş olan yapılandırılmış veriler kullanılmaktadır. Örneğin, müşteriler satırlarda bulunurken müşterilerin aldıkları ürünler ise sütunlarda bulunmaktadır. Fakat bugün artık e-postalar aracılığıyla üretilen verilerin büyük çoğunluğu metin biçiminde olup yapılandırılmamıştır. Ayrıca resimler, ses dosyaları, videolar, PDF, HTLM ve web günlükleri yapılandırılmış veriler olmayıp karar vermede kullanılabilmektedir. Bu tür veriler "etiket" (#tag) sözcüğü ile özetlenmektedir. Bugün verilerin %80 ile %90' arasındaki bir oranı yarı yapılandırılmış ve yapılandırılmamış verilerden oluşmaktadır[90]. Oysaki bugün hala istatstikte yoğun olarak kullanılan SPSS, SAS, STATISTICA ve MINITAB gibi istatistik programlarında analiz yapabilmek için verilerin nümerik formatta olması gerekmektedir. Durum böyle olunca yeni nesil istatistikçilerin bu yapıdaki veriler için yeni veri setleri oluşturmaları gerekmektedir[91].

2.1.20. SQL

SQL, Yapılandırılmış Sorgu Dili anlamına gelmektedir. SQL, verinin yönetilmesi ve tasarlanması için kullanılan bir veritabanı yönetim sistemidir. SQL müşterilerin; depolayarak, işleyerek ve ilişkisel veri tabanlarında depolanan veriyi almak için veritabanı sunucuları ile iletişim kurmak için kullandıkları dildir. Örneğin ilişkisel veritabanı yönetim sistemleri (RDBMS); Oracle, Mssql, IBM DB2, Microsoft SQL

[89]MARR Bernard, *"Big Data Terminology: 16 Key Concepts Everyone Should Understand (Part II)"*, 17 May 2017, http://data-informed.com/big-data-terminology-16-key-concepts-everyone-should-understand-part-ii/, (10.06.2017).
[90] NARİN Bilge, *"Big Data / Büyük Veri"*, 25 Mar 2015, http://es.slideshare.net/BilgeNarin1/big-data-24-mart-2015 (29.09.2016).
[91] NARİN Bilge, *"Big Data"*, 24 Mart 2015, http://es.slideshare.net/BilgeNarin1/big-data-24-mart-2015, (29.09.2016).

Server, MySQL, Microsoft Access, IBM Informix, Sybase, Firebird ve PostgreSQL, dil olarak SQL kullanmakta ve bunların içlerindeki veri tablo şeklinde yazılmıştır.

2.1.21. NoSQL

NoSQL, ilişkisel veritabanı yönetim sistemlerine (RDBMS) bir alternatif olarak ortaya çıkmıştır. NoSQL, internetteki artan veriyi depolayabilmek ve hızlı veri akışına sahip sistemlerin ihtiyaçlarını karşılamak için yatay ölçeklemeye başvuran sistemlerdir[92]. Bu özellik her gün terabaytlarca veriyi işleyen Facebook, Google ve Amazon gibi büyük firmaların NoSQL veri tabanlarını tercih etmelerinde etkin rol oynamıştır. Bunlar aynı anda birden fazla sunucu ile birlikte çalışabilmekte ve çok büyük ve karmaşık veriler üzerinde işlemler yapabilmektedir. Bu yönüyle bu veri tabanları veri seli ile mücadele de kuruluşlar için önemli bir araç olarak ortaya çıkmıştır. NoSQL veri tabanları SQL dilini kullanmadıkları için bunlara "Not Only SQL" adı verilmiştir. NoSQL veritabanı yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış bütün verileri çok hızlı bir şekilde özümseyebilmekte ve yüksek performanslı sorgulama kapasitesi sunabilmektedir. NoSQL veri tabanlarına örnek olarak; Cassandra, HBase, Oracle NoSQL, MongoDB, memsql, Neo4j ve nuodb gibi araçlar verilebilir. Bu veri tabanlarının her birinin kendine özgü mimarileri bulunmaktadır[93]. Örneğin Cassandra yatay ölçeklenebilme özelliği sayesinde kümeye (cluster) yeni sunucular eklenmesine olanak sağlayarak kapasitenin artmasına izin verir. Ayrıca Cassandra doğrusala yakın ölçeklendirme sayesinde yüksek performansın artmasını sağlar.

2.1.22. Veri Bilimi'nde Python ve R Dilinin Önemi

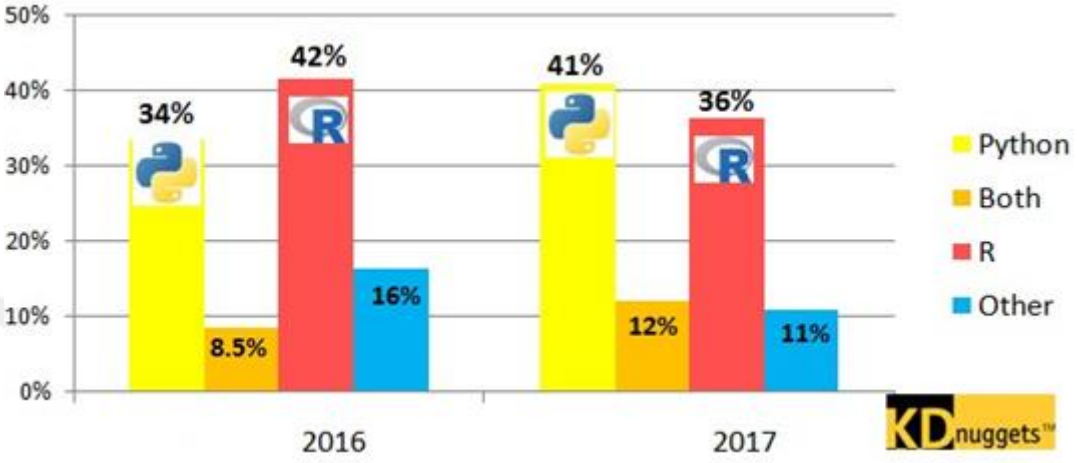
Python, genel amaçlı bir web programlama dili olarak popülerken, özellikle istatistiksel hesaplama için geliştirilen R, veri görselleştirme için mükemmel özellikleriyle popülerdir. Python ve R dili, veri bilimin de istatistiksel hesaplamalar için kullanılan temel araçların başında gelmektedir. Veri bilimciler arasında R veya Phyton'dan hangisinin daha iyi olduğu konusunda tartışmalar hala sürmektedir. Ancak

[92] İLTER Hakan – Hakan SARIBIYIK – Emre YAZICI – Erkan ÜLGEY – Yasemin Kaya – İmran KOCABIYIK – Emre Ekış – Hüseyin Babal – Harun YARDIMCI – Cevat UZUN –Ahmet ARSLAN – Ayhan DEMİRCİ – Erdem EĞAOĞLU – Ümit ÜNAL – Arif BALIK, *"NoSQL"*, 2012, http://devveri.com/nosql-nedir (25.09.2016).

[93] STROUD Forrest, *"NoSQL Database"*, 2017, Nosql Database: http://www.webopedia.com/TERM/N/nosql_databese.html, (10.03.2017).

her iki programlama dili de birbirlerinin tamamlayıcı nitelikte anahtar özelliklerine sahiptirler.

Aşağıdaki Şekil 2.8'de 2016 ve 2017 yıllarında KDnuggets kullanıcılarının Analitik, Veri Bilimi ve Makine Öğrenmesi için kullandıkları programların yüzde değerleri verilmiştir.



Şekil 2.8: 2016 ve 2017 yıllarında Python, R, Her İkisi ve Diğer Platformlar için Analitik, Veri Bilimi ve Makine Öğrenmesi Kullanılma Yüzdesi
Kaynak: KDnuggets, *"Python overtakes R, becames the leader in Data Science, Machine Learning platforms"*, Aug 2017.

Şekil 2.8'deki grafik incelendiğinde, 2016'da Python %34 ile 2. sırada yer alırken R'nin %42 ile 1. sırada; 2017'de bu oranın Python için %41, R için %36 olduğu görülmektedir. Ayrıca, hem R'yi hem de Python'u önemli oranlarda kullanan KDnuggets okuyucularının payı da 2017'de %8,5'den %12'ye yükselirken, diğer araçları ağırlıklı olarak kullanma oranları %16'dan %11'e düştüğü görülmektedir.

2.1.22.1. Python Dili ile Veri Bilimi

Veri bilimi; istatistik hesaplama, tahmin modelleri oluşturma, verilere erişme ve manipüle etme, açıklayıcı modeller oluşturma, veri görselleştirmeleri yapma, modelleri üretim sistemlerine entegre etme ve verilere ilişkin çok daha fazlası gibi birkaç birbiriyle ilişkili ancak farklı faaliyetlerden oluşmaktadır. Python programlama, veri bilimcilerine, tüm bu işlemleri veri üzerinde gerçekleştirmelerine yardımcı olan bir dizi kütüphane sağlamaktadır.

Python, sözdizimi basitliği ve farklı ekosistemlerde çalışabilmesinden dolayı geniş bir popülerlik kazanmıştır. Bu nedenle Python veri bilimi için genel amaçlı çok paradigmalı programlama dilidir. Python programlama, programlayıcıların veriyle

oynamalarına yardımcı olabilir; ihtiyaç duydukları her şeyi veri ile çözme, veri sürtüşme, web sitesi silme, web uygulaması oluşturma, veri mühendisliği ve daha pek çok şey yapabilmektedir. Python dili, programcıların bakımı kolay, büyük ölçekli sağlam kod yazmalarını kolaylaştırır.

Google Müdürü olan Peter Noryig; "Python programlama, başından beri Google'ın önemli bir parçasıydı ve sistem büyüdükçe ve geliştikçe değişmedi. Bugün düzinelerce Google mühendisleri Python dilini kullandı ve biz bu dilde becerileri olan daha fazla kişi arıyoruz" demiştir.

R dilinden farklı olarak, Python dilinde dâhili paketler bulunmamakla birlikte, veri bilimcilerinin yararlı istatistiksel ve makine öğrenme görevlerini yerine getirmek için kullanabilecekleri Scikit, Numpy, Pandas, Scipy ve Seaborn gibi kütüphaneleri desteklemektedir. Python programlama, sözde koda benzer ve İngilizce dili gibi mantıklıdır. Python da kodda kullanılan ifadeler ve karakterler matematiksel olabilir, ancak mantık koddan kolaylıkla anlaşılabilmektedir.

Python dilini Veri Bilimi Programlama Dilleri Kralı yapan nedir?

Python Yazılım Vakfı üyesi, Brian Curti; "Python programlamada her şey bir nesnedir. Python dilinde birkaç programlama paradigması kullanarak uygulamalar yazmak mümkündür, ancak nesne yönelimli çok net ve anlaşılabilir bir kod yazmayı sağlar" demektedir. Python dilinin özellikleri aşağıdaki gibidir:

1) Genişlik

PyPi olarak bilinen Python dili için genel paket indeksi, yaklaşık 300 farklı kategori altında listelenen yaklaşık 40.000 eklentiye sahiptir. Dolayısıyla, bir geliştirici ya da veri bilimcisi Python dili ile bir şeyler yapmak zorunda kalırsa, bir başkasının geliştirdiği eklentilere kolayca erişilebilmesinden dolayı sıfırdan eklenti yazmaya gerek duyulmaz. Python programlama genel olarak, CGI ve web geliştirme, sistem test ve otomasyon ve ETL'de oyun oynamaya kadar çeşitli görevler için kullanılır[94].

2) Verimli

Python geliştiricileri, bugünlerde büyük verileri tanımlamak ve işlemek için çok zaman harcamaktadırlar. İşlenmesi gereken artan veri miktarı ile birlikte, programcılar için bellek içi kullanımı verimli bir şekilde yönetmek son derece önemlidir. Python dili

[94] DEZYRE, "*Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (10.09.2016).

hem fonksiyonlara hem de tekrar tekrar işlemeye yardımcı olan ifadelere, yani her seferinde bir öğe üreten üreticilere sahiptir. Bu durumda, bir dizi veriye uygulanacak çok sayıda süreç olduğunda, Python dilindeki üreticiler, kaynak veriyi bir kerede toplamakta ve kaynakların tümü işleme zincirinde tek seferde geçirildiğinden büyük avantajlar sağlanılır.

3) Uzman Rehberlik Altında Kolay Uydurılabilir-Okuyun, Kolaylaştırın

Sözdizimi açık ve okunabilir olduğundan uzman rehberliği altında öğrenmeyi kolaylaştıran Python dili geniş bir popülerlik kazanmıştır. Veri bilimciler, endüstriye yönelik uzman odaklı Python programlama dersleri alarak, bilimsel hesaplamada Python ile uzmanlık bilgisi ve master programlama becerisi kazanabilmektedir. Python sözdiziminin okunabilirliği, diğer akran programcıların önceden yazılmış Python programlarını daha hızlı bir şekilde güncellemelerini kolaylaştırır ve ayrıca yeni programları hızlı bir şekilde yazmaya yardımcı olur.

Tüm bu avantajlara rağmen Python programlama masaüstü ve sunucu platformlarında popülarite kazanmış ancak Python dili kullanılarak geliştirilen çok az mobil uygulama olduğu için Pyhton mobil bilgi işlem platformlarında hala zayıf kalmaktadır. Python programlama, nadiren web uygulamalarının istemcisi tarafında bulunabilmektedir[95].

2.1.22.2. R Dili ile Veri Bilimi

R programlama dili, S. adlı bir programlama dilinin bir dalı olup, R, S'nin açık kaynaklı bir uygulamasıdır. R, S-plus'dan büyük ölçüde yalnızca komut satırı biçiminde farklıdır. R, Yeni Zelanda Auckland Üniversitesi'nde profesör olan Ross Ihaka ve Robert Gentleman tarafından geliştirilmiştir[96]. Bu profesörlerin adlarının ilk harflerinden dolayı programa R ismi verilmiştir. Programın geliştirilmesindeki temel amaç, öğrencilerin veri analizi yapıp, grafik çizebilecekleri ücretsiz bir yazılım geliştirmektir[97]. R, öncelikle istatistikçiler tarafından benimsenmiş ve günümüzde istatistiksel hesaplama için kullanılan bir programdır.

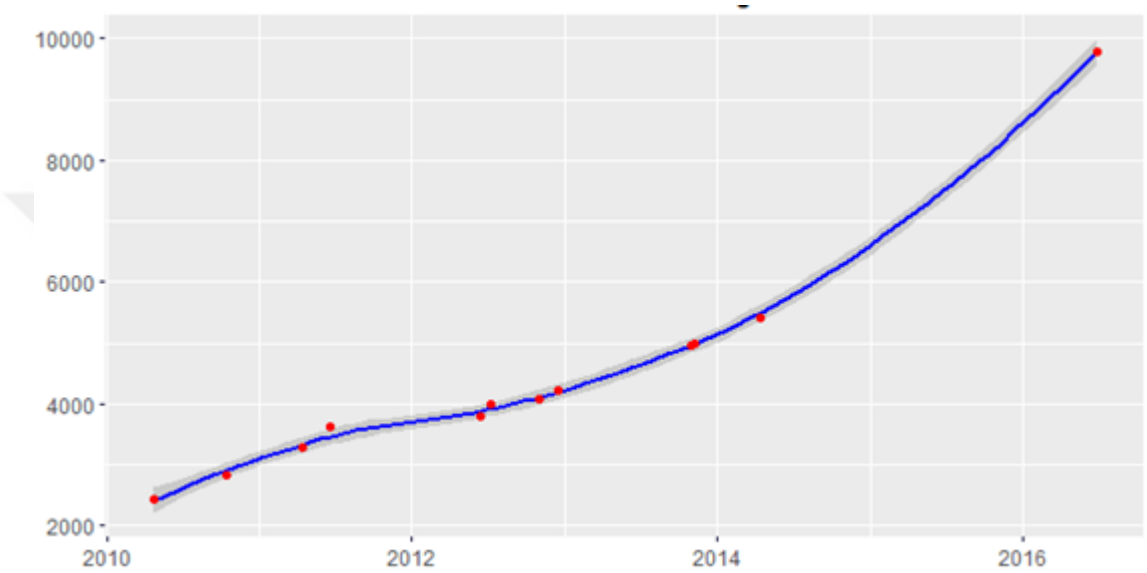
---

[95] DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).

[96] MARTİN Trevor, *''The Undergraduate Guide to R- Abeginner's introduction to the R programming language''*, Princeton University, 2016.

[97] GÜRSAKAL, *"R İle Programlama"*, 1. Baskı, Bursa, Dora Yayınevi, 2014, syf. 5.

R, Linux, Windows ve Mac'te bulunan istatistiksel hesaplama ve grafik için açık kaynak programlama dili ve ortamıdır. R dili, geliştiricilerin, veri ve kodların çapraz platform dağıtımını ve testini sağlayarak işlevselliği yeni boyutlara taşımasını sağlayan yenilikçi bir paket sistemine sahiptir. R paketleri; R fonksiyonları, veriler ve kodlardan oluşmaktadır. Paketlerin bilgisayarda saklandığı dizine library denir[98]. R, 27 Aralık 2016 itibariyle Veri Bilimi ve analizi için 10.000'e yakın ücretsiz paketi desteklemektedir (Bkz. Şekil 2.9)[99].

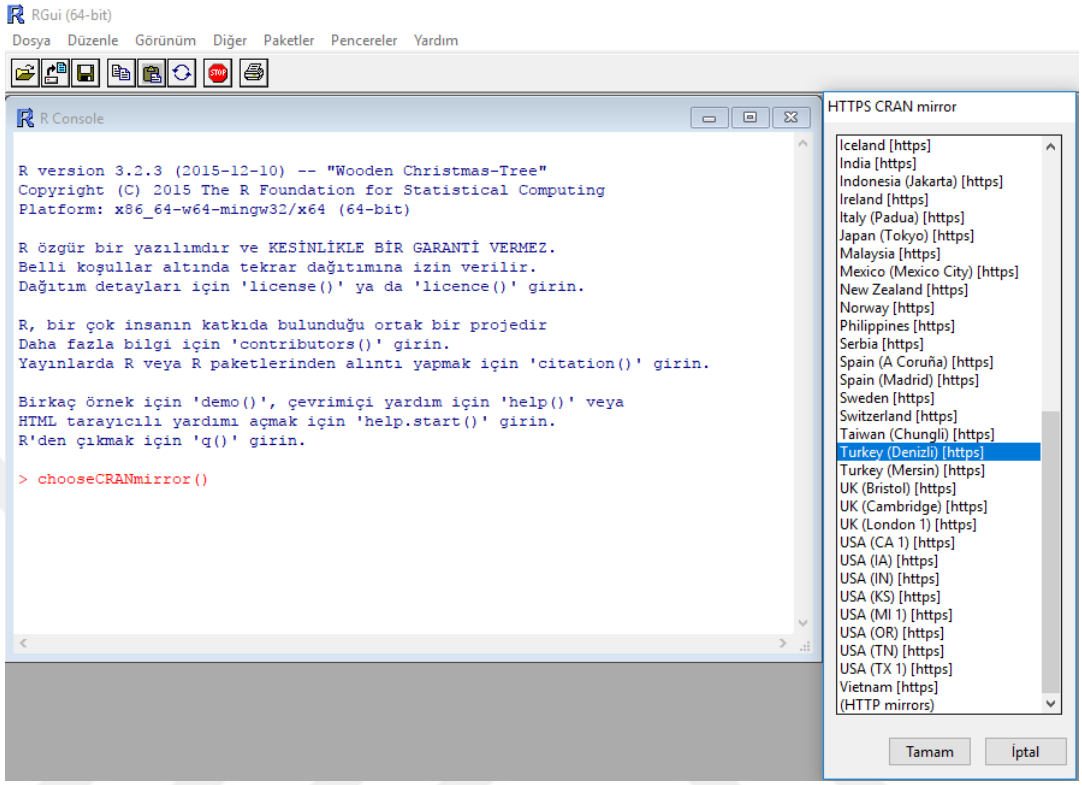


Şekil 2.9: R Paketlerindeki Artış
Kaynak: BHALLA Deepanshu, *"Companies Using R"*, Aralık 2016.

R paketlerinde R fonksiyonları, veriler ve kodlar iyi tanımlanmış bir format içerisinde bulunur. Bu paketlerden kullanılmak istenilen herhangi biri R'ye indirilip kurarak kullanılabilir. R'de paket kurarken coğrafi olarak bize en yakın olan CRAN'ı kullanmak daha mantıklıdır. Örneğin, CRAN'ı kullanırken ülkemizde "Turkey" seçeneğinden istenilen R paketini kurmak mümkündür (Bkz. Şekil 2.10). R, keşifsel veri analizi dilinin C, C ++ ve Java gibi nesne tabanlı programlama dilleriyle kolayca entegre

[98] DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).
[99] BHALLA Deepanshu, "Companies Using R", Aralık 2016, http://www.listendata.com/2016/12/companies-using-r.html, (13.12.2017).

olabilen mükemmel bir programlama dilidir. R dili, programcıların matematiği koda çevirmesini kolaylaştıran dizi yönelimli bir söz dizimine sahiptir[100].

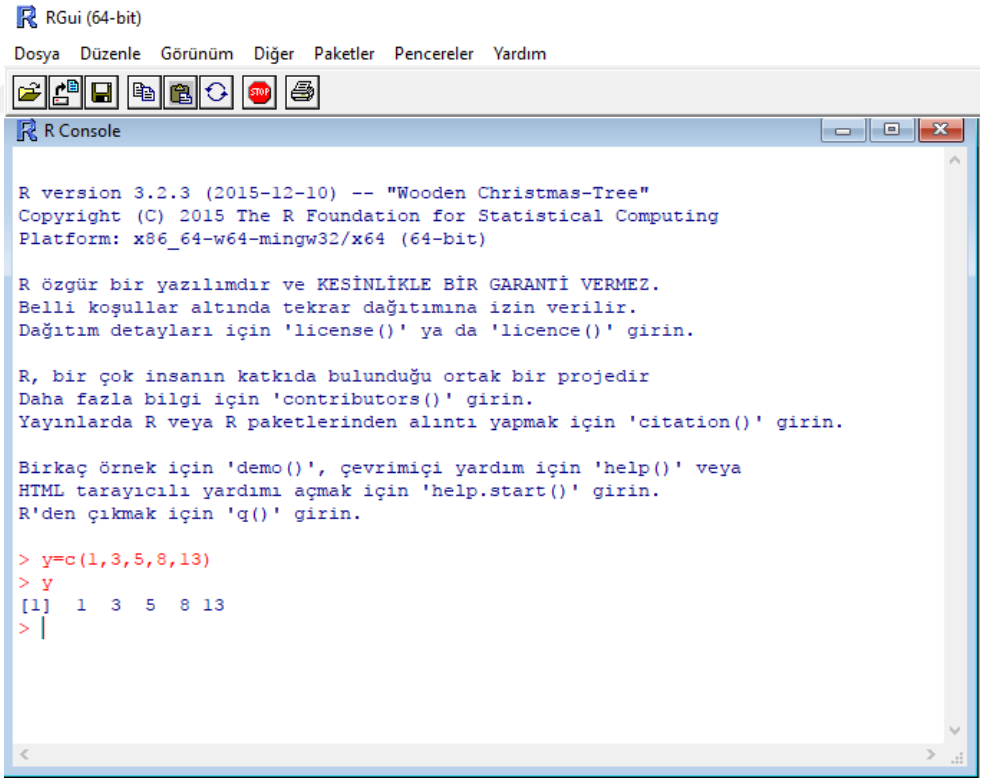


Şekil 2.10: R'de CRAN ile Bağlantı Ekranı

Milyonlarca veri bilimcisi ve istatistikçi, istatistiksel hesaplama ve niceliksel pazarlamayla ilgili büyük sorunları ortadan kaldırmak için R programlamayı kullanıyor. Günümüzde R dilini kullanan şirketler ve kullandıkları alanlar şöyledir: LinkedIn, Twitter, Bank of America, Facebook, Yhoo, Amazon, Airbnb, Google, Microsoft, Ford and Drug Administration, Ford Motor Company, Llyod ve Uber gibi finans ve işletme analiz odaklı kuruluşlar için önemli bir araç hâline gelmiştir. Bu şirketlerden bazısının R'yi kullandıkları alanlar şöyledir: Google reklam kampanyalarında yatırım gelirlerini hesaplamada, ekonomik etkinlik tahmininde, TV reklamlarının etkinlik analizinde, Facebook statü güncellemelerinde, sosyal network grafiklerinde, Microsoft istatistiksel analiz için, Bank of America raporlamada, Ford Company veri temelli karar vermede, Llyod's sigortalamada ve Uber istatistiksel analiz için kullanmaktadır[101].

[100] DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).
[101] GÜRSAKAL Necmi, *"R İle Betimsel İstatistik"*, 1. Baskı, Bursa, Dora Yayınevi, 2015, syf. 41.

R, kullanıcılara ücretsiz ve etkileşimli bir istatistiksel platform sağlamaktadır. R'nin ücretsiz olması onu diğer istatistik programlarından farklı kılmaktadır. Bu sayede istatistiksel hesaplamalar için R ücretsiz olarak indirilip kullanılabilmektedir. R istatistiği amaçlayan, nesne-yönelimli bir programlama dilini kullanmaktadır. Nesneler çeşitli sınıflarda bulunur ve bir nesne çeşitli durumlarda olabilir ve çeşitli davranışlar sergileyebilir. Nesne-yönelimli programlama dillerinde, nesneler birbirleriyle iletişim halindedirler. Nesneler dış ortam ile iletişimlerini bazı yöntemler vasıtasıyla yapmaktadır. Bu yöntemler, nesneler ile dış ortam arasında bir kullanıcı arayüzü (interface) oluşturmaktadır. Nesnelerin veri elemanları mevcut olup, birtakım sınıfların üyesidir. R'de nesne; değişken, değişken kümesi veya istatistiksel bir model olabilmektedir. Ayrıca, analiz çıktıları da bir nesne olabilmektedir. R'de nesneleri oluşturmak için yapılan işlemlere ise fonksiyon denir (Örneğin: >nesne=fonksiyon) (Bkz. Şekil 2.11). Bununla birlikte R: Vektörler, matrisler, faktörler, listeler, veri çerçeveleri gibi nesnelerin üzerinde çalışmaktadır. Genellikle R komutları bir fonksiyonu bir nesneye uygulamayı sağlamaktadır[102].



Şekil 2.11: R'de İşlemler

[102] GÜRSAKAL Necmi, *''R İle Programlama''*, 1. Baskı, Bursa, Dora Yayınevi, 2014, syf. 6-7.

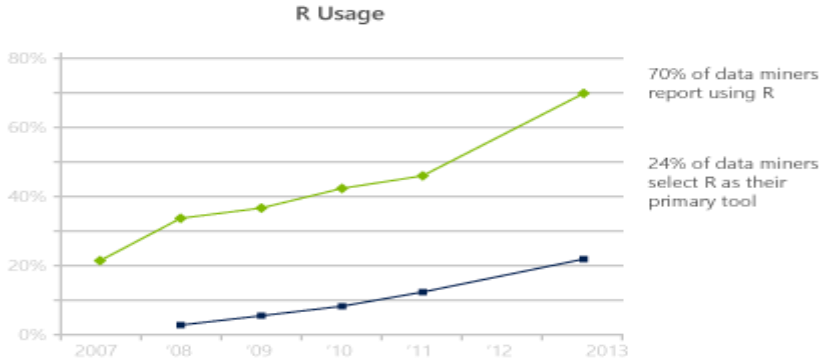
R, başlangıçta kodlayıcı olmayanlara karşı düşünülebilir bir komut satırı üzerinden girdi alan ilk programlama dilidir. Ancak yeni başlayanlar doğrudan veri görselleştirme ve istatistiksel analiz için hazır komutlar içeren önceden tanımlanmış yazılım paketleri arayabilmektedir. Önceden ayarlanmış R paketleri, R programlamayı eğlenceli ve interaktif bir şekilde öğrenmek için yeni başlayanlar tarafından kullanılabilir. R yazılım paketleri, uzman kodlama dünyasıyla ticari kara kutu çözümlerinin kolaylığı arasında bir orta yer işlevi görmektedir.

R, çeşitli görevleri gerçekleştirmek için kullanılabilecek geniş bir ön tanımlı işlev kütüphanesine sahip olan bir programlama dili olarakta görülebilir. Bu önceden tanımlanmış fonksiyonların temel odağı istatistiksel veri analizidir. Bunlar R'nin standart istatistiksel teknikler için tamamen araç kutusu olarak kullanılmasına izin vermektedir. Bununla birlikte, herhangi bir düzeyde R'yi iyi kullanmak için R programlamaya ilişkin bazı bilgilerin bilinmesi gerekmektedir. Özellikle ileri düzey kullanıcılar için, R'nin ana itirazı (diğer veri analiz yazılımlarına karşıt olarak), veri analizine uygun bir programlama ortamıdır[103]. R, neredeyse tüm Internet'e bağlı cihazlarda, yazılım uygulamalarında ve akıllı telefonlarda büyük veri analizi ve makine öğrenmesi yapabilmektedir. Bu sayede R, büyük veri setlerindeki yapıları bulmada etkin bir güce sahiptir.

R dili, dünya çapında 2 milyondan fazla istatistikçi ve veri bilimcisi tarafından kullanılmaktadır. Özellikle ticari uygulamalar için R dilinin daha geniş ölçüde benimsenmesiyle bu istatistiksel yazılımın kullanımı katlanarak artmaktadır (Bkz. Şekil 2.12). R programlama dili, akademik ortamda küçük çaplı istatistiksel analiz için geliştirilmiştir. R dili, verileri görselleştirmek, geniş veri kümelerini keşfetmek ve yeni istatistiksel modeller oluşturmak için güçlü bir istatistiksel hesaplama aracıdır. R dili, özellikle olası en iyi sonuçlar için çeşitli istatistiksel ve tahmin edici modelleri karıştırmak ve eşleştirmek amacıyla veri analizi için tasarlanmıştır. R programlama komut dosyaları, üretim dağıtımlarını ve tekrarlanabilir araştırmaları desteklemek için kolaylıkla otomatikleştirilebilir. R programlamanın başka bir özelliği de yeniden üretilebilir olmasıdır; kod ve veriler, aynı sonuçların tekrar üretilmesi için izleyebilen ilgilenen bir üçüncü partiye verilebilmektedir. Böylece, veri bilimcileri, verileri ayıklayacak, analiz edebilecek ve raporlama için bir HTML, PDF veya PPT üreten bir kod yazacaktır. Başka

[103] SARKAR Deepayan, *"An Introduction to R"*, May 2012, http://www.isid.ac.in/~deepayan/R-tutorials/labs/01_introduction_lab.pdf, (10.01.2017).

herhangi bir üçüncü taraf ilgilendiğinde, orijinal yazar benzer sonuçları üretmek için kodu ve verileri üçüncü bir kişi ile paylaşabilmektedir[104].



Şekil 2.12: R Kullanımının Zamana Göre Artışı
Kaynak: https://azurecomcdn.azureedge.net/cvtba, (20.01.2017).

R, 20 yıl içinde açık kaynak topluluğuna katkı sağlayan güçlü bir iş analitiği aracı olarak yükseliştedir. R dili, farklı kullanıcılar için farklı yüzleri bulunduğundan, en güçlü ve popüler veri bilimi araçları arasında bulunur. R programlama dili; SAS, SPSS, Minitab veya Matlab gibi pahalı istatistiksel programlama araçlarına alternatif olarak geliştirilmiş ve 1997 yılından bu yana kullanılmaktadır[105]. Bu paket programların aksine R ücretsizdir. R bu özelliği sayesinde diğer istatistik paket programlarından daha avantajlıdır. R'nin bir başka avantajı da diğer paket programlara göre R, kullanıcıları düşündürmektedir. R'de bir iş yapmak için düşünmek, paket programlarını kullanmak için sadece belirli komutların kullanılması gerekmektedir. Bununla birlikte R, grafik özelliği sayesinde diğer paket programlardan daha üstündür. R'nin bir diğer üstünlüğü ise kullanıcılara yeni fonksiyonlar ekleme imkânı sağlamasıdır[106].

2.1.23. Tableau

Tableau; Tableau Desktop, Tableau Sever ve Tableau Public dâhil olmak üzere büyük ölçekli veri setlerini işlemek için üç temel ürüne sahiptir. Tableau Desktop, verileri görselleştirmeyi ve farklı ve sezgisel bir şekilde bakmayı kolaylaştıran bir görselleştirme aracıdır. Bu Araç, kullanıcıya verilerin tüm sütunlarını vermek üzere optimize edilmiştir

[104] DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).
[105] DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).
[106] GÜRSAKAL Necmi, *"R İle Programlama"*, 1. Baskı, Bursa, Dora Yayınevi, 2014, syf. 8.

ve kullanıcıların bunları karıştırmasına izin vermektedir. Tableau Server tarayıcı tabanlı analitik sağlayan bir iş zekâsı sistemidir. Tableau Public ise etkileşimli görseller oluşturmak için kullanılır. Ayrıca Tableau, Hadoop altyapısını da içine almaktadır. Tableau, sorguları yapılandırmak için Hive'ı kullanılır ve bellek içi analize ilişkin bilgileri önbelleğe almaktadır. Önbellekleme, bir Hadoop kümesinin gecikmesini azaltmaya yardımcı olmaktadır. Bu nedenle, kullanıcılar ve büyük veri uygulamaları arasında etkileşimli bir mekanizma sağlamak mümkündür.

Tableau, kişilerin verileri görmesini ve anlamasını sağlamaktadır. Yazılım ürünleri, verilerin gücünü günlük olarak insanların kullanmasına imkân sağlamaktadır. Bu geniş bir kullanıcı popülasyonunun verilerle etkileşime girmesine, soru sormasına, sorunları çözmesine ve değer yaratmasına olanak sağlar. Stanford Üniversitesi'nde geliştirilen teknolojiye dayanan Tableau, geleneksel iş zekâsı uygulamaları ile ilgili karmaşıklığı, esnekliği ve masrafı azaltmaktadır. Excel kullanabilen herkes Tableau Desktop'ı, sürükleyip bırakan bir kullanıcı arabirimi kullanarak zengin, etkileşimli görselleştirme ve güçlü gösterge tabloları oluşturmak için Tableau'yu kullanabilir. Ayrıca, Tableau Server veya Tableau Online'ı kullanarak kuruluşlar arasında güvenli bir veri paylaşımı yapmakta mümkündür.

Bugün Tableau; Dünya Bankası, Coca-Cola, Exxon Mobil, Homeland Security, Pfizer, Fannie Mae, Gallup, Nike ve Adobe de dâhil olmak üzere 28.000'den fazla müşteri hesabına sahiptir. Allrecipes yöneticileri tüketicilerin 360 derecelik görüşlerini geliştirmek için Tableau'yu kullanıyor. Cornell Üniversitesi'ndeki 600'den fazla çalışan Tableau ile birçok analiz yapmaktadırlar. Bu çalışanlar Tableau ile katılımcı ilişkilerini yönetiyor, fakülte maaş istatistiklerini görselleştiriyor ve öğrencilerin hangi sınıflarda olduğunu izleyebiliyorlar[107].

Tableau, Google BigQuery'ye yerel, optimize edilmiş bir bağlayıcıya sahiptir ve hem canlı veri bağlantısını hem de bellek içi özetleri desteklemektedir. Tableau'nun veri harmanlaması, kullanıcılara, 40'tan fazla veri kaynağındaki verilerin BigQuery de veri eşleştirmesine olanak tanımaktadır. Tableau Server veya Tableau Online'ı kullanarak

[107] MastersInDataScience*, "using Tableau for Data Science"*, 12 sep 2014, http://www.mastersindatascience.org/data-scientist-skills/tableau/, (04.02.2017).

bulutta yayınlanan görselleştirme için Google BigQuery'ye doğrudan bağlanabilmekte mümkündür[108].

2.1.24. BigQuery

BigQuery Mayıs 2012 de Google tarafından geliştirilmiş, büyük veri kümelerinin etkileşimli analizini sağlayan bir web hizmetidir[109]. BigQuery servisi Google'ın altyapısını kullanarak büyük veri setlerinin hızlı bir şekilde analiz edilmesini sağlar. BigQuery, iç içe geçmiş verileri depolamak için kolona yönelik bir düzen kullanan dağıtılmış ve ölçeklenebilir bir sorgu sistemi olan Dremel üzerinde kurulmuştur[110]. BigQuery'nin tercih edilmesindeki en büyük etken Dremel'i kullanmasıdır. Google tarafından gerçekleştirilen Dremel altyapısı EşleİndirgeAltyapısına göre üç avantaja sahiptir. Birincisi, Dremel kolon bazlı veri modelini kullandığı için satır bazlı veri modelini kullanan EşleİndirgeAltyapısına göre daha hızlı çalışmaktadır. Bundan dolayı Dremel, büyük veri setleri üzerindeki analitik işlemlerde çok hızlıdır. İkinci olarak, kolon bazlı veri modelinde isim verileri kolon bazında tutulduğu için tekrar eden veri sayesinde sıkıştırma durumunda satır bazlı veri modeline göre Dremel daha avantajlıdır. Kolon bazlı veri modellerinin dezavantajı ise az veri ile sorgulama yapıldığında veriye birden fazla okuma ile ulaşılacağından sorgu performansında bir düşüş olur. Sonuç olarak kolon bazlı veri modelinin büyük veri setlerini okurken tüm kolonlar yerine belirli kolonlardaki veriye erişilmesinin tercih edilmesi performansın artmasını sağlamaktadır. Dremel SQL'e benzer bir sorgulama dili kullanmaktadır. Bundan dolayı EşleİndirgeAltyapısı ile uyumlu olan sistemlerdeki gibi SQL formatında sorgulama yapılmasına imkân sağlayan Hive ve Pig tarzındaki dillere gerek duymaz. Bu özelliği sayesinde sorgulama işlemlerinde EşleİndirgeAltyapısı ile uyumlu yapılara göre daha hızlıdır. Üçüncü olarak, ölçeklenebilirlik açısından; Dremel Google'a göre büyük ölçekli sistemler ile test edilmiş tek yöntemdir. Dremel, BigQuery servisleri ile büyük veri analizleri Google sunucuları üzerinden

[108] FENG Jeff – Marc LOBREE – Tino TERESHKO – Mike GRABOSKI, *"Google BigQuery & Tableau: Best Practices"*, 2017, http://www.tableau.com/sites/default/files/media/whitepaper_googlebigquerytableaubestpr/, (06.01.2017).
[109] Google BigQuery, *"Developers Google BigQuery"*, October 2012, http://developers.google.com/bigquery/, (05.02.2017).
[110] MELNİK Sergey – Andrey GUBAREV – Jing Jing LONG – Geoffrey Romer – Shiva SHIVAKUMAR – Matt TOLTON – Theo VASSILAKIS, *"Dremel: Interactive Analysis of Web-Scale Datasets"*, Proceedings of the VLDB Endowment, Vol. 3, No. 1, Singapore, 2010.

yapıldığı için ayrıca bir veri merkezi kurulmasına gerek olmadığı için herhangi bir sermaye ayırmaya gerek duyulmaz[111].

BigQuery, projenize ve bütçenize uyması için ölçeklenebilir, esnek fiyatlandırma seçenekleri sunmaktadır. BigQuery de veri depolama, ek içerik akışı ve veri sorgulama için ücret alınır ancak veri yükleme ve aktarma ücretsizdir. BigQuery, günlük maliyetlerinizi istediğiniz miktarda tutmanıza olanak tanıyan maliyet kontrolü mekanizmaları sağlar. Yerel para birimi cinsinden ödeme yaparken Google, önde gelen finansal kuruluşlar tarafından yayınlanan dönüşüm oranları uyarınca, listelenen fiyatları geçerli yerel para birimine dönüştürmektedir.

Aşağıdaki Tablo 2.2'de, BigQuery fiyatlandırması özetlenmiştir. BigQuery'nin kota politikası bu tablodaki işlemler için geçerlidir.

İşlem	Maliyet	Notlar
Depolama	Aylık 1 GB $0,02	Her ay ilk 10 GB ücretsiz
Uzun Vadeli Depolama	Aylık 1 GB $0,01	Uzun vadeli depolama fiyatı için Bkz. https://cloud.google.com/bigquery/pricing#long-term- storage
Akış Ekleme (Streaming Insert)	1 GB $0,05	Depolama alanı fiyatlandırması için Bkz. https://cloud.google.com/bigquery/pricing#storage
Sorgular (Queries)	1 TB $5	Ayda ilk 1 TB ücretsizdir. İsteğe bağlı fiyatlandırma için Bkz. https://cloud.google.com/bigquery/pricing#on_demand_pricing Yüksek hacimli müşteriler için sabit fiyatlı fiyatlandırma da mevcuttur.
Veri Yükleme	Ücretsiz	BigQuery'ye veri yüklemek için Bkz. https://cloud.google.com/bigquery/loading-data.
Veri Kopyalama	Ücretsiz	Mevcut tabloya kopyalama yapmak için Bkz. https://cloud.google.com/bigquery/docs/tables#copyingtable.
Veri Aktarma	Ücretsiz	BigQuery'den veri dışa aktarmak için Bkz. https://cloud.google.com/bigquery/docs/exporting-veri.
Meta veri işlemleri	Ücretsiz	Listeleme, çağırma, yapıştırma, güncelleme ve silme çağrıları.

Tablo 2.2: BigQuery Fiyatlandırma Tablosu
Kaynak: https://cloud.google.com/bigquery/pricing, (12.05.2017).

[111] DERİNÖZ Cenk, "Google BigQuery Servisi İle Büyük Veri İşlemleri Ve Sorgu Sonuçlarının BIME İş Zekası Ürünü İle Görselleştirilip Android Tabanlı Mobil Cihazlar Üzerinden İzlenmesi", Data & Analytics, Nisan 22, 2014.

BigQuery kolon bazlı veri yapısına sahip olduğu için Tablo 2.2'deki ücretlendirme de tabloda bulunan tüm kolonlar baz alınmaz, sadece sorgulama içerisindeki kolonlar baz alınmaktadır. Bundan dolayı ücretlendirme yapılırken seçilen kolonlarda işlenen toplam veri miktarı dikkate alınmaktadır.

2.1.24. 1. BigQuery Bileşenleri

BigQuery'nin altyapısında projeler, tablolar ve veri kümeleri olmak üzere üç temel bileşen bulunmaktadır.

1. Projeler

Projeler, Google Bulut Platformu'ndaki üst düzey yapılardır. Projeler, faturalandırma ve yetkili kullanıcılar hakkında bilgi depolarlar ve BigQuery verilerini içermektedir. Her projenin bir arkadaşlık ismi ve farklı bir ID'si vardır. BigQuery de herhangi bir projeye kayıt olmak için aşağıdaki adımlar izlenmelidir[112].

1. Adım: *Google APIs Console*'a giriş yapılır (https://console.cloud.google.com).

2. Adım: Yeni bir Google API Console projesi oluşturun veya var olan projeyi kullanın.

3. Adım: API Services table'a git sol üstteki *Ürünler ve hizmetler* menisünü tıkla, API Yöneticisi'ni ve daha sonra Google API'lerini tıkla.

4. Adım: BigQuery'i açarak *Google Cloud API'leri* altındaki BigQuery API'sine giriş yaparak bir sonraki sayfada API'yi etkinleştir'i tıkla.

5. Adım: Arzu edilirse hizmet koşulları incelenip onaylanır.

2. Tablolar

Bir BigQuery tablosu, satırlar halinde düzenlenmiş bireysel kayıtlara ve her bir sütuna atanan bir veri türüne (ayrıca alan adı da denir) sahip standart, iki boyutlu bir tablodur. Bir kayıt içindeki ayrı alanlar iç içe geçmiş ve tekrarlanan alanları içerebilmektedir. Her tablo; alan adlarını, türlerini ve diğer bilgileri açıklayan bir şema tarafından tanımlanmaktadır. Şemayı daha sonra değiştirmek gerekiyorsa şema güncellenebilir. İlk tablo oluşturma isteğinde bir tablonun şemasını belirtebilir veya bir

[112] Google BigQuery Export, "BigQuery veri yönetimi", 2017, https://support.google.com/analytics/answer/3416092?hl=tr&ref_topic=3416089, (12.05.2017).

şema olmadan bir tablo oluşturabilir ve şemayı, sorguyu ya da tabloya ilk giren işi yükleyerek bildirilebilir. BigQuery aşağıdaki tablo türlerini destekler[113]:

Yerel tablolar: Yerel BigQuery deposuyla desteklenen tablolar.

Harici tablolar: BigQuery dışındaki depolama alanıyla desteklenen tablolar.

Views: Bir SQL sorgusu tarafından tanımlanan sanal tablolar.

BigQuery'deki tablolar, ilişkisel veritabanlarındaki tablolara benzemektedir. Bu tablolar yapılandırılmış verilerin satır ve sütun koleksiyonlarıdır. Verinin yapılandırılmış olması, tablodaki tüm veriler için geçerli olan bir şemaya sahip olması anlamına gelmektedir. BigQuery tabloları, satır düzeyinde güncellemeleri desteklemez. Tablolar eklenebilir, böylece oluşturulduktan sonra büyümeye devam edebilirler. Eğer verilerden periyodik olarak oluşturulan varsa ek ilave yapmak faydalı olabilir; böylece yeni veriler geldikçe tabloya eklenebilir. Ayrıca tablolar kesilebilir, yani tek bir atomik işlem ile tablolar silinebilir ve yeni verilerle değiştirilebilir[114].

3. Veri Kümeleri

BigQuery veri kümeleri tabloların birleşiminden oluşmaktadır. Bunlar, bir veritabanına benzer olarak düşünülmüş tabloların mantıksal gruplamalarıdır. Bununla birlikte, veri kümeleri, veritabanlarının aksine, tabloların nerede ve nasıl depolandığına dair herhangi bir bilgi vermez. Ayrıca veriler farklı veri kümelerinde birleşebilmekte, buna ilişkin herhangi bir kısıtlama da yoktur.

Tüm temel tablolar için veri kümeleri, erişimi denetleyen BigQuery de birincil paylaşım birimidir. Veri kümeleri, veri kümesi için okuyucuları, yazarları ve sahiplerini belirten ACL'lere sahiptir. Okuyuculara ve yazarlara veri kümelerin de bulunan tablolardaki verileri okumaya ve bu tablolar da sorgulama yapmaya izin verilir. Bununla birlikte veri kümeleri sahiplerine ve yazarlarına tabloları oluşturma ve veri kümesindeki ACL'yi değiştirmeye de izin verilir. Varsayılan olarak, bir veri kümesindeki ACL, yalnızca Google Developers Console'da kurulan proje izinlerini belirtmektedir. Proje sahipleri, veri seti sahiplerine eşleme yapar; Proje editörleri veri seti yazarlarına eşlenir ve proje izleyicileri veri seti okuyucularıyla eşleşmektedir. Bununla birlikte, bu ACL'ler değiştirilebilir ve proje izinleri girdi veri kümesinden kaldırılabilir.

---

[113] Google Cloud Platform, *"Using Tables"*, 2017, https://cloud.google.com/bigquery/docs/tables, (25.05.2017).

[114] TIGANI Jordan – Siddartha NAIDU, *"Google® BigQuery Analytics"*, Published by John Wiley & Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, Indiana Published simultaneously in Canada, ISBN: 978-1-118-82482-5, 2014.

Depolama

BigQuery, sıkıştırma, şifreleme, çoğaltma, performans ayarlama ve ölçeklendirme gibi yapısal verileri depolamanın teknik yönlerini yönetir. BigQuery, verileri Kapasitör sütuniçi veri biçiminde saklar ve tabloların, bölümlerin, sütunların ve satırların standart veritabanı kavramlarını sunmaktadır.

BigQuery depolama alanına toplu yüklemeler veya akışla veri yükleyebilir, tabloları kopyalamak, SQL kullanarak sorgu tabloları oluşturmak, SQL DML ile verileri değiştirmek, verileri dışa aktarmak veya depolanan verileri Kimlik ve Erişim Yönetimi (IAM) izinlerini kullanarak diğerleriyle paylaşmak gibi veri işlemleri gerçekleştirmek mümkündür. Ayrıca, BigQuery, BigQuery depolama alanındaki verileri sorgulamayı da desteklemektedir.

İşler (Jobs)

İşler, BigQuery'de veri yükleme, verileri dışa aktarma, verileri sorgulama veya verileri kopyalamak için yürütülen işlemlerdir. Bazen işlerin tamamlanması yapılan işin büyüklüğüne bağlı olarak uzun zaman alabilmektedir. Yapılan bu işler eşzamansız olarak yürütülür ve işlerin durumları için sorgulamalar yapılabilmektedir. BigQuery, bir projeyle ilişkili tüm işleri, Google Bulut Platformu Konsolu aracılığıyla erişilebilen bir geçmişi kaydetmektedir.

BigQuery'deki işler sonsuza kadar devam etmektedir. Buna, başarılı olup olmaması ya da çalışıyor olup olmaması veya tamamlanmış işler de dâhildir. Bir proje sahibi değilseniz, projenizle ilgili herhangi bir işte tüm eylemleri gerçekleştirebildiğiniz sürece, yalnızca başlattığınız işler hakkında liste veya bilgi alabilirsiniz.

Yuvalar (Slots)

Bir BigQuery yuvası, SQL sorgularını çalıştırmak için gereken bir hesaplama kapasitesi birimidir. BigQuery, sorgu boyutuna ve karmaşıklığına bağlı olarak her sorgu tarafından kaç yuvanın kaç tane yuva gerektireceğini otomatik olarak hesaplar. Çoğu kullanıcı, varsayılan yuva kapasitesini yeterli bulmamaktadır. Çünkü daha fazla yuvaya erişim, sorgu başına daha hızlı performansı garanti etmez. Bununla birlikte, daha büyük bir yuva havuzu, çok büyük veya çok karmaşık sorguların performansını ve aynı anda eşzamanlı iş yükünün performansını arttırabilmektedir.

BigQuery, alan kotanınızı, müşteri geçmişine, kullanıma ve harcamaya dayalı olarak otomatik olarak yönetir. Aylık en az 40.000 ABD doları tutarındaki analitik

harcama yapan müşteriler için, BigQuery tahsis edilen yuvaların sayısını artırmanın çeşitli yollarını sunmaktadır[115].

2.2. Büyük Veri Analizinde Kullanılan Teknikler

Bugün istatistik ve bilgisayar bilimlerinde kullanılan araçlara bağlı olarak veri analizinde farklı birçok teknik kullanılmaktadır. Araştırmacılar özellikle verilerin yeni kombinasyonlarını analiz etmek için yeni teknikleri geliştirerek mevcut olanları da geliştirmeye devam ediyor. Bugün için büyük miktardaki veriyi analiz eden en gelişmiş teknikler şunlardır: Yapay Sinir Ağları, Tahmini Analiz Yöntemleri, İstatistikler ve Doğal Dil İşleme'dir. Büyük veri işleme yöntemleri, uygulamalı matematik, istatistik, bilgisayar bilimleri ve ekonomi gibi farklı disiplinlerden yararlanmaktadır. Bu disiplinler Veri Madenciliği, Sinir Ağları, Makine Öğrenmesi, Sinyal İşleme ve Görselleştirme Yöntemleri gibi veri analiz tekniklerinin temelini oluşturmaktadır. Bu yöntemlerin çoğu birbiri ile ilişkili olup veri işleme sırasında eşzamanlı olarak kullanılır. Dikkat edilirse bu teknikler büyük veri kullanılmasını gerektiren tekniklerin tamamı değildir. Bazıları küçük veri setlerine de etkili bir şekilde uygulanabilmektedir. Örneğin, A/B testi ve regresyon analizi küçük veri setlerine de uygulanabilmektedir. Ancak burada listelenen tekniklerin tamamı büyük verilere uygulanabilir[116].

2.2.1. A / B Testi

A/B testi hangisinin daha iyi performans gösterdiğini belirlemek için bir web sayfasının veya uygulamanın iki sürümünü birbiri ile karşılaştırma yöntemidir. A/B testi, esas olarak, bir sayfanın iki veya daha fazla varyantının rasgele kullanıcılara gösterildiği bir deneydir ve hangi varyasyonun belirli bir dönüşüm hedefi için daha iyi performans gösterdiğini belirlemek için istatistiksel analiz kullanılır[117]. Bölme (split) testi veya bucket testi olarakta bilinen bu test genellikle dijital pazarlamacılar tarafından

[115] Google Cloud Platform, *"Slots"*, 2017, https://cloud.google.com/bigquery/docs/slots, (29.08.2017).
[116] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.
[117] OPTIMIZELY, "A/B Testing", 20.11.2017, https://www.optimizely.com/optimization-glossary/ab-testing/, (20.11.2017)

kullanılmaktadır. Bu yöntem genellikle web sitesi tasrımında, e-posta pazarlamada ve pazarlama kampanyalarında kullanılmaktadır. A/B testinde temel amaç, karşılaştırma yapılan etkenler (yazılı içerikler, görsel içerikler, satış sayıları, karşılama sayıları, butonlar vs.) içerisinde en iyi performansı göstereni veya en iyi dönüşümü sağlayanı bularak en iyi sonucu veren etkeni siteye uygulamaktır[118].

2.2.2. İlişkili Kurallı Öğrenme

Bu teknik kümelerdeki ilginç ilişkileri keşfetmek için kullanılır. Birliktelik kuralları ile büyük veritabanlarındaki değişkenler arasındaki ilişkileri keşfetmek mümkündür. Bu teknikler, üretmek ve olası kuralları test edecek bazı algoritmalar içermektedir. Örneğin; bir perakendeci, pazarlama için pazar sepeti analizi uygulaması ile hangi ürünlerin birlikte satılacağını belirleyebilmektedir.

2.2.3. Sınıflandırma

Sınıflandırma tekniği bir dizi kategorilere ayırmak için kullınılır. Bu teknik yeni veri noktalarına sahip, daha önceden kategorize edilmiş veri noktalarını içeren bir eğitim setine dayanmaktadır. Örnek olarak özel müşteri segment (iş kolu) davranışının tahmini verilebilir ki, burada kesin bir hipotez ya da objektif bir sonuç yoktur. Yine müşterilerin satın alma kararları, abone kayıp ve tüketim oranı sınıflandırmaya örnek olarak verilebilir. Bu tekniklerde genellikle bir eğitim seti mevcut olduğundan *denetimli öğrenme (supervised learning)* olarak tanımlanmaktadır. Ayrıca, bu teknikler *denetimsiz öğrenmenin (unsupervised learning)* bir türü olan *kümeleme analizine* zıt olup veri madenciliği (data mining) için kullanılır[119].

2.2.4. Kümeleme Analizi

Kümeleme Analizi, nesneleri sınıflandıran istatistiksel bir yöntemdir. Bu yöntem benzer nesneleri daha küçük gruplar halinde çeşitli alt gruplara ayırır ki, bu nesnelerin benzer karakteristik özellikleri daha önceden bilinmemektedir. Kümelemede amaç sınıflar arasındaki benzerliğin minimum, sınıfın kendi içerisinde benzerliğinin

[118] MACAR Barış, "A/B Testi Nedir? Ne için kullanılır", 11.01.2016, http://www.webmasto.com/ab-testi-nedir-ne-icin-kullanilir-infografik, (20.11.2017).
[119] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, competition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

maksimum olmasıdır. Kümeleme analizine bir örnek olarak hedefe yönelik pazarlama için tüketicilerin kendi benzerliklerine göre gruplara ayrılması verilebilir. Bu yöntem *denetimsiz öğrenmenin* bir türü olduğundan burada eğitim verisi kullanılmaz. Kümeleme analizi sınıflandırmaya zıt olup genellikle veri madenciliğinde kullanılmaktadır[120].

2.2.5. Kalabalığın Gücü

Genellikle Web gibi sosyal medya aracıları açık bir çağrı yoluyla kişi ya da toplulukların çok büyük bir grubu tarafından veri toplamak için kullanılan bir tekniktir[121]. Bugün teknolojinin gelişmesiyle birlikte dünya Web 2.0 denilen internet çağına girmiştir. Daha önce sadece bilgi verme amaçlı olan Web 1.0 siteleri bugün artık çok geride kalmıştır. Web 2.0 bu siteyi ziyaret edenlerin siteye etkin bir şekilde katkıda bulundukları, ilaveler yapabildikleri web siteleridir. Bu tür sitelere örnek olarak; Google Adsense, Vikipedi, Flickr, WordPress ve Blogger verilebilir. Ayrıca, bu tür Web 2.0 siteleri devasa büyüklükteki veriyi toplama kapasitesine sahiptirler[122].

2.2.6. Veri Füzyonu (Kaynaştırma-Birleştirme) ve Veri Entegrasyonu

Veri Füzyonu, birçok kaynaktan gelen verilerin entegrasyonunu ve analizini yapan bir dizi tekniktir. Bu şekilde geliştirilen teknikler ile elde edilen veriler, tek bir kaynaktan analiz yapılarak elde edilen veriden daha etkili ve potansiyel olarak daha doğrudur. Sinyal işleme teknikleri veri füzyonunun bazı türleri için kullanılabilmektedir. Bu tekniğe örnek olarak, bir Arıtma Tesisi gibi karmaşık bir dağıtılmış sistemin performansı hakkında bütünleşik bir perspektif geliştirmek için birleştirilen Nesnelerin İnternet'inden gelen sensör verisi verilebilir. Yine bir pazarlama kampanyasının müşteri duyarlığı ve satın alma davranışı üzerindeki etkisini belirlemek için doğal dil işleme yöntemiyle analiz edilen sosyal medyadan alınan veriler gerçek zamanlı satış verileriyle birleştirilebilir[123].

---

[120] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

[121] HOWE Jeff Howe, *"The Rise of Crowdsourcing",* Wired, Issue 14, 06.01.2006, https://www.wired.com/2006/06/crowds/, (02.09.2016).

[122] EL Siraceddin, *"Nedir Bu Web 2.0 Teknolojisi?",* 21 Haziran 2008, http://sanalkurs.net/nedir-bu-web-2-0-teknolojisi-2212.html, (02.09.2016).

[123] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

2.2.7. Veri Madenciliği

Veri Madenciliği, veritabanı yönetimi ile istatistik ve makine öğrenme (machine learning) yöntemlerini birleştirilerek büyük veri setlerinden desenleri ayıklamak için kullanılan bir dizi tekniktir. Bu teknikler, ilişkili öğrenme, kümeleme analizi, sınıflandırma ve regresyondır. Veri Madenciliği'ne örnek olarak; müşterilerin satın alma davranışlarını modellemek için pazar sepeti analizinin kullanılması, insan kaynaklarının veri madenciliğini kullanarak en iyi çalışanlarının karakterlerini belirlemesi ya da müşteri verisi kullanarak bir teklife olası verilebilecek cevapların belirlenmesi verilebilir.

2.2.8. Toplu Öğrenme

Toplu Öğrenme, birden fazla tahmin modeli için kurulan modellerden herhangi birinin elde edilebilir modelden daha iyi bir öngörü performansına sahip olduğunu söylemektedir. Bu modellerin her biri istatistik veya makine öğrenmesi kullanılarak geliştirilmiştir. Toplu öğrenme bir tür denetimli öğrenmedir.

2.2.9. Genetik Algoritmalar

Genetik Algoritmalar, optimizasyon için kullanılan bir teknik olup, bu doğal evrim sürecinden veya "en güçlünün hayatta kalmasından" esinlenmiştir. Bu teknikte, olası çözümler birleştirilebilir ve mutasyon geçirebilir "kromozomlar" olarak kodlanmıştır. Bu bireysel kromozomlar nüfusun her bir bireyinin dayanıklılığını ya da performansını belirleyen bir modellenen "çevre" içinde hayatta kalmak için seçilir. Genellikle "evrimsel algoritma" türü olarak tanımlanan genetik algoritmalar doğrusal olmayan (nonlinear) problemlerin çözümü için çok uygundur. Genetik algoritmalara örnek olarak, üretimde iş planlaması iyileştirilmesi ve yatırım portföyünün performansının optimize edilmesi verilebilir[124].

2.2.10. Makine Öğrenme

Yapay zekâ olarak da adlandırılan makine öğrenme; algoritmaların tasarımı ve geliştirilmesi ile ilgili bilgisayar biliminin bir alt bilim dalıdır. Bu algoritmalar bilgisayarların ampirik verilere dayalı davranışları evrimleştirmeye izin vermektedir.

[124] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, competition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

Makine öğrenme araştırmalarının en önemli odak noktası otomatik olarak karmaşık desenleri tanımak ve verilere dayalı akıllı kararlar almaktır. Makine öğrenmeye örnek olarak, doğal dil işleme verilebilir.

2.2.11. Doğal Dil İşleme

Doğal Dil İşleme (NLP), bilgisayar bilimi ve dil biliminin bir alt bilim dalından gelen tekniklerin bir kümesi olup, beşeri (doğal) dil analizinde bilgisayar algoritmalarını kullanmaktadır. Birçok NLP tekniği makine öğrenme türleri arasında yer almaktadır. NLP'ye örnek olarak, müşterilerin marka kampanyasına tepkilerini belirlemek için sosyal medya duygu analizinin kullanılması verilebilir.

2.2.12. Sinir Ağları

Bilişimsel modeller, verilerdeki desenleri bulmak için biyolojik sinir ağlarının yapısından ve çalışmasından esinlenerek geliştirilmiştir. Sinir ağlarına örnek olarak, bir beynin içindeki hücreler ve bağlantıları verilebilir. Sinir ağları tekniği nonlineer (doğrusal olmayan) desenleri bulmakta oldukça başarılıdır. Ayrıca sinir ağları, örüntü tanıma ve optimizasyon için de kullanılabilmektedir. Bazı sinir ağı uygulamaları denetimli öğrenmeyi içerirken bazısı da denetimsiz öğrenmeyi içermektedir. Bununla birlikte, sinir ağlarına örnek olarak, belirli bir şirketten ayrılma riskiyle karşı karşıya olan yüksek değerli müşterilerin ve sahte sigorta taleplerinin belirlenmesi verilebilir.

2.2.13. Ağ Analizi

Ağ Analizi, bir grafik ya da ağda ayrık düğümler arasındaki ilişkileri karakterize etmek için kullanılan teknikler kümesidir. Sosyal ağ analizinde, bir toplulukta ki veya kuruluşta (grupta) ki bireyler arasındaki ilişkiler analiz edilir. Örneğin, bilgi nasıl transfer edilir veya kimin kim üzerinde en büyük etkiye sahip olduğunu ağ analizi tekniği ile bulmak mümkündür.

2.2.14. Optimizasyon

Optimizasyon, eldeki sınırlı kaynakların en etkin şekilde kullanılması anlamına gelmektedir. Optimizasyonu matematiksel olarak bir fonksiyonun maksimize veya minimize edilmesi olarak tanımlamakta mümkündür. Optimizasyona örnek olarak;

maliyet, hız ya da güvenilirliği vermek mümkündür. Optimizasyon uygulama örnekleri; geliştirici işlemsel süreçler olarak; zaman planlama, dağıtım ve zemin düzenleme ve stratejik kararlar olarak; ürün yelpazesi stratejisi, bağlantılı yatırım analizleri ve Ar-Ge portföy stratejisini içermektedir. Bununla birlikte genetik algoritmalar da optimizasyon tekniğine örnek olarak verilebilir[125].

2.2.15. Örüntü Tanıma

Örüntü Tanıma, belirli bir algoritmaya göre verilen giriş değeri için çeşitli çıktı değeri atayan makine öğrenmesi teknikleri kümesidir. Sınıflandırma teknikleri örüntü tanımaya örnek olarak verilebilir.

2.2.16. Öngörü Modellemesi

Öngörü Modellemesi, bir matematiksel model oluşturmada ya da tahmin sonuçları arasından en iyi tahmin modelini seçmede kullanılan tekniklerin kümesidir. Müşteri ilişkileri yönetimine örnek bir uygulama olarak; benzerlik (likelihood) tahmini ile müşteride olacak abone kaybı için tahmin modellerinin kullanılması ya da benzer olarak bir müşteriye başka bir ürünün çapraz olarak satılması verilebilir. Ayrıca, regresyonu birçok öngörü modellemesi tekniklerine örnek olarak vermek mümkündür.

2.2.17. Duygu Analizi

Duygu Analizi, doğal dil işleme uygulaması ve diğer analitik teknikler; metin (text) kaynaklarından öznel bilgilerin belirlenmesi ve ayıklanması için kullanılır. Bu analizlerin kilit unsurları; özellik, görünüş ya da ürün tanımlamayı içermektedir. Bu duygu türü, "polarite (kutupluluk)" (yani, pozitif, negatif veya nötr) ve derecesi ve gücü olarak ifade edilmekte ve belirlenmektedir. Duygu analizini şirketler; sosyal medyayı (örneğin; bloglar, mikroblog ve sosyal ağlar) analiz etmek için, farklı müşterilerin ve hissedarların ürün ve eylemlere nasıl bir tepki verdiklerini belirlemek için kullanmaktadır[126].

---

[125] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, "Big Data: The next frontier for innovation, compettition, and productivity", Report McKinsey Global Institute, JUNE 2011.

[126] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

2.2.18. Sinyal İşleme

Sinyal işleme, elektrik mühendisliğinden ve uygulamalı matematikten gelen tekniklerin kümesidir. Bu teknikler ayrık (discrete) ve sürekli sinyalleri analiz etmek için geliştirilmiştir. Yani analog fiziksel büyüklüklerin (dijital olarak temsil edilse bile) temsilleri; radyo sinyalleri, ses ve görüntüdür. Bu kategori sinyal algılama tekniklerini içermekte ve bunlar sinyal ve ses arasındaki farkı ölçebilmektedir[127].

2.2.19. Mekânsal Analiz

Mekânsal analiz, insan davranış kalıplarını ve mekânsal ifadesini, matematik ve geometri bakımından, yani konumsal analiz açısından açıklamaya çalışan bir coğrafi analiz türüdür. Mekânsal analizin sonuçları, analiz edilen nesnelerin konumuna bağılı olup, bu tekniği uygulamak için nesnelerin konumlarına ve özelliklerine erişmek gerekmektedir. Konum verilerine örnek olarak, adresler veya enlem/boylam koordinatları da dâhil olmak üzere verileri yakalayan coğrafi bilgi sistemleri (GIS) verilebilir. Mekânsal analiz uygulamalarına örnek olarak mekânsal regresyonlar (Örneğin, bir ürünün yer ile ilişkili tüketici istekliliği nasıl olur?) veya simülasyonlar (Örneğin, bir imalat tedarik zinciri ağı, farklı yerlerdeki sitelerle nasıl bir performans gösterebilir?) verilebilir.

2.2.20. İstatistikler

İstatistikler, anketlerin ve deneylerin tasarımı da dâhil olmak üzere verilerin toplanması, organizasyonu ve yorumlanması ile ilgilenen bir alandır. İstatistiksel teknikler, genellikle değişkenler arasındaki ilişkilerin olma ihtimallerini değerlendirmek için (sıfır hipotezi) ve bu değişkenler arasındaki ilişkilerin altında yatan nedensel ilişkinin bir türü hakkında bir karara varmak için (örneğin; istatistiksel olarak anlamlı) kullanılır. İstatistiksel teknikler I Tip hataların ve II Tip hataların olasılığını azaltmak için kullanılır. İstatistiklere örnek bir uygulama olarak geliri en çok arttıracak pazarlama materyali türünün belirlenmesi için A/B testinin kullanılması verilebilir[128].

---

[127] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

[128] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, compettition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

2.2.21. Denetimli Öğrenme

Denetimli Öğrenme, makine öğrenmesi tekniklerinin kümesi olup bu eğitilen bir veri kümesinden bir fonksiyonun ya da ilişkinin çıkarılmasıdır. Denetimli Öğrenmeye örnek olarak sınıflandırma ve destek vektör makinesi verilebilir[129]. Denetimli Öğrenme yöntemi Denetimsiz Öğrenmenin farklı bir versiyonudur.

2.2.22. Denetimsiz Öğrenme

Denetimsiz Öğrenme, makine öğrenme tekniklerinin bir kümesidir. Bu teknik çözümlenememiş verideki gizli yapıyı bulur. Kümeleme analizi (denetimli öğrenmenin aksine) denetimsiz öğrenmeye örnek olarak verilebilir[130].

2.2.23. Simülasyon

Karmaşık sistemlerin davranışlarının modellenmesi için genellikle öngörü, tahmin ve senaryo planlaması kullanılır. Örneğin, Monte Carlo simülasyonları, tekrarlanan rastgele örneklemeye dayanan bir algoritma sınıfıdır. Yani, simülasyonlar binlerce farklı varsayımlara dayalı olarak çalıştırılıyor. Sonuçta bir histogram elde edilir ve bu histogram sonuçları bir olasılık dağılımını vermektedir. Simülasyon uygulamaları çeşitli girişimlerin başarısı konusunda belirsizlikleri verilen mali hedeflere ulaşmanın olasılığını değerlendirmektedir[131].

2.2.24. Zaman Serileri Analizi

Zaman Serileri, ardışık eşit zaman aralığındaki veri noktalarının dizilerini analiz etmek ve veriden anlamlı sonuçlar elde etmek için istatistik ve sinyal işleme teknikleri kullanılır. Zaman serisine örnek olarak bir borsa endeksinin saatlik borsa değeri ya da her gün belli koşullar altında tanısı konulan hasta sayısı verilebilir. Zaman serisi tahmini; aynı veya başka bir dizi bilinen geçmiş değerlere dayalı bir zaman serisinin gelecekteki değerlerini tahmin etmek için modelin kullanılmasıdır. Bu tekniklerden bazıları örneğin

---

[129] CORTES Corinna – Vladimir VAPNIK,"Support-vector networks", Machine Learning, Issue 3, Volume 20, 20 February 1995, pp 273-297.

[130] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, competition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

[131] MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, competition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

yapısal modelleme yaparak serileri; trend, mevsimsellik ve kalıntı bileşenlerinden ayrıştırır. Zaman serileri uygulamalarına örnek olarak tahmini satış rakamları ya da bulaşıcı bir hastalık için tanısı konulacak insan sayısının tahmin edilmesi verilebilir.

2.2.25. Görselleştirme (Visualization), Büyük Veri ve Görselleştirme

Görselleştirme, büyük veri analizlerini iletmek, anlamak ve geliştirmek için resimler, diyagramlar ya da animasyonlar oluşturmak için kullanan bir tekniktir. Görselleştirme, insan ve elektronik veri işlemenin güçlü yönlerini birleştiren bir teknoloji sunmaktadır. Görselleştirme, insanların ve makinelerin, en etkili sonuçlar için farklı yetenekleri kullanılarak işbirliği yaptığı, yarı otomatikleştirilmiş bir analitik sürecin aracı hâline gelmiştir. Görselleştirme de kullanıcı, analizin yönlendirilmesinde nihai otoriteye sahiptir. Bununla birlikte, sistemin belirli görevleri yerine getirmesi için etkin etkileşim araçlarına da ihtiyaç vardır. Görselleştirme özellikle çok boyutlu veri setlerinin analizinde büyük öneme sahiptir. Çünkü görselleştirme verideki karmaşık ilişkileri keşfetmemize ve anlamamıza yardımcı olmaktadır. Büyük veri setlerini tek başına analiz etmek hem önemli hem de zorken, bu teknikle aynı anda birden fazla veri seti dikkate alındığı için pek çok sorunun üstesinden kolayca gelinir. Bugün, VisualCue ve veri görselleştirme metodları kullanılarak büyük miktardaki veriyi resim, diyagram ve renklere dönüştürmek mümkündür[132].

[132] WANG Lidong – Guanghui WANG – Cheryl Ann ALEXANDER, *"Big Data and Visualization: Methods, Challenges and Technology Progress"*, Digital Technologies, Vol. 1, No. 1, 27 June 2015, pp. 33-38.

ÜÇÜNCÜ BÖLÜM
KUVVET YASASI DAĞILIMI*

Genel olarak, kuvvet yasası bir olayın frekansının olayın artan boyutundan daha hızlı azaldığını tanımlar[133]. Bu yasa, farklı yönleri ve özellikleri dikkate alan: uzun kuyruk, 80/20 kuralı ve Pareto ilkesi gibi isimlerden biri ile de anılmaktadır. Teknik olarak, bir "kuvvet yasası", birinin diğerinin gücü olarak değişen iki büyüklük (nicelik) arasındaki matematiksel ilişkiyi ifade etmektedir. Bu yasanın nasıl işlediğini anlamak için dünyadaki şehir büyüklüklerini düşünün: birkaç mega şehir (Tokyo, Mexico City, New York, Londra, İstanbul vb.), yüzlerce modern orta büyüklükteki şehir, binlerce küçük şehir ve yüz binlerce köy vardır. Veya Twitter takipçilerini düşünün: on milyonlarca takipçisi olan birkaç kişi, milyonlarca ya da yüz binlerce takipçisi olan çok daha fazla kişi ve birkaç yüz takipçiye veya daha az takipçiye sahip büyük bir kitle vardır. Burada ki dağılımların her ikisi de kuvvet yasasına uygun olup, Şekil 3.1'dekine benzer bir dağılımı gösterir[134].

Şekil 3.1: Kuvvet Yasası Dağılımı
Kaynak: https://www.efrontlearning.com/blog/wp-content/uploads/2014/11, (20.06.2017).

*Power Law Distribution

[133] HINCHCLIFFE Dion, *"Twenty-two power laws of the emerging social economy"*, 5 October 2009, http://www.zdnet.com/article/twenty-two-power-laws-of-the-emerging-social-economy/, (20.06.2017).

[134] ANDRIOTIS Nikos, *"The power-law distribution (and you)"*, 2014, https://www.efrontlearning.com/blog/2014/11/the-power-law-distribution.html, (20.06.2017).

Kuvvet yasalarını anlama da bir diğer yöntem, 80/20 kuralıdır. Pareto ilkesi olarakta bilinen 80/20 kuralı, belirli bir olayın %80'inin toplam gözlemlerin %20'si tarafından açıklanabileceğini gösteren istatistiksel bir veri dağılımıdır[135]. İtalyan ekonomist ve matematikçi Vilfredo Pareto, 1897 yılında yaptığı çalışmada İtalya'daki servetin %80'ine nüfusun %20'sinin[136]; İngiltere'deki toprakların %80'ine İngiltere'deki nüfusun %20'sinin sahip olduğunu ve gelir dağılımına ilişkin daha sonra kendi ülkesi İtalya dâhil diğer ülkelerdeki araştırmaları incelediğinde bu oranların genel olarak aynı çıktığını gözlemledi[137]. Pareto daha sonra, bahçesinde yetiştirdiği bezelye tohumlarının %20'sinin, ürünün %80'ini verdiğini belirlemesiyle birlikte bu araştırmalarından önemli azınlık ile önemsiz çoğunluğa yönelik matematiksel bir modelin var olabileceğini keşfetti[138] ve ulusal servete yönelik dağılım teorisini içeren modelini, "Cours d'Economie Politique" isimli eserinde yayınladı[139].

Kuvvet yasası dağılımlarının bir diğer ilginç sonucu "uzun kuyruk" olarak adlandırılmasıdır. Daha küçük frekanslara sahip çok sayıda öğenin bulunduğu bu dağılımın sonu uzundur. E-öğrenme pazarına bir örnek vererek, bu kavram daha net hâle getirilebilir. Bilindiği gibi öğrencilerin çoğu sadece birkaç popüler dersi (örneğin; İşletme yönetimi, programlama, vb.) alır ancak gittikçe daha küçük kitlelere hitap eden yüzlerce ders daha vardır (örneğin; Soyut Matematik veya Saz çalma). Bunun gibi sayısız öğrenciyi hedef alan pek çok ders "uzun kuyruk"lu bir dağılıma sahiptir[140].

3.1. Tanımlar

Bilim insanları, atomun veya parçacıklarının durumlarını, hayvanların, bitkilerin veya bakterilerin popülasyonlarını, borsadaki hissesenedi fiyatlarını veya İnternet üzerinden gönderilen mesajların varış süreleri gibi ampirik niceliklerin istatistiksel dağılımlarını gözlemleyerek birçok şey öğrenmektedir. Sözkonusu olan bu niceliklerin çoğu, belirli bir ortalama değerin etrafında kümelenmiş bir dağılıma sahiptir. Diğer bir

[135] INVESTOPEDIA, *"80 – 20 Rule",* 2017, http://www.investopedia.com/terms/1/80-20-rule.asp, (20.06.2017).
[136] LAURA Peters, *"What is happining to the 80/20 rule?"*, Semiconductor International, 25 (12): 17, 2002.
[137] TATIKONDA L. U. – D. O'BRIEN – R. J. TATIKNDA, "*Succeeding with 80/20 rule. Management Accounting"* (February) 1999, ss. 40-44.
[138] DEIRIC Mccann, *"80-20 vision"* Dairy Industries International, 66 (9), 2001, syf. 25.
[139] CRAFT Ralph – Charles LEAKE, *"The Pareto principle in organizationel decision making"* Management Decision 40 (8), 2002, ss. 729-733.
[140] ANDRIOTIS Nikos, *"The power-law distribution (and you)"*, 2014, https://www.efrontlearning.com/blog/2014/11/the-power-law-distribution.html, (20.06.2017).

deyişle, bu dağılımlar, ortalamadan uzakta olası bir ihtimali bulundurmakta ve dolayısıyla ortalama, bu durumda gözlemlerin çoğunu temsil eder. Örneğin, çoğu Amerikalı yetişkin erkeğin yaklaşık 180 cm uzunluğunda olduğunu söylemek mümkündür. Çünkü hiç kimse bu ortalamanın çok ötesinde bir sapma göstermez. Hatta son derece nadir büyük sapmalarda bile, her iki yönde ortalamadan yaklaşık iki birimlik bir sapma gözlenir ve basit bir standart sapma söz konusu olmaktadır. Diğer taraftan, tüm dağılımlar normal dağılım modeline uymaz ve buna sebep olan değerler genelde ortalama ve standart sapma ile ifade edilmediği için sorunlu veya kusurlu olarak kabul edilir. Bu sorunlu değerler, aynı zamanda tüm bilimsel gözlemlerin en ilginç olanlarıdır. Bu ilginç gözlem değerleri, diğer gözlem değerleri gibi basitçe ifade edilemediğinden, daha çok çalışmayı gerektiren karmaşık temel süreçlere işaret etmektedir[141].

Dağılımlar arasında, kuvvet yasası bazen şaşırtıcı fiziksel sonuçlara yol açan matematiksel özellikleri ve doğa ve insan yapımı olguların çeşitli görünümleri yıllar boyunca özel bir ilgi uyandırmıştır. Örneğin, güneş ışınlarının boyutları, şehir nüfusu ve deprem şiddetleri, protesto ve çatışma şiddeti gibi büyüklüklerin dağılımlarının kuvvet yasasına uyduğu düşünülmektedir. Bu gibi büyüklükler ortalamalar ile ifade edilemezler. Yine, örneğin, 2000 yılında yapılan ABD Nüfus Sayımı'na göre, Birleşik Devletler'deki bir şehir, kasaba veya köyün ortalama nüfusu 8226'dır. Fakat bu ortalama değer çoğu zaman karar vermek için doğru değildir. Çünkü toplam nüfusun önemli bir kısmı, nüfusu katlanarak artan mega şehirlerde (New York, Los Angeles, vb.) yaşamaktadır[142].

Matematiksel olarak, eğer x bir olasılık dağılımından çekiliyorsa, bu takdirde x büyüklüğü bir kuvvet yasasına uygundur.

$$p(x) = x^{-\alpha} \tag{3.1}$$

Burada α, üs veya ölçekleme parametresi olarak bilinen dağılımın sabit bir parametresidir. Ölçekleme parametresi, bazı istisnalar hariç genellikle $2 < \alpha < 3$ aralığında yer almaktadır.

[141] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.
[142] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Uygulamada, birkaç ampirik fenomen (olgu), x'in tüm değerleri için kuvvet yasasına uymaktadır. Genellikle kuvvet yasası, sadece bazı x_{min}'den daha büyük olan değerler için geçerlidir. Bu gibi durumlarda, dağılımın kuyruğu bir kuvvet yasasını belirtmektedir.

Bu bölümde, literatürde çokça kullanılan kuvvet yasası ile karşılaştığımızda nasıl ayırt edilebileceği anlatılmaktadır. Pratikte, gözlemlenen bir niceliğin, bir kuvvet yasası dağılımından çekildiğinden emin olabilmek kısmen mümkündür. Burada söylenebilecek en önemli şey, gözlemlerin, yani x'in (3.1) denklemi biçimindeki bir dağılımdan alındığı hipoteziyle tutarlı olmasıdır. Bazı durumlarda, diğer bazı hipotezler ekarte edilebilir. Bu bölümde, benzer sonuçlara ulaşılmasına imkân sağlayan bir dizi istatistiksel teknik açıklanacaktır. Bunun yanısıra bulunan kuvvet yasalarının parametrelerinin hesaplanma yöntemleri ayrıntılı olarak açıklanacaktır.

Kuvvet yasası dağılımı iki şekilde karşımıza çıkar: sürekli gerçel (reel) sayılar içeren sürekli dağılımlar ve ilgili niceliklerin sadece bir dizi kesikli değer alabildiği kesikli dağılımlar, yani tipik olarak pozitif tamsayı değerleri alabilen dağılımlardır.

x, ilgilendiğimiz dağılımın sayısını göstersin. Bu takdirde, sürekli bir kuvvet yasası dağılımının olasılık yoğunluğu $p(x)$,

$$p(x)dx = \Pr(x \leq X < x + dx) = Cx^{-\alpha}dx \tag{3.2}$$

(3.2)'deki denklem gibi tanımlanabilir. Burada X, gözlemlenen değer ve C bir normalleştirme sabitidir. Denklem (3.2)'deki olasılık yoğunluğu birbirinden uzaklaştıkça $x \to 0$ olur. Bundan dolayı (3.2) denklemi herzaman $x \geq 0$ koşulunu sağlamaz. Bu durumda kuvvet yasası dağılımı için x'in başka bir alt sınırlarının olması gerekmektedir. İşte bu alt sınır x_{min} ile sınırlandırılır. Daha sonra, $\alpha > 1$ şartı sağladığında, normalleştirme sabiti kolayca hesaplanarak,

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} \tag{3.3}$$

(3.3) denklemi elde edilir.

Kuvvet yasası dağılımlı verilerin analizinde izlenecek aşamalar

Verilerin kuvvet yasasına uygunluğunun analiz edilmesinde izlenecek olan aşamalar aşağıdaki gibidir.

1. Maksimum Olabilirlik Tahmincileri (MLEs) kullanılarak kuvvet yasası modelindeki x_{min} ve α parametreleri tahmin edilir.
2. Kolmogorov-Smirnov yöntemi kullanılarak veri ile kuvvet yasası arasındaki uyum iyiliği hesaplanılır. Elde edilen p-değeri 0.05'den büyükse, verilerin kuvvet yasasına uygun olduğu, aksi takdirde uygun olmadığı söylenilir.

Eğer x kesikli ise, bu takdirde x sadece kesikli bir dizi değer alabilir. Eğer olasılık dağılımında sadece tamsayılı değerler kullanılırsa:

$$p(x) = \Pr(X = x) = Cx^{-\alpha} \tag{3.4}$$

(3.4) elde edilir ve bu dağılım sıfırdan farklıdır. Dolayısıyla, kuvvet yasası dağılımı üzerindeki alt sınır $x_{min} > 0$ olmalıdır. Normalleştirme sabiti hesaplanarak:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} \tag{3.5}$$

(3.5) elde edilir. Burada,

$$\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty} (n + x_{min})^{-\alpha} \tag{3.6}$$

olup, (3.6) denklemi genelleştirilmiş ya da Hurwitz zeta fonksiyonu olarak adlandırılmaktadır. Aşağıdaki Tablo 3.1 de birçok dağılım için faydalı olabilecek bazı temel fonksiyonel formlar ve normalleştirme sabitleri verilmiştir.

Genellikle, bir kuvvet yasası dağılımı değişkeninin kümülatif dağılım fonksiyonunu veya CDF'sini de dikkate almak gerekmektedir. Burada ifade edilen $P(x)$, sürekli ve kesikli durumların her ikisi için $P(x) = \Pr(X \geq x)$ olarak tanımlanabilir. Örneğin, sürekli durumda,

$$P(x) = \int_x^{\infty} p(x')dx' = \left(\frac{x}{x_{min}}\right)^{-\alpha+1} \tag{3.7}$$

ve kesikli durumda,

$$P(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{min})} \tag{3.8}$$

olur.

	İsim	Dağılım $p(x) = Cf(x)$	
		$f(x)$	C
Sürekli	Kuvvet yasası (Power law)	$x^{-\alpha}$	$(\alpha-1)x_{min}^{\alpha-1}$
	Kesmeli kuvvet yasası (Power law with cutoff)	$x^{-\alpha}e^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda x_{min})}$
	Üssel (Exponential)	$e^{-\lambda x}$	$\lambda e^{\lambda x_{min}}$
	Gerilmiş üssel (Stretched exponential)	$x^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{min}^{\beta}}$
	Log-normal	$\frac{1}{x}\exp\left[-\frac{(lnx-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[erfc\left(\frac{lnx_{min}-\mu}{\sqrt{2}\,\sigma}\right)\right]^{-1}$
Kesikli	Kuvvet yasası	$x^{-\alpha}$	$1/\zeta(\alpha, x_{min})$
	Yule dağılımı (Yule distribution)	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha-1)\frac{\Gamma(x_{min}+\alpha-1)}{\Gamma(x_{min})}$
	Üssel	$e^{-\lambda x}$	$\left(1-e^{-\lambda}\right)e^{\lambda x_{min}}$
	Poisson	$\mu^x/x!$	$\left[e^{\mu}-\sum_{k=0}^{x_{min}-1}\frac{\mu^k}{k!}\right]^{-1}$

Tablo 3.1: Bazı İstatistiksel Dağılımlar için Temel $f(x)$ Fonksiyonu ve Uygun C Normalleştirme Sabiti
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Tablo 3.1'de kuvvet yasası dağılımı ve bazı istatistiksel dağılımlar için temel $f(x)$ fonksiyonu ve uygun C normalleştirme sabiti verilmiştir. Burada sürekli durum için $\int_{x_{min}}^{\infty} Cf(x)dx = 1$ ve kesikli durumu için $\sum_{x=x_{min}}^{\infty} Cf(x) = 1$ dir.

Sürekli dağılım formülleri, denklem (3.3)'ten de görüleceği üzere, kesikli dağılımlardan daha basittir. Matematiksel olarak, genellikle kesikli ve sürekli kuvvet yasası dağılımları yaygın olarak kullanılmaktadır. Ancak şuna dikkat edilmelidir; kesikli

bir kuvvet yasasını sürekli bir yaklaşımla hesaplamanın birkaç farklı hesaplama yöntemi vardır. Fakat bu yöntemlerin bir kısmı iyi sonuçlar verirken, diğerleri için aynı şey geçerli değildir. Nispeten daha güvenilir yöntemlerden biri de, bir kuvvet yasası için, x değerlerinin sürekli bir kuvvet yasasından üretilmiş gibi ele alınıp, daha sonra x'in en yakın tamsayıya yuvarlanmasıdır. Bu yaklaşım pek çok uygulamada oldukça iyi sonuçlar vermektedir.

3.2. Ampirik Verilere Kuvvet Yasalarının Uydurulması

Araştırmalar kuvvet yasasını belirten amprik dağılımların genellikle ölçekleme parametresi α'nın ve bazen de ölçekleme bölgesi x_{min}'nin alt sınırı için bazı tahminler vermektedir. Bu amaç için en sık kullanılan yöntem basit histogramdır. (3.1) denkleminin her iki tarafının logaritması alınırsa, kuvvet yasası dağılımının $lnp(x) = \alpha lnx + constant$ formuna uyduğunu ve bunun çift logaritmik bir doğru üzerinde düz (doğrusal) bir doğruyu izlediği görülür. Bu nedenle kuvvet yasası dağılımını araştırmanın en yaygın yolu, ilgili x miktarını ölçerek, x'in frekans dağılımını temsil eden bir histogram oluşturmak ve bu histogramı çift logaritmik eksenler üzerinde çizmektir. Eğer bunu yaparken, yaklaşık olarak düz bir doğruya düşen bir dağılım bulunursa, o zaman büyük ihtimalle dağılımın düz doğrusunun mutlak eğimi ile verilen bir ölçekleme parametresi α olan bir kuvvet yasasını belirteceği söylenebilir. Tipik olarak bu eğim, histogramın logaritması üzerinde bir en küçük kareler doğrusal regresyon (least-squares linear regression) yöntemi uygulanarak bulunur. Bulunan bu kuvvet yasası dağılımı, Pareto'nun 19. yüzyılın sonundaki zenginlik dağılımı üzerine kurulmuştur[143].

3.2.1. Ölçekleme Parametresinin Tahmin Edilmesi

Ölçekleme parametresi α'yı tahmin etmek için verideki kuvvet yasası dağılımı için x_{min} alt sınırının bir değerinin olması gerekir. Bir an için bu değerin bilindiğini farzedilsin. Bilinmediği durumda ise bu değer veriden de tahmin edilmektedir.

Kuvvet yasası dağılımları gibi parametreleştirilmiş modelleri gözlemlenen verilere uydurmak için tercih edilen yöntem, büyük örneklem büyüklüğünün sınırında kesin parametre tahminleri yapabilen *maksimum olabilirlik* (maximum likelihood)

[143] ARNOLD Barry C., *"Pareto Distributions"*, International Cooperative Publishing House, Fairland, Maryland USA, 1983.

yöntemidir[144]. Verilerin tam olarak $x \geq x_{min}$ için bir kuvvet yasasına uygun olan bir dağılımdan alındığı varsayılırsa, hem kesikli hem de sürekli durumlarda ölçekleme parametresinin maksimum olabilirlik tahmincileri (maximum likelihood estimators-MLEs)'nden türetilebilir.

Sürekli veri: Veriler sürekli olduğunda, Muniruzzaman tarafından 1957'de[145] türetilen ölçekleme parametresi için maksimum olasılık tahmincisi, bilinen Hill tahmincisine[146] eşdeğerdir. Varsayalım ki sürekli kuvvet yasası dağılımı (3.3)'teki gibi olsun. Burada α, ölçekleme parametresi ve kesme parametresi x_{min}, kuvvet yasası dağılımının bulunduğu minimum değerdir. n gözlemleri $x_i \geq x_{min}$ olan bir veri kümesi göz önüne alındığında, α değerini, verilerin oluşturmuş olma olasılığı en yüksek olan kuvvet yasası modeli için görmek istenilir. Verilerin modelden alınma olasılığı,

$$p(x/\alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} \tag{3.9}$$

olur. Bu olasılık, modeldeki verinin *olabilirliği (likelihood)* olarak adlandırılır. Verilerin büyük olasılıkla, ölçekleme parametresi (α), model tarafından üretilmektedir. Bu da fonksiyonun maksimum olmasını sağlamaktadır. Genellikle, α'nın maksimum olduğu yerde ise olabilirlik algoritması olan $\mathcal{L}$ ile çalışılır.

$$\begin{aligned} \mathcal{L} = lnp(x \setminus \alpha) &= ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \left(\frac{x_i}{x_{min}}\right)^{-\alpha} \\ &= \sum_{i=1}^{n} \left[\ln(\alpha - 1) - lnx_{min} - \alpha ln \frac{x_i}{x_{min}}\right] \\ &= n \ln(\alpha - 1) - n\, lnx_{min} - \alpha \sum_{i=}^{n} ln \frac{x_i}{x_{min}} \end{aligned} \tag{3.10}$$

olur. $\partial \mathcal{L} \,/\, \partial \alpha = 0$ seçilir ve α için çözüm yapılırsa, ölçekleme parametresi için *maksimum olabilirlik tahmini (maximum likelihood estimate)* veya MLE[147]:

---

[144] LARRY Wasserman, *"All of Statistics: A Concise Course in Statistical Inference"*, (SpringerVerlag, Berlin, ISBN 978-0-387-21736-9, 2003.

[145] MUNIRUZZAMAN A. N. M., *"On Measures of Location and Dispersion and Test of Hypotheses in a Paretto Population"*, Vol. 7, İssue:3, Page(s):115_123, Issue published: 1 July 1957, page(s):115-123.

[146] HILL Bruce M., *"A Sımple General Approch to Inference about the Tail of a Distribution"*, The Annals of Statistics, Vol. 3, No. 5, 1975, page(s):1163-1174.

[147] MUNIRUZZAMAN A. N. M., *"On Measures of Location and Dispersion and Test of Hypotheses in a Paretto Population"*, Vol. 7, İssue:3, Page(s):115_123, Issue published: 1 July 1957, page(s):115-123.

$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} ln\frac{x_i}{x_{min}}\right]^{-1} \tag{3.11}$$

dir. Denklem (3.11)'deki $x_i,\ i = 1 \ldots n$ olmak üzere, öyle ki x'in tüm gözlenen değerleri için $x_i \geq x_{min}$ şeklindedir. Denklem (3.11) de ve başka yerlerde veriden türetilmiş tahminleri belirtmek için $\hat{\alpha}$ gibi "şapkalı" semboller kullanılacaktır. Şapkasız semboller ise pratikte genellikle bilinmeyen gerçek değerleri ifade etmektedir. Denklem (3.11), iyi bilinen Hill tahmincisine eşit olup,[148] asimptotik olarak normal[149] ve tutarlı[150](yani; n en büyük değerini aldığında $\hat{\alpha} \to \alpha$ olur.)'dır. En çok benzerlik oranı ile elde edilen $\hat{\alpha}$'daki standart hata maksimum:

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + 0\left(\frac{1}{n}\right) \tag{3.12}$$

dır. Denklem (3.12)'deki yüksek dereceli korelesyon pozitiftir. ($\alpha \leq 1$ olan dağılımlar normalleştirilemediğinden ve bundan dolayı doğada oluşamayacağından bu hesaplamalar $\alpha > 1$ varsayımı altında yapılır. Bir olasılık dağılımının, x aralığı bazı kesmeler ile sınırlandırılmış olması halinde, dağılım $x^{-\alpha}$'ya giderken $\alpha \leq 1$ dir. Ancak böyle bir dağılım için başka maksimum olabilirlik tahmincilerine ihtiyaç duyulur.)

Kesikli (Discrete) veri: Eğer x kesikli, bir tamsayılı değişken ise bu takdirde MLE yöntemini anlamak daha zordur[151]. Bundan dolayı özel olarak $x_{min} = 1$ durumu ele alınırsa, α için uygun çözümün transandantal denklemi:

$$\frac{\zeta'(\hat{\alpha})}{\zeta(\hat{\alpha})} = -\frac{1}{n}\sum_{i=1}^{n} lnx_i \tag{3.13}$$

olur. Denklem (3.13), $x_{min} > 1$ koşulunu sağlamaktadır. Ancak (3.13) denklemindeki zeta fonksiyonları genelleştirilmiş zetalar ile değiştirilerek[152]:

$$\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} = -\frac{1}{n}\sum_{i=1}^{n} lnx_i \tag{3.14}$$

---

[148] HALL Peter, *"On some Simple Estimates of an Exponent of Regular Variation"*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 44, No. 1, 1982, pp. 44-37.

[149] MASON David M., *"Law of Large Numbers for Sum of Extreme Values"*, The Annals of Probability Vol. 10, No. 3, Aug. 1982, pp. 754-764.

[150] HILL Bruce M., *"A Sımple General Approch to Inference about the Tail of a Distribution"*, The Annals of Statistics, Vol. 3, No. 5, 1975, page(s):1163-1174.

[151] SEAL H. L., *"Maximum Likelihood Fitting of the Discrete Pareto Law"*, Journal of the Institute of Actuaries, Vol. 78, Issue 1, June 1952, pp. 115-121.

[152] CLAUSET Aaron – Maxwell YOUNG – Kristian Skrede GLEDITSCH, *"On the Frequency of Severe Terrorist Events"*, Journal of Conflict Resolution, Vol. 51, Number 1, 2007, pp. 58-87.

(3.14) elde edilir. Uygulamada, $\hat{\alpha}$'nın değerlendirilmesi için (3.14) denkleminin nümerik olarak çözülmesi gerekmektedir. Alternatif olarak, olasılık fonksiyonunun doğrudan sayısal maksimizasyonu ile veya onun algoritmasının eşdeğeri ile α aşağıdaki gibi tahmin edilebilir:

$$\mathcal{L}(\alpha) = -n\, ln\zeta(\alpha, x_{min}) - \alpha \sum_{i=1}^{n} lnx_i \tag{3.15}$$

Kesikli durumda, $\hat{\alpha}$ standart hatasına ilişkin bir tahmin bulmak için, log-likelihood ikinci dereceden (quadratic) bir yaklaşım uygulanır, maksimum değer ve olası tahmin için Gaussian formunun standart sapması hata tahmini olarak alınmaktadır. Sonuçta standart sapma,

$$\sigma = \frac{1}{\sqrt{n\left[\frac{\zeta''(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} - \left(\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})}\right)^2\right]}} \tag{3.14}$$

olur ve $\hat{\alpha}$'yı bulduktan sonra hesaplama oldukça basitleşmektedir. Alternatif olarak, denklem (3.12) yeterli büyüklükteki n ve x_{min} için benzer sunuçları vermektedir.

Kesikli durumda $\hat{\alpha}$ için tam kapalı form ifadesi olmasa da, yaklaşık ifadesi, daha önce bölüm 3.2'de bahsedilen yaklaşım kullanılarak tahmin edilebilir. Yani, bu yaklaşımda gerçekte kuvvet yasası ile dağılmış tamsayılar, sürekli reel sayıların kendisine en yakın tamsayıya yuvarlanması ile elde edilmiştir.

Kesikli kuvvet yasasının ölçekleme parametresi için yaklaşık tahminci

$f(x)$ diferansiyellenebilir bir fonksiyon, $F(x)$ sonsuz integrallenebilir ve öyle ki $F'(x) = f(x)$ ise, bu takdirde,

$$\begin{aligned}\int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f(t)dt &= F\left(x+\frac{1}{2}\right) - F\left(x-\frac{1}{2}\right) \\ &= \left[F(x) + \frac{1}{2}F'(x) + \frac{1}{8}F''(x) + \frac{1}{48}F'''(x)\right] + \\ &\quad -\left[F(x) - \frac{1}{2}F'(x) + \frac{1}{8}F''(x) - \frac{1}{48}F'''(x)\right] + \cdots \\ &= f(x) + \frac{1}{24}f''(x) + \cdots\end{aligned} \tag{3.15}$$

olur. (3.15) denklemini x tamsayısı üzerinden toplarsak, aşağdaki,

$$\int_{x_{min}-\frac{1}{2}}^{\infty} f(t)dt = \sum_{x=x_{min}}^{\infty} f(x) + \frac{1}{24}\sum_{x=x_{min}}^{\infty} f''(x) + \cdots \tag{3.16}$$

(3.16) elde edilir. Örneğin, bazı α sabiti içi $f(x) = x^{-\alpha}$ ise, bu takdirde,

$$\int_{x_{min}-\frac{1}{2}}^{\infty} t^{-\alpha} dt = \frac{\left(x_{min} - \frac{1}{2}\right)^{-\alpha+1}}{\alpha - 1}$$

$$= \sum_{x=x_{min}}^{\infty} x^{-\alpha} + \frac{\alpha(\alpha+1)}{24} \sum_{x=x_{min}}^{\infty} x^{-\alpha-2} + \cdots \tag{3.17}$$

$$= \zeta(\alpha, x_{min}) \left[1 + O\left(\frac{1}{x_{min}^2}\right)\right]$$

(3.17) elde edilir. Denklem (3.17) de ikinci toplamdaki tüm veriler için $x^{-2} \leq x_{min}^{-2}$ koşulu kullanılmıştır. Böylece,

$$\zeta(\alpha, x_{min}) = \frac{\left(x_{min} - \frac{1}{2}\right)^{-\alpha+1}}{\alpha - 1} \left[1 + O\left(\frac{1}{x_{min}^2}\right)\right] \tag{3.18}$$

(3.18) elde edilmiştir. Denklem (3.18), kesikli kuvvet yasasının ölçekleme parametresi α için maksimum olabilirlik tahmincisine bir yaklaşım elde etmek için kullanılabilir. Ayrıca (3.18) denklemi (3.14)'teki x_{min} çok büyük olduğunda geçerlidir. Böylece, (3.14) denklemindeki zeta fonksiyonlarının oranı,

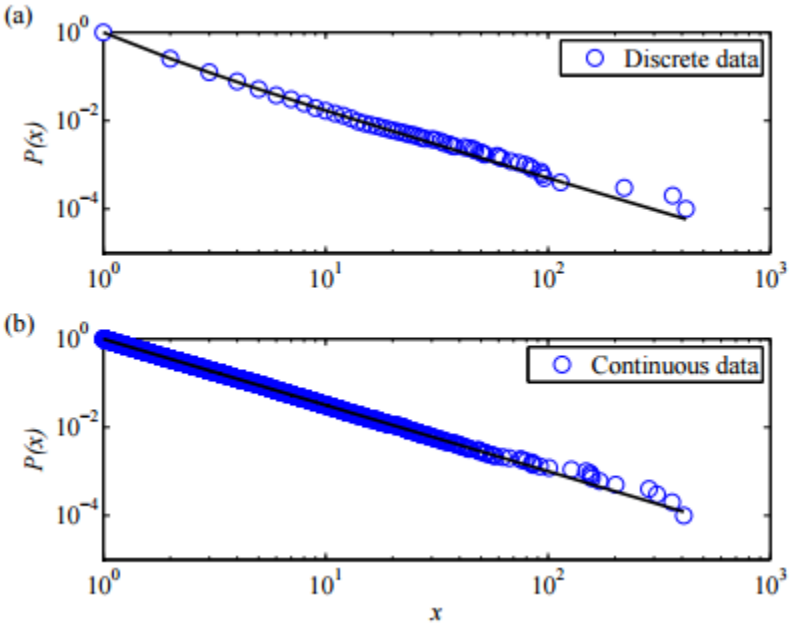
$$\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} = -\left[\frac{1}{\hat{\alpha} - 1} + \ln(x_{min} - \frac{1}{2})\right]\left[1 + O\left(\frac{1}{x_{min}^2}\right)\right] \tag{3.19}$$

(3.19) hâline gelmektedir. (3.19) denklemindeki x_{min}^2'nin mertebesinin büyüklüğü 1'in mertbesinin büyüklüğü ile kıyaslandığında çok küçük olduğundan ihmal edilir ve denklem,

$$\hat{\alpha} \backsimeq 1 + n\left[\sum_{i=1}^{n} ln \frac{x_i}{x_{min} - \frac{1}{2}}\right]^{-1} \tag{3.20}$$

(3.20) elde edilir. Dikkat edilirse burada, kesikli durumda sürekli durumdan farklı olarak sadece payda da $-\frac{1}{2}$ vardır. Diğer taraftan, $-\frac{1}{2}$ dışında geri kalan sürekli ve kesikli durum için MLE aynıdır. Sayısal olarak, (3.20)'deki tam kesikli MLE ile (3.14)'teki sürekli MLE denklemleri karşılaştırıldığında, $x_{min} \gtrsim 6$ olduğunda (3.20) denklemi daha iyi sonuçlar vermektedir. Ayrıca, (3.20) ifadesi, tam kesikli MLE'den daha kolay değerlendirilebilir ve yüksek doğruluk gerekmeyen durumlarda yararlı olabilmektedir. Pratikte (3.20)'deki tahminci yaklaşık %1 anlamlılık düzeyinde veya $x_{min} \gtrsim 6$ olduğunda oldukça iyi sonuçlar vermektedir.

Bazı araştırmacılar ise ortaya başka bir yaklaşım atmışlardır. Bunlar kesikli verinin sürekli olduğunu iddia ederek $\hat{\alpha}$'yı hesaplamak için sürekli veriler için (3.11) denklemindeki MLE'yi kullanmışlardır. Fakat bu yaklaşım, denklem (3.20)'deki $\hat{\alpha}$'ya göre daha iyi sonuç vermediğinden ve uygulanmasının kolay olmaması nedeniyle çok fazla tavsiye edilmemektedir.



Şekil 3.2: Kesikli ve Sürekli Veriler için Kümülatif Yoğunluk Fonksiyonları
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Şekil 3.2'deki mavi noktalar, dağıtılan sentetik veri kümeleri için kümülatif yoğunluk fonksiyonlarını ($P(x)$ fonksiyonunu) temsil etmektedir. (a) bir kesikli kuvvet yasasını ve (b) bir sürekli kuvvet yasasını göstermektedir. Ayrıca, her iki grafik için de $\alpha = 2.5$ ve $x_{min} = 1$ dir. Grafiklerdeki düz çizgiler ise verilere en iyi uyumu temsil etmektedir[153].

3.2.2. Ölçekleme Parametresi Tahmcilerinin Performansı

Bu kısımda, yukarıda açıklanan tahmincilerin çalışmasını göstermek için, sentetik kuvvet yasası verilerinin bilinen ölçekleme parametrelerinin tahmin gücü test edilecektir. Pratikte, yapılan hesaplamalarda verilerin kuvvet yasası ile dağıldığı önceden bilinmemektedir. Bu durumda, MLE'ler bize uyumların yanlış olduğuna dair herhangi bir

[153] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

bilgi vermemekte: kuvvet yasasının, veriler için iyi bir model olup olmadığını söylemezler, sadece verilerin kuvvet yasasına uygunluğunu söylerler.

Kuvvet yasası, dağıtılmış rassal sayılar üretme yöntemi kullanarak, her biri $\alpha = 2.5, x_{min} = 1$ ve $n = 10.000$ olan sürekli ve kesikli iki kuvvet yasası dağılımını veriler üretsin. MLE'ler rassal olarak üretilen bu veriye uygulandığında, sürekli durumda $\hat{\alpha} = 2.50(2)$ ve kesikli durumda $\hat{\alpha} = 2.49(2)$ olarak hesaplanır. Bu tahminler, verilerin üretildiği bilinen gerçek ölçekleme parametresiyle iyi bir uyum içindedir. Tablo 3.3'te tahmini parametreler kullanarak iki veri setinin uyumu ile birlikte dağılımları gösterilmiştir. Genellikle, CDF (Cumulative Density Function)'nin görsel biçimi, özellikle dağılımın kuyruğu sonlu örneklem büyüklüklerine bağlı olarak dalgalanmalara karşı PDF (Probability Density Function)'den daha uzundur[154].

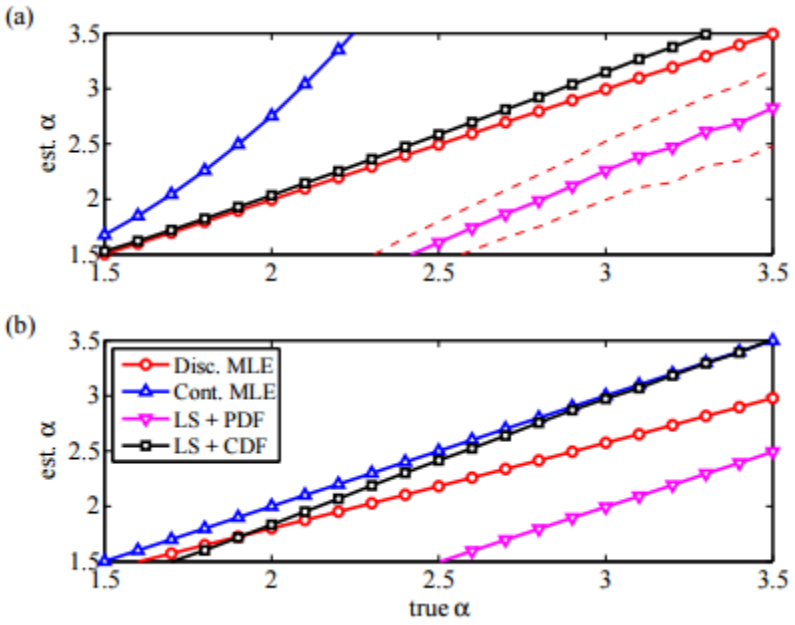
Metod	Notlar	$\hat{\alpha}$ (kesikli)	$\hat{\alpha}$ (sürekli)
LS+ PDF	sabit genişlik (const. width)	1,5(1)	1,39(5)
LS+ CDF	sabit genişlik (const. width)	2,37(2)	2,480(4)
LS+ PDF	log. genişlik (log. width)	1,5(1)	1,19(2)
LS+ CDF	rank-freq.	2,570(6)	2,4869(3)
cont. MLE	–	4,46(3)	**2,50(2)**
disc. MLE	–	**2,49(2)**	2,19(1)

Tablo 3.2: Kesikli ve Sürekli Sentetik Veriler Kullanılarak Tahmin Edilen Ölçekleme Parametresi Değerleri
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Tablo 3.2'de tahmin edilen ölçekleme parametresi değerleri $\alpha = 2.5,\ x_{min} = 1$ ve $n = 10\,000$ olmak üzere, kesikli ve sürekli sentetik veriler için çeşitli tahmincilerin kullanılmasıyla bulunan değerlerdir. Tablo 3.2'de MLE ile verilen sonuçlar, lineer regresyona dayalı birkaç alternatif yöntem kullanarak tahmin edilen ölçekleme parametreleri ile karşılaştırılmıştır. Bu yöntemler, log dönüşümlü histogram eğimine düz çizgi uydurma, "logaritmik gruplar" ile bir histogramın eğimine uyma (gruplama genişliği x ile orantılı olup, bu sayede histogramın kuyruğundaki dalgalanmalar azalır), sabit genişlikteki gruplar ile hesaplanan CDF eğimine uyması ve herhangi bir grup olmaksızın hesaplanan CDF eğimine uymasıdır. Tablo 3.2'de görüldüğü üzere, MLE'ler en iyi

[154] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

sonucu verirken, regresyon yöntemleri tümüyle yanlı sonuçlar vermektedir. Belki de bu yöntemler içinde kesikli durumda yanlı tahmin üreten, ancak sürekli durumda makul ölçüde iyi sonuç veren CDF'leri ayrı tutmak gerekmektedir. Dahası, tahminin yanlı olduğu her durumda, hata tahmini yanlılığa dair herhangi bir uyarı vermez. Bu durumda bizleri, sonuçların önemli ölçüde hatalı olduğu konusunda uyaracak hiçbir bilgi yoktur. Şekil 3.3'te $n = 10\,000$ gözlemli bir sentetik veri seti için tahmincilerin ($\hat{\alpha}$), gerçek α'nın bir fonksiyonu olarak aldıkları değerler gösterilmiştir[155].



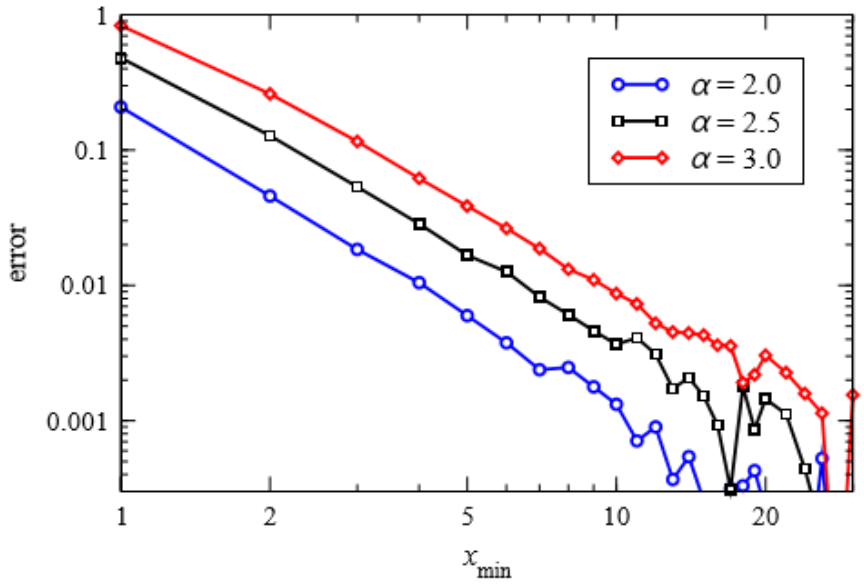
Şekil 3.3: Ölçekleme Parametresi Tahmin Değerleri
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Şekil 3.3'teki tahmini ölçekleme parametresi değerleri Tablo 3.2'deki dört yöntem kullanılarak tahmin edilmiştir (PDF için logaritmik gruplara dayanan yöntemler ve CDF için sabit genişlikli grupları temel alan yöntemler hariç). Şekil 3.3'te $n = 10\,000$ gözlem için $x_{min} = 1$ olan (a) kesikli ve (b) sürekli kuvvet yasası dağılımları için verilmiştir[156].

Son olarak maksimum olabilirlik tahmincilerinin yalnızca büyük örneklem büyüklüğünün asimptotik limitinin, $n \to \infty$ iken $O(\frac{1}{n}) \to 0$ olup, yansız olduğu garanti edilir. Diğer taraftan, sonlu veri setleri için yanlılık mevcuttur. Fakat seçilecek herhangi

[155] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.
[156] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703

bir x_{min} değeri ile $O(\frac{1}{n})$ yanlılığı ortadan kalkmaktadır. Genellikle çok küçük veri kümeleri için bu yanlılık önemli olabilmektedir. Ancak bazı durumlarda göz ardı edilebilir, çünkü $O(\frac{1}{n^2})$ olduğunda yanlılığı ortadan kalkan tahmincinin istatistiksel hatası çok daha küçük olmaktadır. Genellikle $n \geq 50$ olduğunda parametre tahminleri yapmak için örneklem büyüklüğü yeterli kabul edilmektedir. Şekil 3.4'te gösterilen örnekler için, α'nın %1'e kadar ki kesin tahminleri verilmiştir. Bundan daha küçük veri setleri ($n \geq 50$) ile çalışılıyorsa dikkatli olmak gerekmektedir. Bununla birlikte, küçük veri setlerini dikkatli kullanmak için daha önemli sebeplerin de olduğu unutulmamalıdır. Yani, gerçekte veriler, kuvvet yasası ile dağılmış olsa bile, bu tür verilere alternatif uyumları gözardı etmemek gerekir ve tersine veriler kuvvet yasasından farklı olan bir dağılımdan alınsa bile bu veriler yine de kuvvet yasasına uygun bir dağılım sergileyebilmektedir[157].



Şekil 3.4: Ölçekleme Parametresi Tahminindeki Hata Değerleri
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Şekil 3.4'te kesikli veriler için yaklaşık MLE kullanılarak, elde edilen tahmini ölçekleme parametresindeki $\hat{\alpha}$ hatası değerleri verilmiştir. Bu hata, (3.20) denkleminde $\alpha = 2,\ 2{,}5$ ve 3 için (1000 tekrarlama için) x_{min}'nin bir fonksiyonu olarak verilmiştir.

[157] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

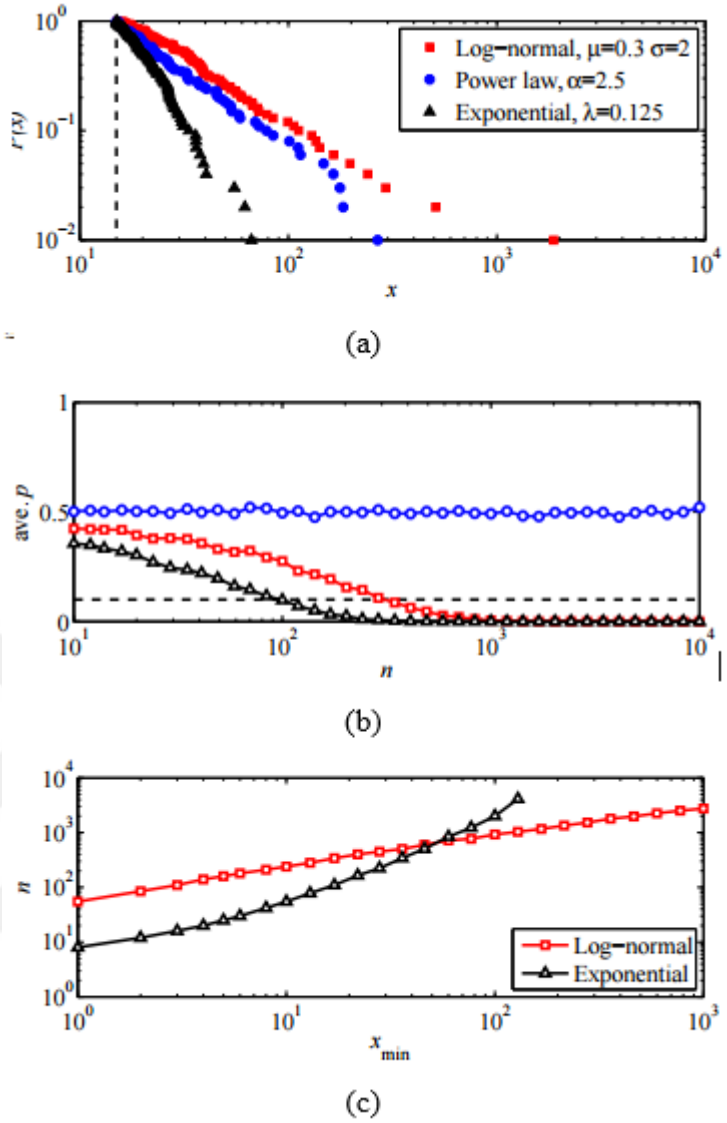
Ortalama hata $O(\frac{1}{x_{min}^2})$ olarak azalır ve $x_{min} \gtrsim 6$ olduğunda α değeri % 1'den daha küçük olmaktadır[158].

3.3. Kuvvet Yasası Hipotezinin Test Edilmesi

Önceki bölümlerde açıklanan yöntemler belirli bir veri kümesine bir kuvvet yasası dağılımı uydurmaya ve parametre tahminleri yapılmasına izin vermekteydi. Bununla birlikte, bu tahminler kuvvet yasasının verilere makul bir uyum gösterip göstermediği konusunda hiçbir şey söylemez. Verilerin gerçek dağılımı ne olursa olsun, her zaman bir kuvvet yasasına uyabilmektedir. Uyumun veriyle iyi uyuşup uyuşmadığını anlamak için bazı yöntemlere ihtiyaç duyulmaktadır.

Görünür de kuvvet yasası ile dağılan verilere ilişkin daha önceki ampirik çalışmaların çoğu, kuvvet yasası hipotezini nicel olarak test etmeye çalışmamıştır. Bunun yerine, tipik olarak, görselleştirmeler temelinde, verilerin nitel değerlendirmelerine güvenilmektedir. Ancak bunlar yanıltıcı olabilir ve daha ayrıntılı incelenmesi gereken kuvvet yasassı iddialarını çürütebilmektedir. Şekil 3.5a, $\alpha = 2{,}5$ olan bir kuvvet yasası dağılımının sırasıyla ortalaması ve standart sapması $\mu = 0{,}3$ ve $\sigma = 2{,}0$ olan bir log-normal dağılımı ve üssel (exponential) parametresi $\lambda = 0{,}125$ olan üssel bir dağılım gösterilmiştir. Biran için bu dağılımların CDF'lerin küçük veri kümelerinden $(n = 100)$ oluştuğu düşünülsün. Her durumda, dağılımların alt sınırı $x_{min} = 15$'e sahiptir. Çünkü bu dağılımların her biri, şekil içinde kullanılan log-log grafiğinde kabaca düz görünmektedir. Dağılımların her biri, dikkatle incelendiğinde, her üçünün de, farklı ölçekleme parametreleri olsa da, kuvvet yasasına uygun bir dağılım sergiledikleri sonucuna varılabilir. Bununla birlikte, log-log grafik üzerinde eğrinin sol yukarıdan sağ aşağıya düz bir çizgi şeklinde olması gerekirken, fakat bu kuvvet yasası için yine de yeterli bir koşul değildir[159].

[158] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.
[159] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.



Şekil 3.5: Log-normal, Kuvvet Yasası ve Üssel Dağılım
Kaynak: CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

Kısacası, verilerin dağılımının sırası ile üssel dağılıma veya log-normal dağılıma uygun olup olmadığını uyumun iyiliği testlerinden, Kolmogorov Smirnov veya Anderson Darling testi ile test etmek mümkündür. Bu testler, ileri sürülen model ile amprik verilerin dağılımı arasındaki uzaklık ölçüsüne dayanmaktadır. Bahsi geçen testlerde ileri sürülen modelden üretilen benzer yapay veri setleri ile ampirik veriler karşılaştırılır. Karar verme aşamasında hesaplanan test istatistiğinin $p-$değeri ile seçilen anlamlılık düzeyi karşılaştırılır. Seçilen anlamlılık düzeyine $(0{,}05\ veya\ 0{,}01)$ göre hesaplanan test

istatistiğinin $p-$değerine bakılarak karar (H_0: $hipotezi\ ya\ red\ edilir\ ya\ da\ red\ edilmez$) kılınır[160].

3.3.1. Uyum İyiliği Testleri

Genellikle, elde edilen veri seti ve verilerin uyduğu varsayıldığı bir kuvvet yasası dağılımı göz önüne alındığında, verilere göre varsayımımızın makul olup olmadığından emin olmak isteriz. Bu tür bir sorun için standart bir yaklaşım, hipotezin makul olup olmadığını ölçen bir $p-değeri$ üreten bir *uyum iyiliği testi* kullanmaktır. Bu testler, ampirik verilerin dağılımı ile varsayımlanan model arasındaki "uzaklık" ölçüsüne dayanmaktadır. Bu uzaklık, aynı modelden alınan karşılaştırılabilir sentetik veri setleri için uzaklık ölçümleri ile karşılaştırılır. Buarda $p-değeri$, ampirik uzaklıktan daha büyük olan sentetik uzaklıkların fonksiyonu olarak tanımlanır. Eğer $p-değeri$ büyük (1'e yakın) ise verilerin kuvvet yasası dağılımından geldiği söylenilebilir. Eğer $p-değeri$ küçük (1'e uzak) ise verilerin kuvvet yasası dağılımından gelmediği söylenebilir[161].

Her sentetik veri kümesi için, Kolmogorov-Smirnov (KS) istatistiğini, veri kümesinin gösterdiği orijinal dağılıma değil, o veri kümesindeki en uygun kuvvet yasasına göre hesaplamak önemlidir. Eğer $p-değerinin$ yansız bir tahmin elde etmesi isteniliyorsa, gerçek veri seti için yapılan hesaplamaların herbir sentetik veri seti için de yapılması gerekmektedir.

Sentetik verilerin üretilmesinde bazı önemli ayrıntılar vardır. Doğru p tahminleri elde etmek için, x_{min}'den küçük ampirik verilere benzer bir dağılıma sahip sentetik verilere ihtiyaç vardır. Ancak bu veriler, x_{min}'deki kuvvet yasasına uygun olmalıdır. Sentetik verileri üretmek için, semiparametrik (semiparametric) bir yaklaşım kullanılmaktadır. Burada gözlenen veri setimizin n_{tail} gözlemi $x \geq x_{min}$ ve toplam gözlem sayısı n'dir. Buna göre, n gözlemli bir veri seti aşağıdaki gibi üretilmektedir. n_{tail}/n olasılığı ile ölçekleme parametresi $\hat{\alpha}$ ve $x \geq x_{min}$ olan bir kuvvet yasasından rassal olarak bir x sayısı üretilir. Diğer taraftan, $1 - n_{tail}/n$ olasılığı ile $x < x_{min}$'e sahip

[160] TÜZÜNTÜRK Selim, *''Firmalarda Organizasyonel Ağ Analizi Ve Bir Uygulama''*, Doktora Tezi, Bursa, 2012, s.75.
[161] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

olan ve gözlemlenen veri kümesinden her seferinde eşit olasılıkla rassal olarak tek bir x_i elemanı seçilir. Bu seçim $i = 1 \ldots n$'e kadar tekrarlanılırsa, x_{min} üzerinde bir kuvvet yasasını izleyen, ancak gözlenen veriler ile aynı (kuvvet yasası olmayan) dağılıma sahip olan bir sentetik veri seti üretilir.

Ayrıca, kaç tane sentetik veri seti üretileceğine karar vermekte önemlidir. En kötü test performansının analizine dayanarak kaç tane sentetik veri setinin üretileceği şu şekildedir: eğer $p - değerlerimizin$ gerçek değerin yaklaşık ϵ'sı kadar doğru olmasını istersek, bu takdirde en az $\frac{1}{4}\epsilon^{-2}$ sentetik veri seti üretilmelidir. Böylece, $p - değerlerinin$ yaklaşık 2 ondalık basamak kadar doğru olması istenilirse, yaklaşık 2500 sentetik küme oluşturmak gerektiğinden $\epsilon = 0{,}01$ seçilmelidir.

$p - değerlerini$ hesapladıktan sonra, bunun kuvvet yasası hipotezini red edecek kadar küçük olup olmadığı veya tersine, söz konusu veriler için mantıklı bir hipotez olup olmadığı hakkında bir karar vermek gerekmektedir. Hesaplamalar sonucunda, eger $p \leq 0.05$ ise kuvvet yasası red edilir. Yani %5'in altında bir olasılıkla verilerin dağılımının kuvvet yasasına uygun olduğu söylenebilir ki bu da çok düşük bir olasılık olup modelin çok zayıf olduğunu göstermektedir.

Diğer taraftan büyük $p - değerleri$ ise verilerin kesinlikle kuvvet yasasına uygun dağıldığı anlamına gelmemektedir. Bunun iki nedeni vardır. Birincisi, gözlemlenen x aralığında verilere eşit veya daha iyi eşleşen diğer dağılımlar olabilmektedir. İkincisi, yukarıda belirtildiği gibi, n küçük değerlerinin ampirik dağılımının bir kuvvet yasasını yakından takip edeceği ve dolayısıyla veri için yanlış bir model olan kuvvet yasası bile büyük $p - değerleri$ için mümkün olabilmektedir.

3.3.2. Kolmogorov-Smirnov D İstatistiği ile Uyum İyiliği Testi

Kuvvet yasası dağılımının gözlemlenen verilere uygunluğunu test etmek için Kolmogorov-Smirnov (KS) Uyum İyiliği Testi kullanılabilir. Bu testte öncelikle, KS D istatistiği hesaplanarak,

$$H_0: Verilerin\ Dağılımı\ kuvvet\ yasası\ dağılımına\ uygundur.$$
$$H_1: Verilerin\ Dağılımı\ kuvvet\ yasası\ dağılımına\ uygun\ değildir.$$

hipotez test edilir. D istatistiği,

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{3.21}$$

şeklinde hesaplanılır. Buradaki $S(x)$, verilerin kümülatif dağılım fonksiyonu ve $P(x)$ ise $x \geq x_{min}$ bölgesinde kuvvet yasası modeline en iyi uyum sağlayan verilerin dağılım fonksiyonudur.

Eğer D istatistiği çok küçük (1'den çok uzak ise) ise, kabaca verilerin kuvvet yasası dağılımına uygun olduğu sonucuna varılır. Eğer D istatistiği çok büyük bir değer (1'e çok yakın) ise, kabaca verilerin kuvvet yasasına uygun olmadığı sonucuna varılır. Clauset, Shalizi ve Newman D istatistiğinin büyüklüğü ile kastedilen büyüklüğü bir $p-değeri$ hesaplayarak cevap vermektedir. Ancak bu $p-değerini$ hesaplamak için herhangi bir formül yoktur. Bunadan dolayı, Clauset, Shalizi ve Newman Monte Carlo Simülasyonu kullanarak bu $p-değerinin$ hesaplanmasını önermektedir [162]. Monte Carlo yöntemi ile gözlenen veriye en iyi uyuma sahip kuvvet yasası dağılımından çekilen yapay veri setleri üretilir. Üretilen bu veri setlerinin her biri için KS D istatistiği hesaplanır. $p-değeri$, gerçek veri seti için gözlenen D istatistiğinden büyük D değerlerinin toplam sayısının, hesaplanan toplam D değerleri sayısına bölünerek bulunur. Burada bulunan $p-değeri$, verilerin ileri sürülen kuvvet yasası dağılımından çekilmiş olma olasılığını belirtmektedir. Eğer $p-değeri$ oldukça küçük (1'e uzak) ise, verilerin kuvvet yasası dağılımından gelmediği söylenir. Eğer $p-değeri$ oldukça büyük (1'e yakın) ise, verilerin kuvvet yasası dağılımından geldiği söylenir.

3.3.3. Uyum İyiliği Testinin Performansı

Uyum iyiliğinin faydasını göstermek ve kuvvet yasasına uygun dağılmayan bir dağılımdan kuvvet yasasına uygun dağılan dağılımları doğru bir şekilde ayırt edebilmek için Şekil 3.5a'daki sürekli kuvvet yasası, log-normal ve üssel dağılımları göz önüne alınsın. Şekil 3.5b n örneklem sayısının bir fonksiyonu olarak, bu üç dağılımdan alınan veri setleri için, hesaplanan ortalama $p-değerini$ göstermektedir. Eğer n küçük ise; yani $n \lesssim 100$ olduğu zaman, log-normal, kuvvet yasası ve üssel dağılım için $p-değerleri$ eşik değer 0,1'den büyüktür. Bu da kuvvet yasası hipotezinin test tarafından ihlal edilmediği anlamına gelmektedir. Çünkü burada örneklem büyüklüğü yeterli büyüklükte olmadığından veri kümeleri doğru ayırt edilemez. Bununla birlikte, örneklem hacmi büyüdükçe, kuvvet yasası dağılımına uymayan iki dağılım için $p-değerleri$

[162] CLAUSET Aaron – Maxwell YOUNG – Kristian Skrede GLEDITSCH, *"On the Frequency of Severe Terrorist Events"*, Journal of Conflict Resolution, Vol. 51, Number 1, 2007, pp. 58-87.

azalmaktadır. Böylece kuvvet yasası modelinin bu veri setleri için zayıf bir uyum iyiliği olduğunu söylemek mümkündür[163].

Bununla birlikte, kesme parametresi x_{min}'den büyük dağılımın sadece bir bölümüne kuvvet yasası formu uyduğundan, x_{min} değeri kaç veri noktası ile çalışılacağını etkili bir şekilde kontrol etmektedir. Eğer x_{min} büyükse, veri kümesinin sadece küçük bir kısımı ondan daha büyük olmaktadır. Bu nedenle x_{min} değeri çok büyük olursa, kuvvet yasasını reddetmek için gereken toplam n değerinin de yeterli büyüklükte olması gerekmektedir. Bu olay, log-normal ve üssel dağılımlar için $p = 0{,}1$ eşik değerinin x_{min}'in bir fonksiyonu olarak elde edilmesi için gerekli olan n değeri Şekil 3.5c'de gösterilmiştir[164].

[163] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.
[164] CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E. J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

DÖRDÜNCÜ BÖLÜM
BIGQUERY İLE BÜYÜK VERİ UYGULAMALARI

Genellikle okuyucular yaşadığı coğrafyadan veya dünyadan haberdar olmak için haberleri alırken, sosyal bilimciler, aynı haberi, anlamlarını incelemek için rapor edilen olaylarla ilgili verileri çıkarmak için kullanır. 2000 yılından önce, araştırmacılar olayları belirlemek için saygın haber kaynaklarına dayanan küçük veri setlerini kodlamaktaydı. O zamandan bu yana, artan oranda gelişmiş algoritmalar ve bilgisayar standartları (Örneğin, Doğal Dil İşleme ve Makine Öğrenmesi), bu olayları benzeri görülmemiş bir ölçeğe genişletmemize izin vermekte ve protesto ve şiddet olaylarını hemen hemen gerçek zamanlı bir hızda önümüze getirmektedir. Dünyada bugün için olay verileri olarak bilinen üç olay kümesi vardır: ABD hükûmeti için Lockheed Martin tarafından geliştirilen Dünya Çapındaki Bütünleşik Kriz Erken Uyarı Sistemi (World-Wide Integrated Crisis Early Warning System-(ICEWS)). ICEWS, olay verisinin "altın standart"ı olarak kabul edilmesine karşın, bugün çok az insan bu veri kümesine erişebilmektedir. İkinci veri kümesi, 2013 yılında yayımlanan Olaylar, Dil ve Ton Küresel Veritabanı (The Global Database on Events, Language and Tone-(GDELT))'dır. Üçüncü veri kümesi, Olay/Yer: Bir Kutudaki Dataset, Linux Seçeneği (Event/Location: Dataset In A Box, Linux-Option-(EL:DIABLO)) ya da temel olarak, bir kutudaki olay verileridir. EL:DIABLO, olay verilerini oluşturmak için gerekli tüm araçları içeren bir sanal makinenin son kullanıcının makinesinde kolaylıkla oluşturulmasını sağlayan bir Vagrant kutudur. ICEWS veri kümesi elle kodlanmışken, GDELT ve EL:DIABLO tamamen bilgisayarlar tarafından kodlanmıştır[165].

Bu bölümde ikinci bölümde anlatılan Google'ın altyapısındaki BigQuery de bulunan GDELT veri seti kullanılarak, birinci, ikinci ve üçüncü bölümde anlatılan konulara paralel olarak bazı analizler yapılacaktır. Uygulamada Clauset, Shalizi ve Newman (2009) tarafından geliştirilen istatistiksel teknikler kullanılarak çatışma ve protestoların kuvvet yasasına uygun dağılıp dağılmadığı test edilecektir.

[165] GIN Jasper, *"Distilling Event Data from News Articles"*, 18 JANUARY 2015. https://www.jasperginn.nl/distilling-event-data-from-news-articles/#fnref:1/, (06/09/2017).

Öncelikle, Ocak 1979 ve Haziran 2017 yılları arasında aylık olarak dünya basınında yer alan haberler dikkate alınarak, dünya çapında yaşanan çatışma olayları incelenecek ve bu çatışmalar çizgi grafiği şeklinde verilecektir. Daha sonra çatışma sayılarının kuvvet yasası dağılımına uygun olup olmadığı test edilecektir.

Daha sonra, yine Ocak 1979 ve Haziran 2017 yılları arasında Türkiye ve Ukrayna ile ilgili yapılan protestoların aylık sayısının normalleştirilmiş bir "yoğunluk" ölçüsüne dönüştürülerek, çizgi grafiği şeklinde sunulacaktır. Ayrıca, bu protesto yoğunluklarının kuvvet yasası dağılımına uygun olup olmadığı test edilecektir.

Bu çalışmadaki analizler SQL, Ecxel, Tableau, SPSS ve R programları kullanılarak yapılmıştır.

4.1. GDELT Projesi

GDELT, insanlık tarihinde bu zamana kadar yaratılmış olan en geniş, en kapsamlı ve en yüksek çözünürlüğe sahip açık veritabanıdır. Bu veritabanı dünyanın tüm bölgelerinden, online ve offline formatta 100'den fazla dilde dünya haber medyasını izleyen bir platformdur. Bu platform 1 Ocak 1979'dan beri günlük olarak güncellenmektedir. GDELT 30 yılı aşkın bir süredir tüm dünyayı kapsayan, çeyrek milyar coğrafi referanslı bir kayıt veritabanını oluşturmaktadır. GDELT olayların altında yatan tüm insanları, organizasyonları, yerleri, temaları ve duyguları birbirine bağlayan devasa ağlar ile birleşerek, sadece yaratmak için eşsiz güçlükleri çözmekle kalmıyor. Aynı zamanda toplumsal ölçekli veriyle nasıl etkileşim kurduğumuz ve düşünüyor olduğumuzun "yeniden gözden geçirilmesi" ni de içermektedir[166].

GDELT Projesi, küresel insan toplumunu ve özellikle iletişimsel söylem ile fiziksel toplumsal ölçekli davranış arasındaki bağlantıyı daha iyi anlama arzusundan yola çıkılarak geliştirilmiştir. GDELT Projesinin vizyonu, küresel dünyanın anlaşılması için yeni bir platform sağlayan tüm açık bilgi kaynaklarını kullanarak hesaplanabilir bir formatta tüm gezegeni kodlamaktır[167]. Google Ideas tarafından desteklenen GDELT Projesi, dünya haber medyasında gerçek zamanlı, açık kaynaklı bir dizin oluşturmayı ve bu kodlanmış meta veriyi dünyayla paylaşmayı amaçlamaktadır. GDELT Proje arşivleri, küresel toplum hakkında dünyanın en büyük açık veri kümelerinden biridir. GDELT'teki

[166] The GDELT Project, *"The GDELT Story"*, 2017, http://gdeltproject.org/about.html#intro, (10.08.2017).
[167] The GDELT Project, *"The GDELT Story"*, 2017, http://gdeltproject.org/about.html#intro, (10.08.2017).

karmaşıklık, büyüme hızı ve analitik yük, verilerin anlaşılması ve erişilebilirliği açısından büyük zorlukları ortaya koymaktadır[168].

GDELT Projesi, 1979 yılından günümüze kadar tüm dünyayı kapsayan 300'den fazla kategoride, insanı, organizasyonu, yeri, temayı ve duyguyu birleştiren muazzam bir ağ diyagramıdır. Bu veri seti yaklaşık çeyrek milyarlık bir olay kaydından oluşmaktadır. Sadece 2015'te bir milyondan fazla duygusal değerlendirme ve 1,5 milyara aşkın yer referansı GDELT'in yapısında işlenmiştir. Yine 2016 yılının ilk yarısında GDELT'te 70 milyondan fazla resim işlenmiştir. GDELT, bu verileri dünyanın her ülkesindeki insan topluluğu ölçekli davranış ve inançların bir kataloğunu oluşturarak her insanı, organizasyonu, yeri, sayıyı, temayı, haber kaynağını ve olayı gezegenin tek bir kitlesine bağlamak için kullanmaktadır. Kısacası, GDELT, dünya çapında neler olup bittiğini, içeriğinin ne olduğu, kimin yer aldığı ve dünyanın her geçen gün nasıl hissettiğini yakalayan bir ağdır[169].

GDELT, 100'ün üzerinde dilde dünyanın dört bir tarafındaki yüz binlerce yayın, basılı ve çevrimiçi haber kaynağına dayanmaktadır. Bu kaynakların sayısı her geçen gün hızla artmaktadır. Dünya çapında çevrilen haberlere ek olarak, geçmişi 1979'a kadar uzanan GDELT'nin arka planında, AfricaNews, Agence France Presse, Associated Press, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Yabancı Broadcast Information Service, New York Times, United Press International ve Washington Post bulunmaktadır[170].

4.1.1. GDELT Analiz Servisi

GDELT Analiz Servisi, hem GDELT Olay Veritabanı (The GDELT Event Database) hem de GDELT Global Bilgi Grafiği (The GDELT Global Knowledge Graph)'ni görselleştirmeye, keşfetmeye ve dışa aktarmaya imkân sağlayan çeşitli araç ve hizmetler sunan ücretsiz bir bulut tabanlı hizmettir.

1) Görselleştirme ve Format Dönüşümü

Olay Veritabanı ve Global Bilgi Grafiği'nin yapısında coğrafi, zamansal, ağ ve bağlamsal görselleştirmeleri yapmak için on dört farklı araç bulunmaktadır. Bu araçlar

[168] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.
[169] The GDELT Project, *"The GDELT Story"*, 2017, http://gdeltproject.org/about.html#intro, (10.08.2017).
[170] The GDELT Project, *"The GDELT Story"*, 2017, http://gdeltproject.org/about.html#intro, (10.08.2017).

hiçbir teknik uzmanlığa gerek duyulmadan sadece istenilen görselleşmeyi seçerek, sorguyu girerek, istenilen sonuçlara birkaç dakika içerisinde ulaşmak mümkündür.

Kullanıcılar genellikle GDELT ile daha kolay çalışmak için merkezi bir araç grubuna ihtiyaç duymaktadır. Özellikle çok boyutlu veri bilgi tabanını, analistlerin ve akademisyenlerin daha iyi anlayabilecekleri dosya biçimlerine ve görselleştirmelere dönüştürmede bu araçlara daha çok ihtiyaç duyulmaktadır. Ayrıca bu araçlar her gün kullandığımız araç kitleri ve yazılımlarla uyumludur. Bu amaçla, her bir araç, ilgili dosya formatları, CSV'den Google Earth'e ve Gephi'ye aktarmaya olanak tanımaktadır. Örneğin; bu araçlar, daha fazla analiz yapmak için bir endüstri çevresinde etkileyici bir ağ oluşturmamıza ve bir Gephi dosyası olarak çıkartabilmemize olanak tanımaktadır[171].

2) Veri Aktarma

Genellikle GDELT Analiz Servisi, bir hipotezi test ederken veya ortaya çıkan bir eğilimi kontrol ederken başlamak için en iyi yerdir. GDELT Analiz Servisi'nde yeni bir sorgu anında test edebilir, sonuçları sadece birkaç dakika içerisinde geri alınabilir, araştırmanın tekrarlamaya değer bir şey olup olmadığını görmek için tekrar tekrar test edilebilmektedir. Analiz Hizmetinde, keşfetmek, görselleştirmek veya analiz etmek istenilen ham veriyi özel dosya formatlarından CSV'ye aktarılarak birçok analizi yapmak mümkündür. Örneğin; belirli bir ülkedeki sivillere yönelik tüm saldırıları dört aylık bir süre içinde bulmaya çalışıyorsanız, eşleşen olayların sayısının yeterli olup olmadığını veya sorgunuzu daha da ayarlamanız gerektiğini belirlemek için TimeMapper aracı kullanılabilir. Daha sonra da Exporter aracını kullanarak sadece eşleşen kayıtları içeren bir CSV dosyası indirmek mümkündür[172].

4.1.2. Ham Veri Dosyaları

Gelişmiş kullanıcılar ve benzersiz kullanım durumlarına sahip kullanıcılar, temel olay ve grafik verilerini CSV formatında indirebilmektedir. 2015 yılında Global Bilgi Grafikleri tek başına 2,5 TB'dan fazla veriye ihtiyaç duymuştur. Bu nedenle, bu veri kümelerini kullanabilmek için geniş veri kümeleri ile çalışan derin teknik bilgi ve kapsamlı deneyime ihtiyaç duyulmaktadır.

[171] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html, (15.08.2017).
[172] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html, (15.08.2017).

4.1.2.1. GDELT 1.0 Olay Veritabanı

GDELT 1.0 Olay Veritabanı, tarihe göre bir dizi sekmeyle sınırlandırılmış dosyalar halinde düzenlenen çeyrek milyar kaydı içermektedir. 31 Mart 2013 tarihine kadar kayıtlar, olayın gerçekleştiği tarihe göre aylık ve yıllık olarak dosyalarda saklanmaktadır. 1 Nisan 2013'ten beri dosyalar günlük olarak oluşturulmuş ve kayıtlar, olayın meydana geldiği tarihten ziyade dünya haber medyasında bulunduğu tarihe kadar saklanmıştır (Olayların % 97'si olay gerçekleştikten sonraki 24 saat içinde bildirilir. Ancak her gün bahsedilen bazı olaylar ilk kez bahsedilen olayların geçmişidir. Eğer bir olaydan daha önce bahsedilmiş ise bu olay tekrar dâhil edilmez.). GDELT 1.0 Olay Veritabanı'ndaki dosyalar ZIP ile sıkıştırılmış sekmeyle ayrılmış formattadır. Ancak .TXT veya .TSV dosyalarını kabul etmeyecek bazı yazılım paketlerine olan ihtiyacı gidermek için ".CSV" uzantısı da bulunmaktadır[173].

Haftanın yedi günü her sabah, en son günlük güncelleme 6:00'da EST ile gönderilir. Bu dosya, önceki günün tarihini "YYYYMMDD.export.CSV.zip" şeklinde (yani, 24 Mayıs 2013 sabahı "20130523.export.CSV.zip" adlı yeni bir dosya olarak adlandırılmaktadır.) adlandırılır. UNIX veya Linux kullanıcıları, her sabah en son günlük güncellemeyi ve işlemi otomatik olarak indirmek için kolayca bir cronjob veya başka otomatik zamanlama süreçleri oluşturabilmektedirler. Bu sayede izleme, öngörme, erken uyarı, uyarı hizmetleri ve diğer uygulamalar yapılabilmektedir[174].

Ayrıca, daha eski akademik olay veritabanlarında yaygın olarak kullanılan "günde bir kez" ülke düzeyinde filtrelemeyi kullanan özel bir GDELT 1.0 "indirgenmiş" olay veri kümesi (1.1GB) bulunmaktadır. Verilerin bu sürümü, daha önce karşılaşılan ve "DATE + ACTOR1 + ACTOR2 + EVENTCODE" adlı veritabanını daraltan önceki olay analiz deneyimine sahip olan toplama düzeyiyle en iyi şekilde eşleşecektir (yani belirli bir günde Rusya'da herhangi bir yerde yapılan her protesto tek bir girişe indirgenir.). Bu sürüm yalnızca önceki nesil akademik olay veritabanlarına dayanan analizlerle uyumluluk için önerilir ve 1 Ocak ile 17 Şubat 2014 arasındaki süreleri kapsar[175].

---

[173] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html#gdeltanalysisservice, (25.08.2017).

[174] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html#gdeltanalysisservice, (25.08.2017).

[175] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html#gdeltanalysisservice, (25.08.2017).

4.1.2.2. GDELT 1.0 Global Bilgi Grafiği (GKG)

GDELT 1.0 Global Bilgi Grafiği, 1 Nisan 2013'te başlamış ve iki paralel veri akışından oluşmaktadır. Biri tüm bilgi grafiğini tüm alanlarıyla kodlamaktadır. Diğeri protestocuların sayısı, öldürülenlerin sayısı veya yerinden olmuş (göç etmiş) veya hastalananların sayısı gibi önceden tanımlanmış bir kategori kümesinin "sayılarını" kayıt eder ve bir grafiğin alt kümesi olarak kodlamaktadır. Bu sayılar birincil GDELT olay akışındaki CAMEO olaylarından bağımsız olarak oluşmaktadır. Bunlar; endüstriyel kazalar (CAMEO'da yakalanmayan) veya doğal afetler nedeniyle yerlerinden edilmiş veya bir salgın hastalığı geçirmiş kişiler olabilmektedir[176].

İkinci dosya, tüm kişileri, kuruluşları, konumları, duyguları, temaları, sayıları, olayları ve kaynakları her gün birleştiren gerçek grafiği içeren tam bir grafik dosyasıdır. Bu dosyalar "YYYYMMDD.gkgcounts.csv.zip" olarak adlandırılır ve her sabah 6 AM EST tarafından haftanın yedi günü yayınlanmaktadır. Global Bilgi Grafiği şu an da "alpha" sürümündedir ve yeni özellikler kazandırıldığında temel algoritmaları genişleyerek zamanla değişmektedir.

4.1.2.3. GDELT 2.0: Gerçek Zamanlı Olarak Küresel Dünyamız

GDELT 2.0, küresel dünyayı izlemek için şimdiye kadar yaratılmış en büyük ve en iddialı platformlardan biridir. Bu platform dünya haberlerini 65 dilde gerçek zamanlı olarak çeviren, 2300'den fazla duygu ve temayı ölçen, Batı dünyasındaki medyanın büyük bir envanteridir. GDELT 2.0, küresel dünyayı algılama biçimimizi ve etkileşimimiz arasındaki dil engelini ortadan kaldırarak, dünyadaki olayların tepkilerini ve duygusal rezonansını derinden inceleyerek yeniden tanımlamamıza izin vermektedir. Kısaca, GDELT, dünyanın her yerinden bir haber raporunu izleyen 15 dakika içerisinde, tüm olayları, sayıları, teklifleri, kişileri, organizasyonları, yerleri, temaları, duyguları, ilgili görüntüleri ve videoları tanımlayarak işlenebilecek hâle getirmektedir. Özellikle sosyal medya mesajları, küresel bağlamda bu platform içerisine yerleştirilmiştir. Bu sayede, dünya üzerinde canlı bir açık meta verisi oluşturularak anlık araştırmalar yapma imkânı doğmuştur[177].

---

[176] The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html#gdeltanalysisservice, (25.08.2017).

[177] The GDELT Project, *"GDELT 2.0: Our Global World in Realtime"*, 19 February 2015, http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/, (25.08.2017).

GDELT 2.0, insan dünyasını "hesaplanabilir" hâle getirmeye çalışan ve insan dünyasını gerçek zamanlı olarak gezegen ölçekte nasıl incelediğimizi temelde yeniden tasarlamak için Google Cloud'un muazzam gücünden yararlanmaya çalışan açık bir veri kümesidir[178].

Yarım yüzyılı aşkın bir süredir, küresel toplumu anlama konusundaki çalışmaların çoğunda İngilizce dilini kullanan Batı medyası takip edilmiştir. GDELT 2.0, dünyayı dinlemenin ne anlama geldiğini yeniden tanımlıyor; bugüne kadar dünyanın en uzak köşelerini gerçek zamanlı olarak seslendirmek için Batı dışındaki medya organlarının en yüksek çözünürlükteki veri stoklardan biri olduğuna inanılıyor. GDELT 2.0'ın getirmiş olduğu en önemli yenilik, GDELT Translingual'dır. " Dünyanın en büyük gerçek zamanlı otomatik haber tercümesi uygulaması" olarak tanımlanan GDELT Translingual, gerçek zamanlı olarak 65 dilde çeviri yapabiliyor. Bu da dünyada İngilizce dışındaki dillerin %98,4'ünü kapsadığını gösteriyor. Ayrıca dünyadaki haber bültenlerini İngilizce kaynaklı haberlerin etkisinde kaldığı düşünülürse, GDELT Translingual dünyayı algılama biçimimizi büyük oranda değiştiriyor. Yine son birkaç yıldır, küresel toplumun incelenmesi için önemli bir teknolojik gelişme olan duygu analizi ortaya çıkmıştır. İnsan duygusu sadece "positive" ve "negative" olmaktan çok daha fazlasını içermektedir. Ayrıca bu duygu kişisel ve kültürel bağlamda tarihin derinliklerinde saklı olan zengin bir kültürel dokuyu barındırmaktadır. Bu amaçla, GDELT 2.0, 15 dilde 2300'den fazla duygu ve temayı birlikte değerlendiren 24 duygusal ölçüm paketini bir araya getirerek küresel duygu çalışmasında yeni bir çığır açmıştır. GDELT 2.0, duygu analizi dünyasındaki tek en büyük dağıtımdır. Bu veri seti çok dili ve disiplini birbirine bağlayan çok sayıda duygusal ve tematik boyutu bir araya getirmekte ve hepsini gerçek zamanlı olarak dünyadaki son dakika haberlerine uygulamaktadır. Bu sayede, duyguyu nasıl düşündüğümüz ve yeni olayları bağdaşlaştırma, yorumlama, yanıtlama ve anlama biçimimizi daha iyi anlamamıza yardımcı olarak bize yeni bir çağın kapısını aralamaktadır[179].

[178] The GDELT Project, *"GDELT 2.0: Our Global World in Realtime"*, 19 February 2015, http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/, (25.08.2017).
[179] The GDELT Project, *"GDELT 2.0: Our Global World in Realtime"*, 19 February 2015, http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/, (25.08.2017).

4.1.3. GDELT + BigQuery = Gezegeni Sorgulayın

Google BigQuery, GDELT gibi büyük veri kümeleri için geliştirilmiş bulut tabanlı bir analitik veritabanıdır. BigQuery, onbinlerce trilyon satırı içeren çok petabaytlı veri setlerine ve arşivlere karşı etkileşimli SQL sorgularını etkinleştirmek için Google'ın altyapısını kullanmaktadır. Sorgular, bir REST API'sı aracılığıyla gönderilir ve gelişmiş sorgular için kullanıcı tanımlı JavaScript işlevleri aracılığıyla genişletilebilen standart bir SQL'de yapılabilmektedir. Birkaç yüz terabayt yeni veri (toplu iş ve akış), müşterilerin hepsi sorgu için anında mevcut olan BigQuery'ye yüklenmektedir. Tek bir sorgu üzerine binlerce işlemci getirilebilir ve veri dizine ekleme veya bölümleme gerekmeden hızlı sonuçlar elde edilebilmektedir[180].

BigQuery yapısında bulunan GDELT sayesinde artık küresel insan topluluğu ve dünya hakkında gerçek zamanlı bilgilere erişmek mümkün hâle gelmiştir. GDELT veri kümesi, İnternet Arşivinin (1.3M hacim) İngilizce ve HathiTrust'un (2.2 milyon ciltlik) İngilizce dilinin tamamına ait toplam koleksiyonlarını kapsayan, iki yüzyıla dayanan 3.5 milyon sayısallaştırılmış kitabı içermektedir. Bu koleksiyonlar, GDELT Global Bilgi Grafiği kullanılarak işlenmiş ve Google BigQuery'de mevcuttur. GDELT'te 2017 yıldan geriye uzanan bir milyondan fazla sayfa, tüm kişilerin, kuruluşların ve diğer isimlerin listesi derlenmiştir. Bu verileri işleyebilmek için coğrafi kod metni halinde toplanarak, 4.500'den fazla duygu ve tema derlenmiştir. Bu hesaplanan meta verilerin tümü, katkıda bulunan kütüphaneler tarafından sağlanan başlık, yazar, yayıncı ve konu etiketleri de dâhil olmak üzere mevcut tüm kitap düzeyinde meta verilerle birleştirilmiştir. İnternet Arşivi verileri, 1800-1922 arasında yayınlanan tüm İnternet Arşivi kitaplarının tam metnini ve "americana" aramasını kullanarak İngilizce tam metin bulunan mevcut Amerikan Kütüphaneleri koleksiyonundaki tüm kitapları içermektedir. HathiTrust için, HathiTrust tarafından özel bir araştırma kapsamında 1800-2015 yılları arasında İngilizce dilinde yayınlanan tüm kitapları içermektedir[181].

[180] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.
[181] Google Cloud Platform, *"GDELT HathiTrust and Internet Archive Book Data"*, 2017, https://cloud.google.com/bigquery/public-data/gdelt-books, (10.08.2017).

4.1.3.1. GDELT, Büyük Veri Sorunlarını Aşmak için Google BigQuery'yi Nasıl Kullanıyor?

GDELT ile çalışırken karşılaşılan en büyük zorluklardan biri, muazzam büyüklükteki bir veri seti ile nasıl etkileşim kurulabileceğidir. Pek az veritabanı platformu, bu karmaşık veri kümesini çok çeşitli erişim modelleriyle ve her gün sorgulara toplanan alanların permütasyon sayısıyla yönetebilmektedir. Google'ın BigQuery veritabanı, GDELT gibi veri kümeleri için özel olarak tasarlanmıştır. Bu sayede BigQuery de bulunan tüm veri kümelerine gerçek zamanlı sorgulama yapılabilmektedir. Bu, GDELT'e nasıl eriştiğinize, hangi sütunlara baktığınıza, hangi tür operatörleri kullandığınıza veya sorgunuzun karmaşıklığına bakılmaksızın, gerçek zamanlı sorgulamalar yapılabilmektedir.

BigQuery aşağıdaki özellikleri sayesinde, kullanıcıların GDELT veri kümeleri ile etkili bir şekilde etkileşimde bulunmalarını sağlamaktadır[182]:

Ölçeklenebilirlik ve Esneklik: GDELT veri setleri toplamak için on trilyon veri noktasını birden fazla biçimde kodlamaktadır. Protesto ya da barış görüşmeleri gibi olayların bazı katalogları gibi bazı veri akışları, RDBMS sistemleri ile birlikte kullanılmak üzere tasarlanmıştır. On yıllar boyunca GDELT yapısında, optimize edilmiş yüksek orandaki yapılandırılmış şemalar bulunmaktadır. GDELT'teki anlatıların ve duyguların katalogları gibi diğer akışlar, çok küçük ölçeklerde kullanılmak üzere tasarlanmış meta verilerdir. Bu ölçekteki kodlamadan çok az örnek vardır. Ayrıca, BigQuery de karmaşıklığın yanısıra, değerlendirilecek boyutların sayısı devamlı genişlemekte ve sürekli genişleme kapasitesine sahip akışkan bir şema gerektirmektedir. Değerlendirilen her bir boyut, diğer bilgilere nispeten yakın olması veya yoğunluğu gibi sayısal bilgileri kodlamaktadır. Bu, BigQuery'nin sağladığı karmaşık yuvalanma, sayısal değerler ve sabit genişlemeyi destekleyen esnek bir veri formatını gerektirmektedir.

Yeni Sütunların Sürekli Eklenmesi: GDELT'in veri setleri, izlenen her haber içeriğinden milyonlarca temanın ve binlerce duygunun varlığını, bağlamını ve yoğunluğunu belirlemeye çalışmaktadır. Tema ve duyguların listesi, her biri sırayla her bir skorla ilişkili sayısal değerleri depolamaktadır. Depolanan bu veriler sürekli artmakta ve herbir satırın analizi için milyonlarca boyutun sorgulanması gerekmektedir. GDELT,

[182] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.

verileri yuvalanmış ayrılmış biçimlerde depolamak ve sorgu zamanında seçili değerleri ayıklamak için BigQuery'yi kullanmaktadır.

Gerçek Zamanlı ve Tarihsel Verileri Kombine Etme: BigQuery'de GDELT'yi kullanmanın en ilginç tarafı, yalnızca karmaşık sorgulama ve veri ayıklama işlemlerini hızlandırması değil, aynı zamanda gerçek dünya analizlerinin tamamen veritabanında yürütülmesine de olanak tanıyan kapıları açmasıdır. Son 38 yılda dünyadaki en önemli çatışma etkileşimini bir ayda hesaplamayı veya bir grup ülke arasındaki ilişkilerin farklı sınıfları üzerinden çapraz sekmeli ilişkilendirmeyi BigQuery ile gerçekleştirmek mümkündür. Bu tür sorgular, tamamen BigQuery'nin içinde çalıştırılabilir ve yalnızca bir kaç saniye içinde sonuçlara ulaşılmaktadır. Bu, gerçek zamanlı olarak küresel ölçekli eğilimler üzerinde "What if " hipotezlerini denememizi sağlamaktadır[183]. Gerçek zamanlı güncellemeler, olayların kırılmasının analizini etkinleştirmek için derhal kullanıma sunulmasını gerektirmektedir. Bundan dolayı gerçek zamanlı ve geçmiş veri depolamalarında birleşik sorgulama yapılmasına olanak tanıyan bir ortam gereklidir. BigQuery, akış ekleri yoluyla bunu desteklemektedir[184].

Birçok Sütun Üzerinde Geçici Endeks İçermeyen Arama: GDELT veri setlerinden bir tanesi, 310 milyondan fazla satır ve 59 sütundan oluşan küresel olayların 38 yıllık bir arşividir. Sorgular genellikle her bir sorgunun farklı sütun kombinasyonlarına erişen birçok sütunu birleştirmektedir. Hiçbir sütun veya sütun kümesi boyut indirgemesi sunmaz; bu da geleneksel endeksli bir RDBMS modelinin kullanılmasını imkânsız kılmaktadır. Bunun yerine BigQuery tarafından kullanılan bir dizin içermeyen sorgu işleme modeli gerekmektedir[185].

Genel Erişim: GDELT'deki tüm veri akışları serbest olup, bunlar açık verilerdir. Bu, veri depolama ve yönetimi ile ilgili kaynakları sorgulamaktan ayıran bir platform üzerinde çalışması gerektiğini ifade etmektedir. Bununla birlikte BigQuery, veri kümelerinin genel erişilebilir olmasını sağlar.

Gelişmiş Hesaplama: GDELT sorguları, belirli bir belgede temaları eşleştirmek, haritalama için coğrafi histogramlar üreterek terabaytlarca verinin işlenmesini sağlayacak

---

[183] The GDELT Project, *"Google BigQuery"*, 2017, http://www.gdeltproject.org/data.html#googlebigquery, (20.08.2017).

[184] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.

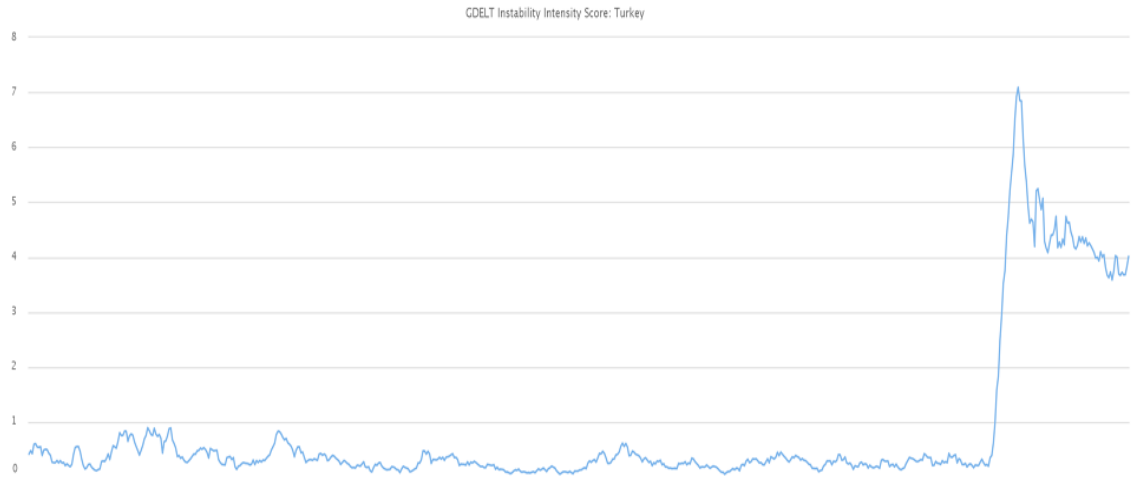
[185] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.

sofistike bir mantık içermektedir. GDELT, sofistike kontrol akışlarını ve algoritmalarını tamamen veritabanı platformu içerisinde kendisi yürütebilir ve hesaplamayı veriye getirebilmektedir. BigQuery'nin kullanıcı tanımlı işlevleri bunu mümkün kılmaktadır.

Veri İçi Veritabanı Analizi: Yukarıda özetlenen daha geleneksel analizlere ek olarak, belirli analiz türleri, veritabanının kendisine karşı etkin bir şekilde analiz etme yeteneğini gerektirmektedir. Örneğin, tarihin döngülerini ve kalıplarını haber medyasında ifade edilen şekilde incelemek için tüm veritabanının hareketli bir pencerede çapraz ilişkilendirme özelliğini kullanmak gerekmektedir. Bunun için şeffaf hesaplama ve veri hareketinin ölçeklendirilmesi gerekmektedir. Bu tür analizler için ihtiyaç duyulan muazzam sayıda işlemci, BigQuery gibi bulut tabanlı bir ortamı gerektirmektedir.

4.1.3.2. BigQuery ve GDELT'in Çalışması

BigQuery, GDELT'in muazzam büyüklükteki arşivlerini gerçek zamanlı olarak keşfetmeyi, görüşlerini interaktif olarak sorgulamayı, analiz etmeyi ve görselleştirmeyi sağlamaktadır. Genellikle, BigQuery, bir ulus içindeki protesto ve çatışmalardaki zaman eğilimlerini incelemek ve güncel huzursuzlukları tarihsel bağlamında incelemek için GDELT'i kullanır. Örneğin, 15 Temmuz 2016'da Türkye de yaşanan darbe girişiminin GDELT'in gözünden nasıl göründüğünü merak edenler aşağıdaki zaman çizelgesini inceleyebilir. Şekil 4.1 Türkiye için GDELT Stability API zaman çizelgesinde görülen ani devasa sıçramayı göstermektedir. Stability API zaman çizelgesi, 20:00 civarında ilk yukarı doğru hareketi göstermektedir. Ancak işler gerçekten 20:45'te kötüye gitmektedir. GDELT tarafından değerlendirilen istikrarsızlık seviyesi 00:15 civarında zirveye ulaşmış ve Cumhurbaşkanı Recep Tayyip Erdoğan'ın o saat başında darbecilerin ülkenin kontrolünü kaybediyor yönündeki haberlerine göre, 01:00 de ülke hızlı bir şekilde istikrara doğru gitmektedir.



Şekil 4.1: GDELT'in Gözünden Türkiye'deki 15 Temmuz 2016 Darbe Gecesi Kararlılık Zaman Çizelgesi
Kaynak: http://blog.gdeltproject.org/wp-content/uploads/2016-turkey-coup-stability-apiti, (20.08.2017).

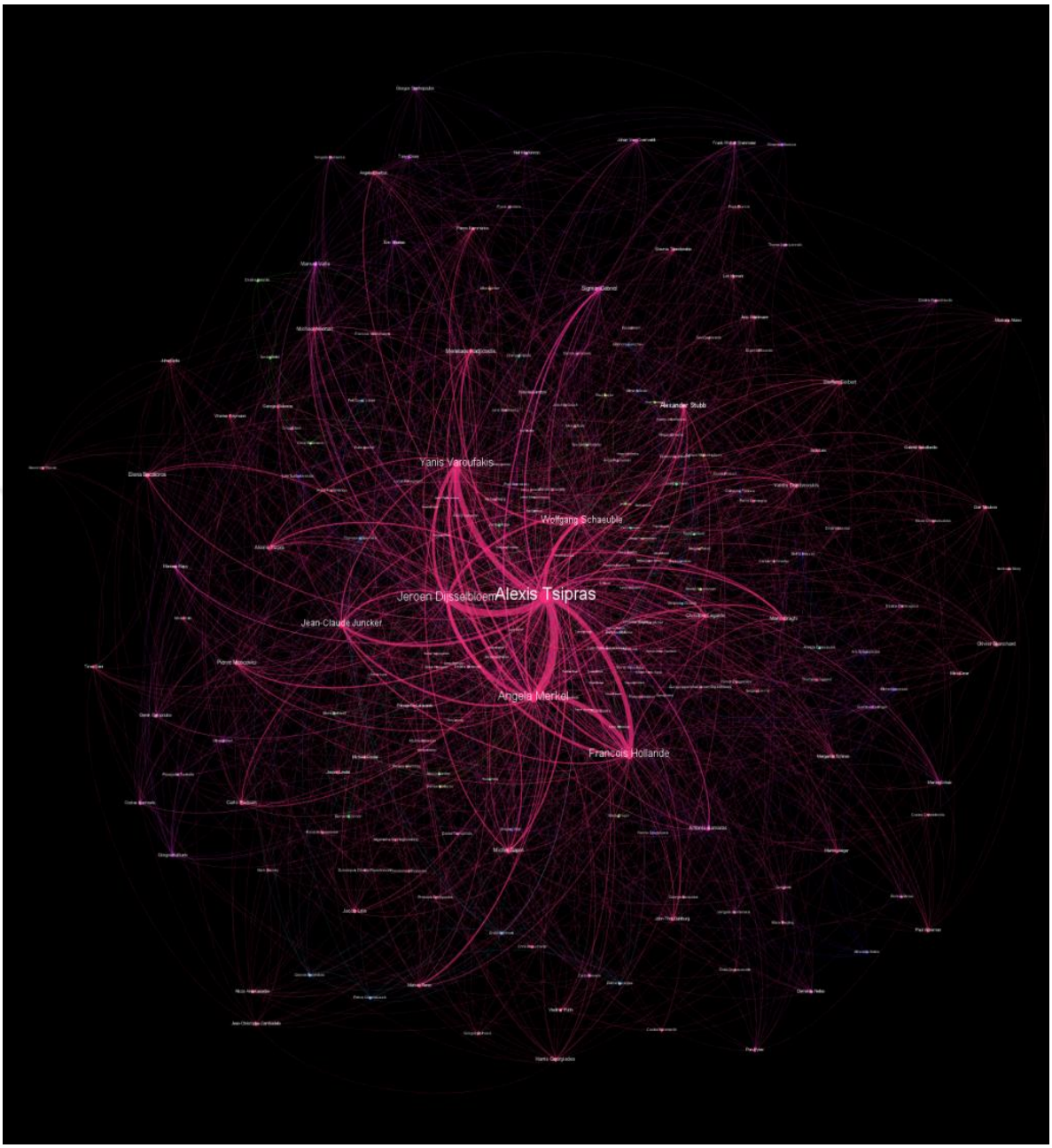
Hızla hareket eden gerçek zamanlı durumu değerlendirmek için GDELT'i kullanırken, GDELT'in şu an da 15 dakikalık bir kalp atışı üzerinde çalışan çok aşamalı bir işlem hattını kullanmaktadır. İçerik, her bir kalp atışında boru hattı etabından kademesine doğru ilerlemektedir. Reuter'in 19:29'daki olayların zaman çizelgesine göre, ilk raporlar, iki önemli köprünün askeri birlikler tarafından kapatıldığı ile ilgili haberlerdir. Sosyal medya bu zamana kadar kapalı köprülerin ve askeri hareketlerin dağınık görüntülerini de göstermeye başlamıştır. Bu en eski medya raporlarının çoğu bu saatlerde ve 19:30'dan biraz sonra ortaya çıkmıştır. Dolayısıyla, GDELT'nin içme suyu boru hattı öncelikle bu içeriği 19:45-20:00 döngüsü boyunca ele geçirmiştir. Daha sonra, yabancı dil materyali, makine çevirisi için GDELT Translingual altyapısına gönderildi ve burada 20:00 ile 20:15 arasında tercüme edilir. Tercümeden sonra işleme konması için 20:15-20:30 saatleri arasında iletilir. Daha sonra, 20: 30-20: 45 saatleri arasında, Stability API zaman çizelgesini içeren GDELT API altyapısı tarafından işlenmiş ve sonuçta, büyük askeri mobilizasyonun Stability zaman çizelgesinde 20:45'te yansıtması sağlanmıştır[186].

Diğer bir GDELT veri seti, izlenen her haber makalesinde bulunan tüm kişileri, organizasyonları, yerleri, temaları ve duyguları kaydetmek ve bu bilgilerle büyük bir meta veri dizini oluşturmaktır. Tek bir SQL satırı ile BigQuery, Yunanistan'ı kurtarma referandumunun haberinde bir arada bulunan, en çok bahsedilen 1500 çiftin isim listesini derlemek için 150 milyon haber kaydını taramıştır. Bu işlem saniyeler içinde

[186] The GDELT Project, *"What Turkey's Attempted Coup Looked Like Through GDELT's Eyes"*, 16 July 2016, http://blog.gdeltproject.org/turkeys-attempted-coup-looked-like-gdelts-eyes/, (20.08.2017).

gerçekleştirilmektedir. BigQuery, ardından ağ diyagramını üretmek için Gephi ile görselleştirmek için tasarlanmış bir CSV dosyası çıktılarını almaktadır. Bu diyagram türü, kullanıcıların belirli bir konunun, merkezi rakamlar olan dünya haber medyası tarafından nasıl sunulduğunu ve birbirleriyle nasıl ilişkili olduklarını keşfetmeye imkân sağlamaktadır. Bu durumda, Almanya'nın Angela Merkel ve Wolfgang Schaeuble, Lüksemburg'un Jean-Claude Juncker ve Fransa'nın Francois Hollande gibi AB liderlerinin Yunanistan'ı kurtarma referandumun da oynadıklar roller Şekil 4.2'de açıkça görülmektedir[187].

[187] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.



Şekil 4.2: 1-15 Temmuz 2015'te Yunanistan Haberlerinde en çok Geçen Kişilerin Şebeke Diyagramı
Kaynak: http://blog.gdeltproject.org/a-network-diagram-of-greece-july-1-15/, (20.08.2017).

GDELT'nin yaygın olarak BigQuery'i kullandığı bir başka yol da, belirli konularla bağlantı içinde görünen yerleri haritalamasıdır. BigQuery'nin kullanıcı tanımlı işlevler kapasitesi, bir sorgunun bir parçası olarak çalıştırılacak ve tüm analitik boru hattının BigQuery'de özel olarak çalışmasına olanak tanıyan keyfi karmaşık JavaScript uygulamalarına (örneğin, her temayı belgedeki en yakın konumla ilişkilendiren karmaşık filtrelemeye) olanak tanımaktadır. Şekil 4.3'te, Şubat ve Haziran 2015 arasında yaban

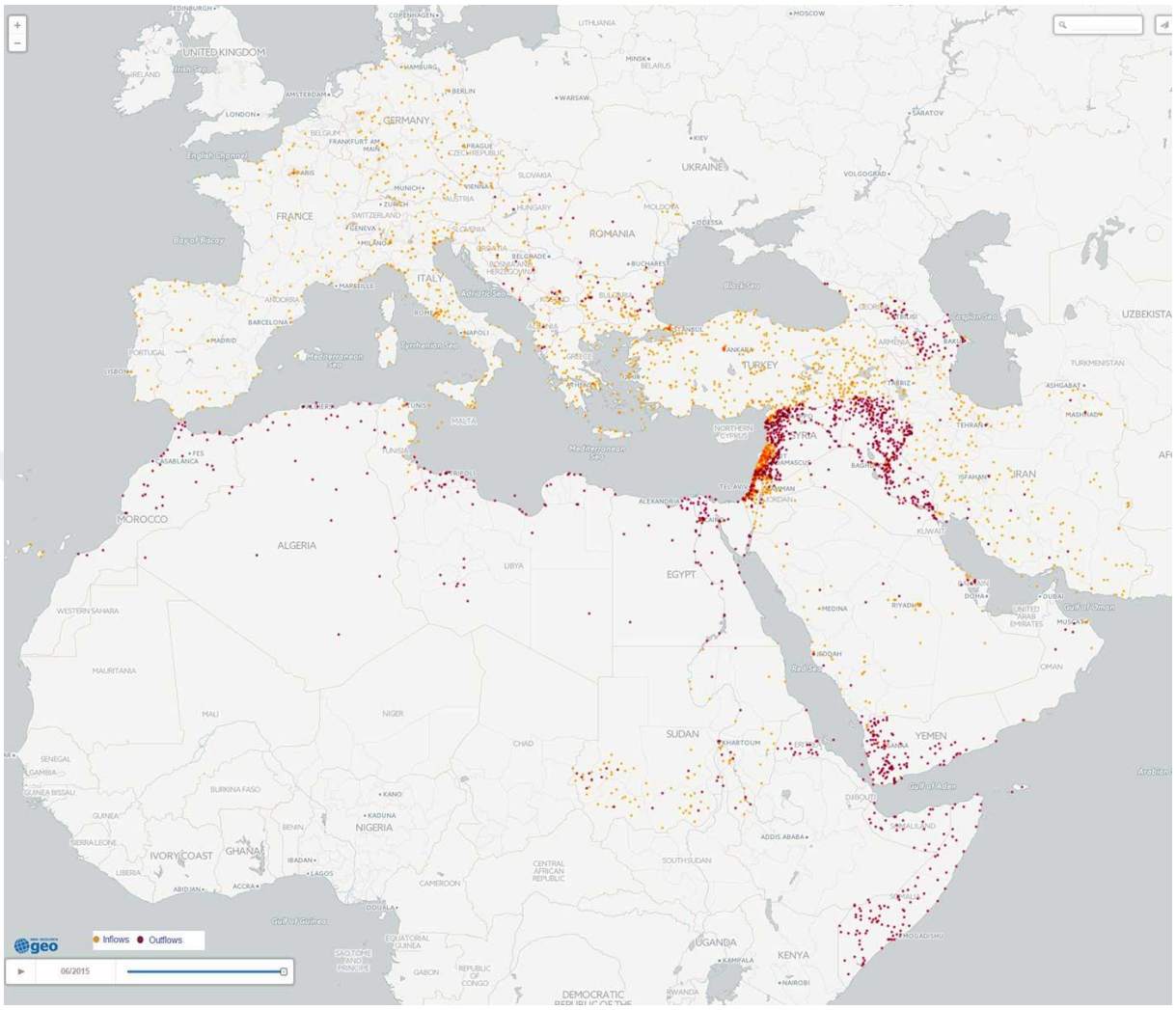
hayatı suçuyla bağlantılı olarak belirtilen tüm yerler CartoDB kullanılarak eşleştirilmiştir. Bu harita, yaygın yaban hayatı suçunun ne ölçüde arttığını anlatmak için kullanılmıştır. GDELT ve BigQuery tarafından üretilen diğer haritalar, anti-tank savar silahları, iklim değişikliği, 200 yıllık kitaplar, Yunan borç krizi ve hatta İslam Devleti ile bağlantılı yerleri içermektedir[188].

Şekil 4.3: Dünya çapında, Şubat ile Haziran 2015 arasında Küresel Haberlerde Vahşi Hayat Suçuyla Bağlantılı Olarak Bahsedilen Yerler
Kaynak: LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.
.

Banco Bilbao Vizcaya Argentaria (BBVA)'nın Cross-Country Emerging Markets Unit'deki araştırma ekibi, Şekil 4.4'te, 14 Ocak-15 Haziran 2015 tarihleri arasında Avrupa mülteci krizinin bir haritasını göstermiştir. Burada BigQuery'yi kullanarak GDELT ile bir dizi huzursuzluk dinamiklerinin analizi yapılmıştır. BBVA, aşağıdaki haritada, Avrupa ve Kuzey Afrika'daki mültecilerin coğrafi girişlerini (turuncu) ve çıkışlarını (kırmızı), bu yılın ilk altı ayı için göstermektedir. Milyonlarca haber raporunda görülen yüzey eğilimleri, önemli ölçüde istikrarsızlığı ve huzursuzluğu tetikleme

[188] LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.

potansiyeline sahip olup ortaya çıkan krizlerin ölçeği ve coğrafi dağılımı hakkında bilgiler vermektedir.



Şekil 4.4. 14 Ocak - 15 Haziran 2015 Tarihleri Arasında Avrupa ve Kuzey Afrika'daki Mülteci Giriş-Çıkış Haritası
Kaynak: BBVA Cross-Country Emerging Markets Unit, used with permission, (20.08.2017).

4.2. Dünya'daki Çatışmaların İncelenmesi

Bugün dünyada çatşma olayları sürekli artmakta ve dünya genel bir savaş ortamına doğru hızla sürüklenmektedir. Özellikle bilişim teknolojilerinde yaşanan gelişmeler medyada devletlerarasındaki çatışma olayları ile ilgili yapılan haberlerin ve yazılan kitapların kolayca dijital ortamda kayıt altına alınmasını sağlamıştır. BigQuery'deki GDELT veri seti bugün dünyada yaşanan çatışma olaylarını gerçek zamanlı olarak takip etmemizi sağlayarak bu olayların gerçek zamanlı olarak izleme imkânı sağlamıştır.

Burada, BigQuery de bulunan GDELT verisi ile 1979-2017 yılları arasında dünya genelinde yaşanan çatışma olayları analiz edilip yorumlanacaktır.

BigQuery ile ilk kez çalışıyorsanız, bir Google projesi oluşturmak için kaydolmak gerekmektedir. Bunun için herhangi bir ücret ödemeye gerek yoktur.

Ardından, aşağıdaki sorguyu kopyalayıp BigQuery sorgu kutusuna yapıştırıp "RUN QUERY" düğmesini tıklayın:

```
SELECT MonthYear, Actor1Name, Actor2Name, Count FROM (
SELECT Actor1Name, Actor2Name, MonthYear, COUNT(*) Count, RANK()
OVER(PARTITION BY MonthYEAR ORDER BY Count DESC) rank
FROM
(SELECT Actor1Name, Actor2Name, MonthYear FROM [gdelt-bq:full.events]
WHERE Actor1Name < Actor2Name and Actor1CountryCode != '' and
Actor2CountryCode != '' and Actor1CountryCode!=Actor2CountryCode),
(SELECT Actor2Name Actor1Name, Actor1Name Actor2Name, MonthYear FROM
[gdelt-bq:full.events] WHERE Actor1Name > Actor2Name and
Actor1CountryCode != '' and Actor2CountryCode != '' and
Actor1CountryCode!=Actor2CountryCode),
WHERE Actor1Name IS NOT null
AND Actor2Name IS NOT null
GROUP EACH BY 1, 2, 3
HAVING Count > 100
)WHERE rank=1
ORDER BY MonthYear
```

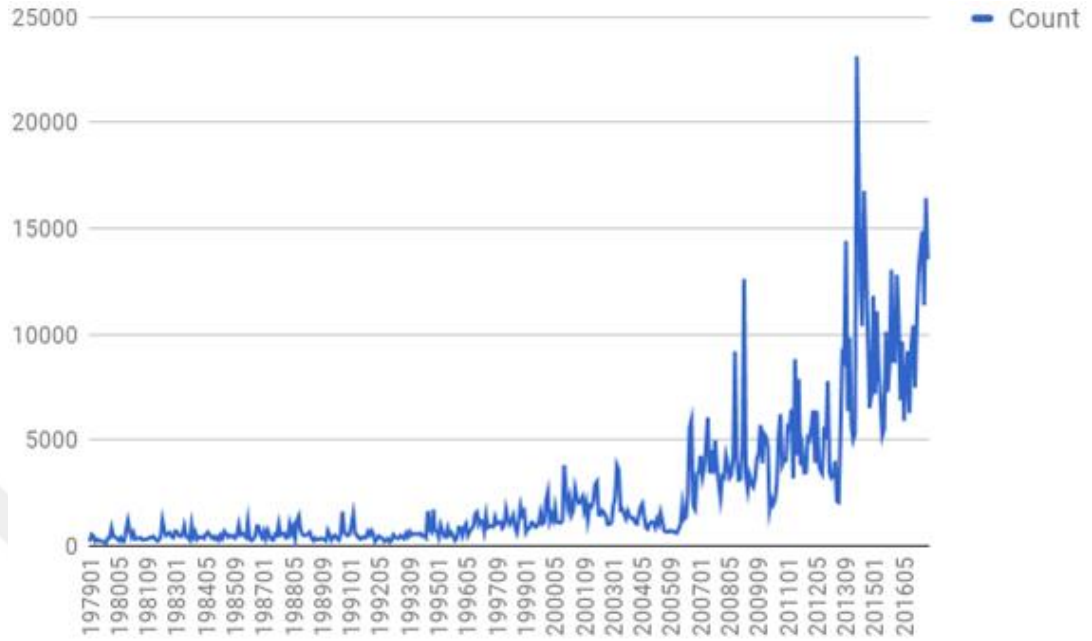
Kaynak: https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html, (26.08.2017).

Analiz sonucunda 1979-2017 yılları arasında dünya çapında gerçekleşen çatışma olaylarını ayrıntılı olarak anlatan 250 milyondan fazla kayıt işlenmiş ve her yıl için en üst düzeyde tanımlayıcı ilişki sayısı bulunmuştur.

Buradaki çatışma olayları; askeri seferberlikler, yardım ya da diplomatik ilişkilerin durdurulması/azaltılması, ambargo, boykot ve yaptırımlar, sokağa çıkma yasağı ve kitlesel gözaltına alınma gibi zorlama ve fiziksel saldırılardır[189].

[189] GDELT PROJECT, "Humanitarian & Crisis Response", 2017, https://www.gdeltproject.org/globaldashboard/, (20.10.2017).

Analiz sonucunda EK 1'deki sonuçlar elde edilmiştir. EK 1'deki sonuçlar grafiğe aktarılırsa; Şekil 4.5'teki grafik elde edilir.



Şekil 4.5: Dünya'daki Çatışma Sayısı Çizgi Grafiği

Şekil 4.5'te aykırı değerlerden 20 Mart 2003'teki "ABD-Irak Savaşı"[190], Ağustos 2008'deki "Güney Osetya Krizi" ile Gürcistan ve Rusya arasında çıkan savaşı[191], 27 Aralık 2008-18 Ocak 2009 tarihleri arasında İsrail'in başlattığı "Gazze Savaşı"[192], 15 Mart 2011'de başlayan " Suriye Savaşı"[193], Kasım 2013'te Ukrayna-Rusya arasında yaşanan kriz ve 2015'te Suriye de devam eden savaşlar belirgin olarak görülmektedir.

Şekil 4.5'teki grafik incelendiğinde grafiğin artan bir trende sahip olduğu görülmektedir. Bu nedenle dünyanın zamanla artan bir düzensizliğe (kaosa) doğru sürüklendiği söylenebilir. Termodinamiğin ikinci yasası olan Entropi, evrendeki herşeyin maksimum düzensizlik ve minimum enerjiye doğru hareket ederek; çoktan aza, sıcaktan soğuğa, doğumdan ölüme, düzenden düzensizliğe veya moleküler rastgeleliğe doğru bir eğilim içinde olduğunu söylemektedir. Bir sistemin entropisi ise sistemin bulunabileceği mikroskobik hallerin toplam sayısına bağlı olup bu sayı aynı zamanda "termodinamik

[190] ÖZŞİMŞİR Şafak, *"2003 ABD-Irak Savaşı ve Nedenleri"*, 30 Ocak 2011. http://www.tuicakademi.org/2003-abd-irak-savasi-ve-nedenleri/, (27.08.2017).
[191] ERKAN Süleyman, *"2008 Rusya-Gürcistan Savaşı ve Uluslararası Toplum"*, Uluslararası İktisadi ve İdari İncelemeler Dergisi, 06.04.2016, s.41-64.
[192] MOLTKE helmuth von, *"İsrail'in Yaptığı 10 Operasyon"*, 18 Temmuz 2014, https://onedio.com/haber/israil-in-gazze-ye-yaptigi-10-operasyon-337586, (27.08.2017).
[193] NTV Haber, *"Suriye iç savaşında 6 yılda neler yaşandı"*, 15 Mart Çarşamba 2017, http://www.ntv.com.tr/dunya/suriye-ic-savasinda-6-yilda-neler-yasandi,Ik0VGc0sPUqbvD, (27.08.2017).

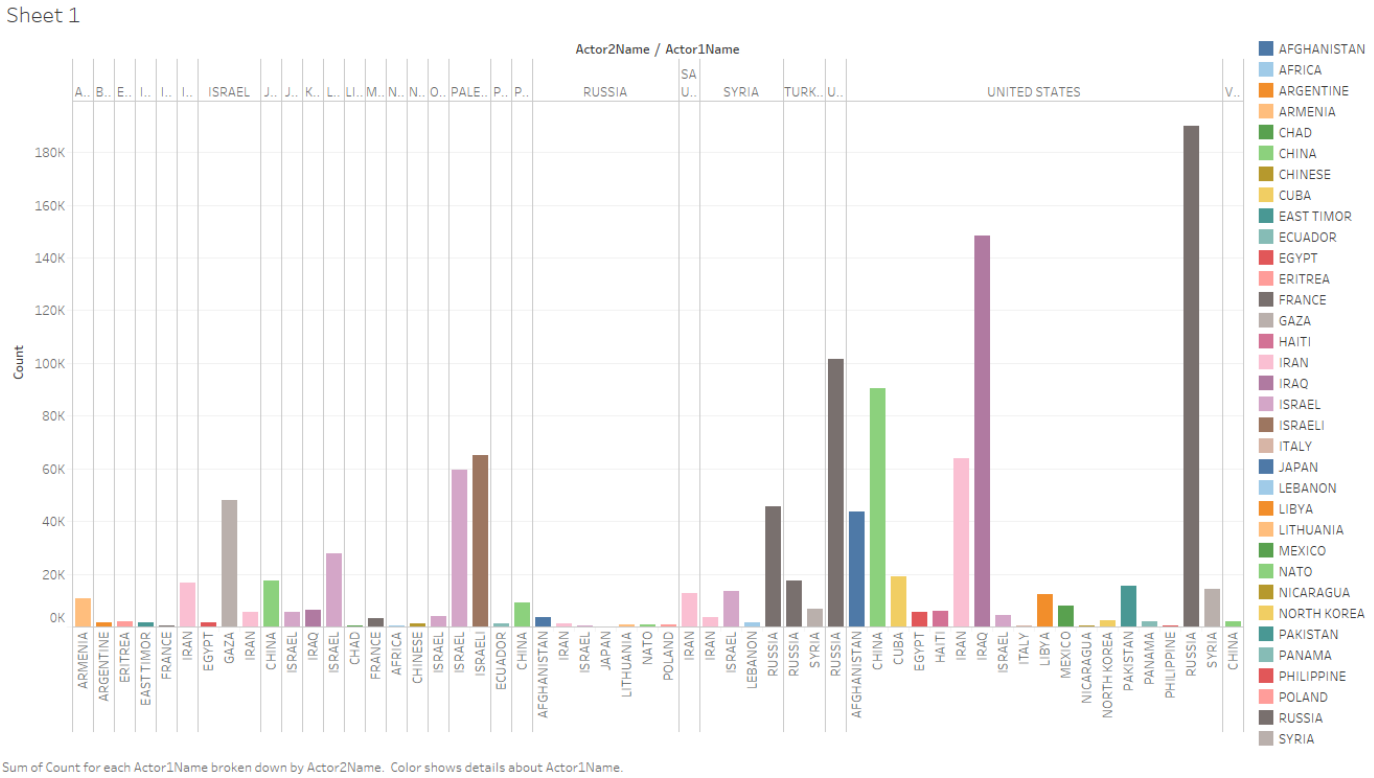
olasılık" olarak bilinmektedir[194]. Termodinamik olasılığın entropi ile ilişkisi aşağıdaki Boltzman eşitliği ile verilir;

$$S = k \ln P$$

Buradaki "S" entropi, k ($k = 1.3806 \times 10^{-23}$) Boltzman sabiti ve "$P$" termodinamik olasılıktır[195].

Bundan dolayı, mikroskopik açıdan sistemin entropisi, düzensizlik arttıkça daha büyük bir değer almaktadır. Zaman oku tek yönde ilerlediğinden entropi azalmaz artarak evren gittikçe genişlemektedir[196]. Dolayısıyla Dünya'daki çatışma, kaos ve düzensizlik sürekli artmaktadır. Bu da önümüzdeki süreçte dünyada yaşanan çatışmaların artarak devam edeceğine işaret etmektedir.

EK 1'deki sonuçlar Tableau programı kullanılarak görselleştirilirse, aşağıdaki Şekil 4.6'daki grafik elde edilir.



Şekil 4.6: Ülkelerarası Çatışma Sayısı Çubuk Grafiği

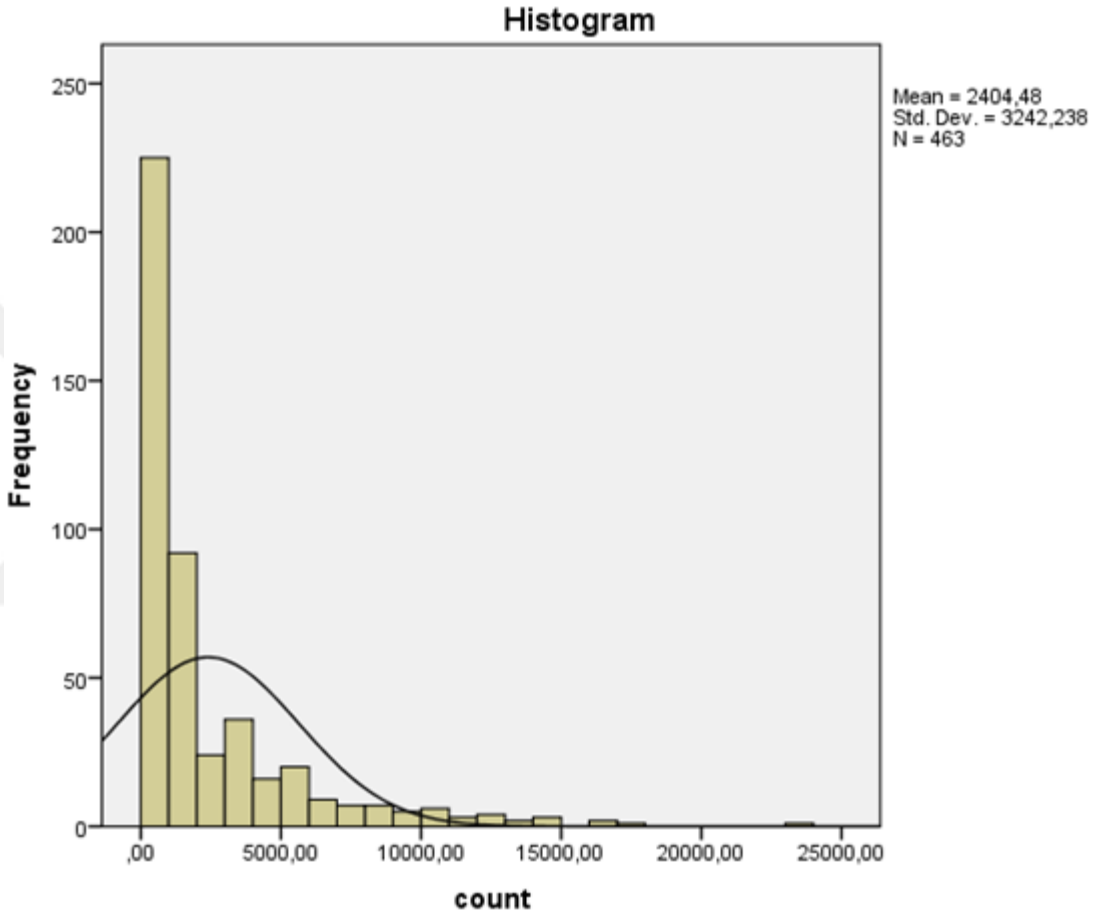
[194] ÇENGEL, Yunus A. - Michael A. BOLES, *"Mühendislik Yaklaşımıyla Termodinamik"*, çev. Taner Derbentli, Literatür Yayıncılık, Đstanbu, 1996.

[195] ÇENGEL, Yunus A. - Michael A. BOLES, *"Mühendislik Yaklaşımıyla Termodinamik"*, çev. Taner Derbentli, Literatür Yayıncılık, Đstanbu, 1996.

[196] GREENE Brian, *"Evrenin Dokusu"*, Tübitak Popüler Bilim Kitapları, 2. Baskı, çevirmen: Murat Alev, Num. 342, ISBN: 978-975-403-547-6, 2013.

Şekil 4.6 incelendiğinde dünyadaki çatışmaların oluşmasında en büyük aktörlerin Amerika ve Rusya olduğunu söylemek mümkündür.

Basit olarak verilerin dağılımını bulmak için verilerin histogram grafiğini çizmek gerekir. Bunun için EK 1'deki sonuçlar kullanılarak histogram çizilirse Şekil 4.7'deki grafik elde edilir.



Şekil 4.7: Dünya'daki Çatışma Sayısı Histogramı

Şekil 4.7'deki histograma göre az sayıda ülke ile ilgili yapılan haber sayısı çok fazla, çok sayıda ülke ile ilgili yapılan haber sayısı azdır. Grafik incelendiğinde, grafiğin sol yukarıdan sağ aşağıya doğru bir çizgi ile tanımlandığı görülmektedir. Buradan haber sayısı verisinin dağılımının Kuvvet Yasası Dağılımı'na uygun olduğu ve dolayısıyla ölçekten bağımsız ağ modeline işaret ettiğini söylemek mümkündür.

4.2.1. Parametre Tahminleri ve Kuvvet Yasası Uygunluk Analizi

Burada, GDELT veri seti üzerinde bazı analizler yapılarak dünyadaki "çatışma" sayılarının kuvvet yasasına uygun bir dağılıma sahip olup olmadığı test edilecektir. Bunun için R yazılımın da bulunan "PoweRlaw" paketi kullanılacaktır. Bu paket, kuvvet yasalarını ve diğer uzun kuyruklu dağılımları basitleştirmeyi amaçlamaktadır. Ayrıca, bu paket, uzun kuyruklu dağılımları uydurma, karşılaştırma ve görselleştirme için R fonksiyonlarını içermektedir.

PoweRlaw paketi, kesikli ve sürekli kuvvet yasası dağılımları için uzun kuyruklu dağılımlara uyacak bir kod sağlamaktadır. Bu çerçevede, çatıma verilerinin kuvvet yasası dağılımına uygun olup olmadığı araştırılacaktır.

Kesikli kuvvet yasası,

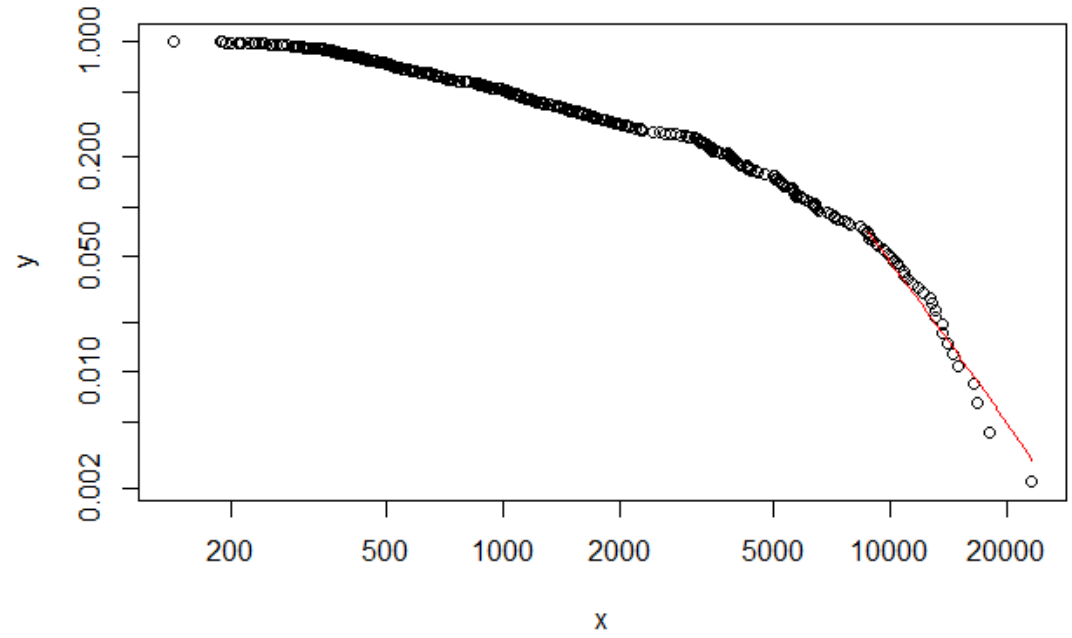
$$p(x) = Cx^{-\alpha} \tag{4.1}$$

şeklindedir. Dağılımın paramerteleri α ve C maksimum olabilirlik yöntemi kullanılarak tahmin edilmiştir. Kesmeli değer, x_{min} ise Kolmogorov-Smirnov istatistiği minimize edilerek tahmin edilmiştir. Maksimum Olabilirlik Tahminleri ve Kolmogorov-Smirnov test sonuçları Tablo 4.1'de verilmiştir.

Tahminciler	$\hat{C}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{n}_{tail}$	$\widehat{D}$ (KS)	$bs_{\hat{p}}$
Tahmin değerleri	1,098	8335	3,789	35	0,089	0,29

Tablo 4.1. Maksimum Olabilirlik Tahminleri ve Kolmogorov-Smirnov Test Sonuçları

Tablo 4.1'deki test sonuçları, x_{min} =8335 ve ölçekleme parametresi α =3,789 olan eşik değer tahminini verir ve sonuçlar Clauset ve arkadaşları (2009) ile eşleşmektedir. Model tahminlerinde hesaplanan Kolmogorov-Smirnov D istatistiğinin küçük bir değer (0,089) olması verilerin kuvvet yasası dağılımına uygun olduğuna işaret etmektedir.



Şekil 4.8: Çatışma Veri Kümesinin Kümülatif Yoğunluk Fonksiyonu (CDF)

Şekil 4.8 incelendiğinde, bütün grafiğin sol yukarıdan sağ aşağıya doğru bir eğri ile tanımlandığı görülmektedir. Bu görsel biçimden verinin dağılımının kuvvet yasası dağılımına uygun olduğu söylenebilir. Ancak hiçbir şekilde yeter koşul değildir.

Herhangi bir veri kümesine bir kuvvet yasası dağılımını uydurmak mümkün olduğu için, gözlenen veri kümesinin gerçekten bir kuvvet yasasını gösterip göstermediğini sınamak gerekmektedir. Clauset, Shalizi ve Newman (2009), bu hipotezin bir bootstrapping yöntemi aracılığıyla uygunluk testi kullanılarak test edilebileceğini söylemektedir. Buna göre, kuvvet yasası dağılımının gözlenen verilere ne kadar iyi uyduğunu test etmek için seçilen parametrelerle kuvvet yasası dağılımından üretilen verilerin aynı dağılımlardan geldiğini görmek için Kolmogorov-Smirnov testi kullanılır. Her bir Kolmogorov-Smirnov testinin red edilip edilmeyeceğini belirlemek için uygun bir anlamlılık düzeyi (0,05 veya 0,01) seçilmektedir. Eğer $p - değeri$ seçilen anlamlılık düzeyinden büyük ise temel hipotez red edilemez ve her iki veri setinin de aynı dağılımı gösterdiği söylenir. Diğer taraftan eğer $p - değeri$ seçilen anlamlılık düzeyinden küçük veya eşitse temel hipoez red edilir ve alternatif hipotez kabul edilip, veri kümelerinin bir kuvvet yasası dağılımından gelmediği söylenir. Eğer, anlamlılık düzeyi yaklaşık olarak 0,01 seçilirse, temel hipotezi reddetmek için test istatistiğinin yaklaşık %10 veya daha az

olması beklenir. Bu durumda, seçtiğimiz parametrelerle kuvvet yasası dağılımı için iyi bir uyum göstermektedir.

$$H_0: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygundur.$$
$$H_1: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygun\ değildir.$$

Tablo 4.1'deki KS test sonuçlarına göre $p = 0{,}29$ olup, anlamlılık seviyesi 0,05 seçilirse $p > 0{,}05$ olduğundan temel hipotez (H_0) red edilmez ve her iki veri setinin de aynı dağılımdan geldiği söylenir. Bu sonuç; Richardson (1948, 1960), Richardson ve Cederman (2003), Newman (2005), Clauset, Young ve Gleditsch (2007) ve Biggs (2016) ile uygunluk göstermektedir.

Burada, çatışma verilerinin kuvvet yasasına uygun çıkması, çatışma olaylarında ülkelerarasında orantısız bir güç etkisi olduğunu göstermektedir. Dolayısıyla Pareto ilkesi olarakta bilinen bu güç dağılımı 80/20 kuralına uygun bir dağılıma sahiptir. Bu bağlamda az sayıda ülkenin dünyadaki çatışmalarda çok büyük bir etkiye sahip olduğu ve bunların dünyadaki çatışmaların büyük çoğunluğunun oluşmasında aktör görevi gördükleri söylenebilir.

4.3. Türkiye'deki Protestoların İncelenmesi

Sosyal hareketler, sosyal bir sorunu çözmek, sosyal bir kurumu yıkmak, değiştirmek veya tâdil etmek amacıyla bir grup şeklinde yapılan hareketlere verilen genel addır[197]. Toplumsal hareketler, 30 Kasım 1999'da Seattle'de yapılan Dünya Ticaret Örgütü zirvesine karşı yapılan kitlesel protestolar ile niteliksel bir artış yaşanmıştır[198]. Bu hareketler, özellikle son dönemlerde belli başlı meydanların işgal edilerek, kollektif eylemler olarak sistem karşıtı popüler hareketler olarak ortaya çıkmıştır. İspanya İndignados hareketi, Yunanistan Aganaktismenoi hareketi, Arap Baharı, Avrupa ve ABD'deki işgal hareketleri, Türkiye Gezi Olayları, Ukrayna Turuncu Devrimi gibi meydan hareketleri kollektif eylemler ve toplumsal hareketler niteliksel bir değişimin ve dönüşümün yaşanmasına sebep olabilmektedir. Protestolar, 1960'ların sonlarında ortaya

[197] SOSYALBİLİMLER, *"Sosyal Hareketler"*, 2017, http://www.enfal.de/sosyalbilimler/s/055.htm, (30.08.2017).
[198] KALFA Ceren – Faruk ATAAY, *"Küresel Toplumsal Hareketler"*, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergis (16) 2008/2:127-149.

çıkan ve 1970'lerin ortalarından beri çeşitlenerek direniş mücadeleleriyle birlikte çokça tartışılmaya başlandı. Halk hareketlerinin yeni ibaresi, genellikle, hak, adalet, özgürlük, demokrasi, çevre, ekoloji, feminizm ve barış gibi temelerdir[199].

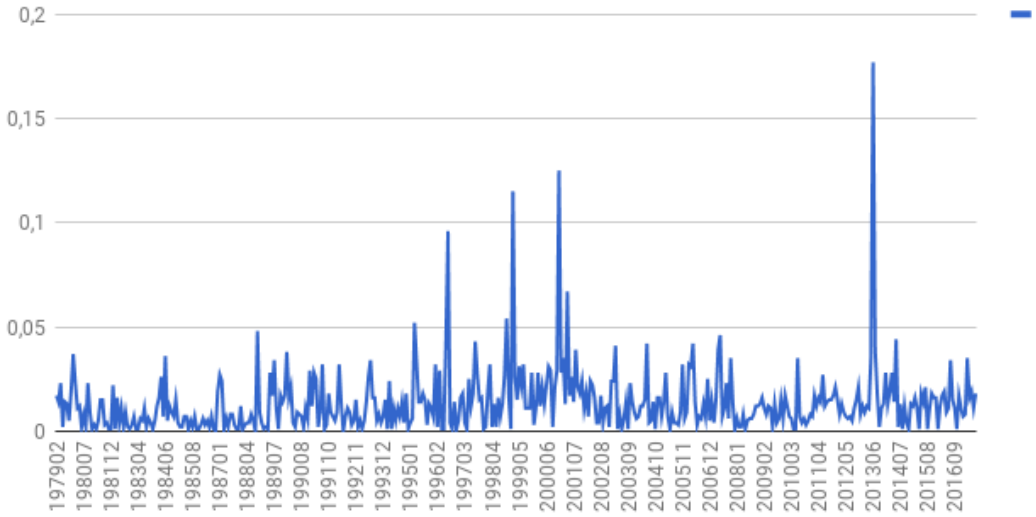
Burada Türkiye tarihinde yaşanan toplumsal protesto hareketleri GDELT'in gözünden incelenerek analiz edilecektir. Buradaki protestolar, haber medyasında bir "protesto" ya da "gösteri" olarak tanımlanan herhangi bir topluluğa atıfta bulunmaktadır. SQL de aşağıdaki kod yazılıp çalıştırıldığında,

```
SELECT MonthYear MonthYear, INTEGER(norm*100000)/1000 Percent
FROM (
SELECT ActionGeo_CountryCode, EventRootCode, MonthYear, COUNT(1) AS c,
RATIO_TO_REPORT(c) OVER(PARTITION BY MonthYear ORDER BY c DESC) norm
FROM [gdelt-bq:full.events]
GROUP BY ActionGeo_CountryCode, EventRootCode, MonthYear
)
WHERE ActionGeo_CountryCode='TU' and EventRootCode='14'
ORDER BY ActionGeo_CountryCode, EventRootCode, MonthYear;
```

Kaynak: https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html, (28.08.2017).
Not: Farklı ülkelerdeki farklı olayları analiz etmek için yukarıdaki koda istenilen ülkenin ve olayın kodunu[200] yazarak yapmak mümkündür.

EK 2'deki sonuçlar elde edilmiştir. Bu sonuçlar kullanılırsa, Şekil 4.9'daki grafik elde edilir.

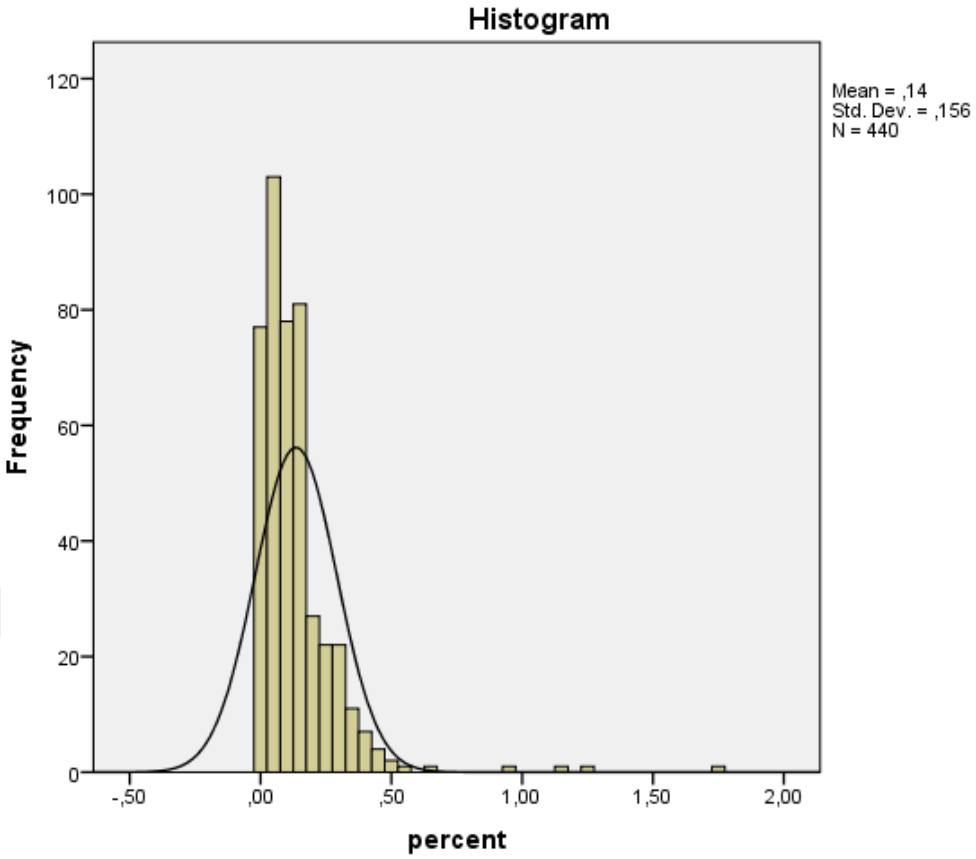
[199] ÖZEN Hayriye, *"Meydan Hareketleri ve Eski ve Yeni Toplumsal Hareketler"*, Mülkiye Dergisi, 2015, 39(2), 11-40.
[200] CAMEO, *"Conflict and Mediation Event Observations Event and Actor Codebook"*, Event Data Project Department of Political Science Pennsylvania State University Pond Laboratory University Park, PA 16802, Philip A. Schrodt (Project Director), March 2012



Şekil 4. 9: Türkiye'deki Protesto Yoğunluk Grafiği

Şekil 4.9'daki aykırı değerlerden 1996'daki Öğrenci Protestoları, 1998 yılında doğu illerindeki "9 Ekim Komplosu" protestoları, 19 Aralık 2000 "Hayata Dönüş Operasyonları" protestoları ve 31 Mayıs 2013'te Taksim Gezi Parkı'nda yapılan "Gezi Parkı" protestolarının yoğun olarak yaşandığı görülmektedir.

Bununla birlikte verilerin dağılımını bulmak için verilerin histogram grafiğini çizmek gerekmektedir. EK 2'deki sonuçlar kullanılırsa aşağıdaki histogram elde edilir.



Şekil 4.10: Türkiye'deki' Protestoların Histogramı

Şekil 4.10 incelendiğinde, düşük şiddetteki protestoların sayısının çok fazla, yüksek şiddettekilerin sayısının çok az olduğu görülmektedir. Grafik incelendiğinde, grafiğin sol yukarıdan sağ aşağıya doğru bir çizgi ile tanımlandığı görülmektedir. Buradan protesto verisinin dağılımının kuvvet yasası dağılımına uygun olduğu söylenebilir.

4.3.1. Parametre Tahminleri ve Kuvvet Yasası Uygunluk Analizi

Burada Ocak 1979 ve Haziran 2017 arasındaki dönemde Türkiye'deki protesto yoğunluğunun kuvvet yasasına uygun bir dağılım sergileyip sergilemediği araştırılacaktır. Bu çerçevede, sürekli kuvvet yasası,

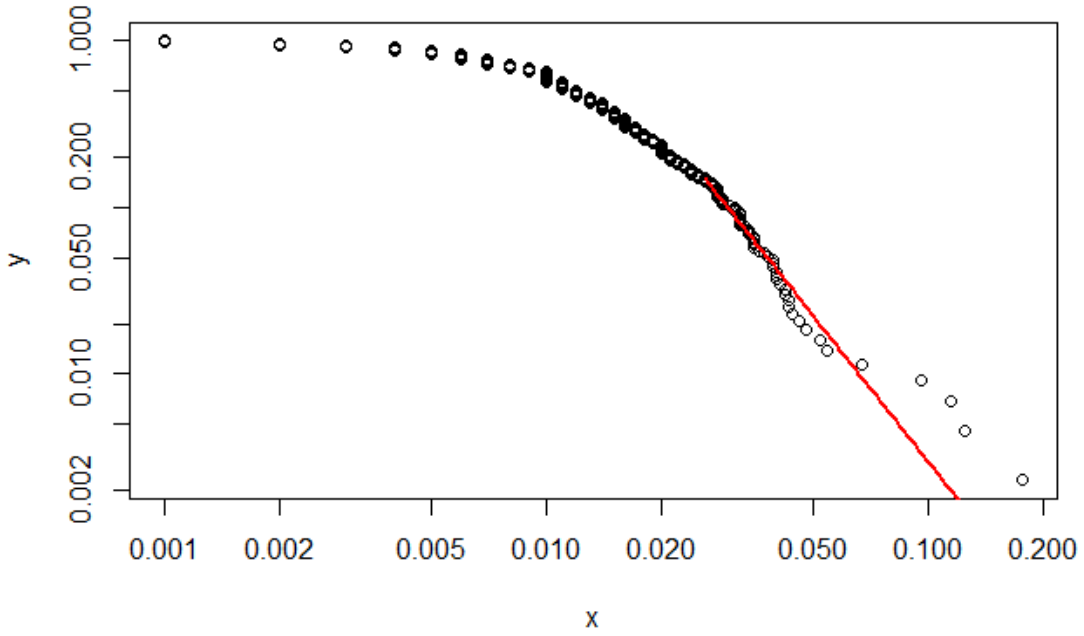
$$p(x) = Cx^{-\alpha} \tag{4.2}$$

şeklinde olup, burada $C = (\alpha - 1)x_{min}^{\alpha-1}$ dir. Model parametrelerinin tahmini yapıldığında, Tablo 4.2'deki sonuçlar bulunmuştur.

Tahminciler	$\hat{C}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{n}_{tail}$	$\hat{D}$ (KS)	$bs_{\hat{p}}$
Tahmin değerleri	1,089	0,026	3,911	65	0,071	0,320

Tablo 4.2: Maksimum Olabilirlik Tahminleri ve Kolmogorov-Smirnov Test Sonuçları

Tablo 4.2 incelendiğinde, $x_{min} = 0{,}026$ ve α =3,911 olan eşik değer tahminleri bulunmuştur. Burada Kolmogorov-Smirnov D istatistiğinin küçük bir değer (0,071) olması verilerin kuvvet yasası dağılımına uygun olduğuna işaret etmektedir.



Şekil 4.11: Protesto Veri Kümesinin Kümülatif Yoğunluk Fonksiyonu (CDF)

Şekil 4.11 incelendiğinde, bütün grafiğin sol yukarıdan sağ aşağıya doğru bir eğri ile tanımlandığı görülmektedir. Bundan dolayı verinin dağılımının kuvvet yasası dağılımına uygun olduğu söylenebilir. Fakat bu hiçbir şekilde yeterli koşul değildir.

Verilerin kuvvet yasasına uygunluğunu test etmek için Kolmogorov-Smirnov testi kullanılır.

$$H_0: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygundur.$$
$$H_1: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygun\ değildir.$$

Tablo 4.2'de görüldüğü üzere $p = 0{,}320$ bulunmuştur. Özel olarak anlamlılık düzeyi 0,05 alınırsa, $p > 0{,}05$ olduğundan, verilerin dağılımı kuvvet yasasına uygundur. Buradan az sayıdaki protestonun çok yoğun ve çok sayıdaki protestonun az bir yoğunluğa sahip olduğu söylenebilir. Buda bize protesto nedenlerinin çoğunun veya bir protestodaki katılımcı sayısının çoğunun az sayıdaki gözlem tarafından açıklanabileceğini göstermektedir.

4.4. Ukrayna'daki Protestoların İncelenmesi

Ukrayna tarihi boyunca her dönemde protesto şiddetinin yoğun olarak yaşandığı ülkelerin başında gelmektedir. Dünya haber medyasında Ukrayna'daki protestolarla ilgili yoğun olarak haberler yapılmakta ve bu haberler dünya haber medyasını takip eden GDELT'in yapısında işlenmektedir. Bugün, Google'ın BigQuery sistemini kullanarak, protestoların yerini ve yoğunluğunu (medya hacminin öngördüğü şekilde) ve sivillere yönelik şiddeti tanımlamak mümkündür. Tüm bunları yalnızca birkaç saniye içinde çeyrek milyar GDELT kaydını tarayarak gözlemlemek mümkündür. GDELT, tüm dünyada, 1 Kasım 2013'ten Temmuz 2017'ye kadar Ukrayna'daki protesto gösterileri ve şiddet olayları ile ilgili 2 milyondan fazla haberi yayınlamıştır.

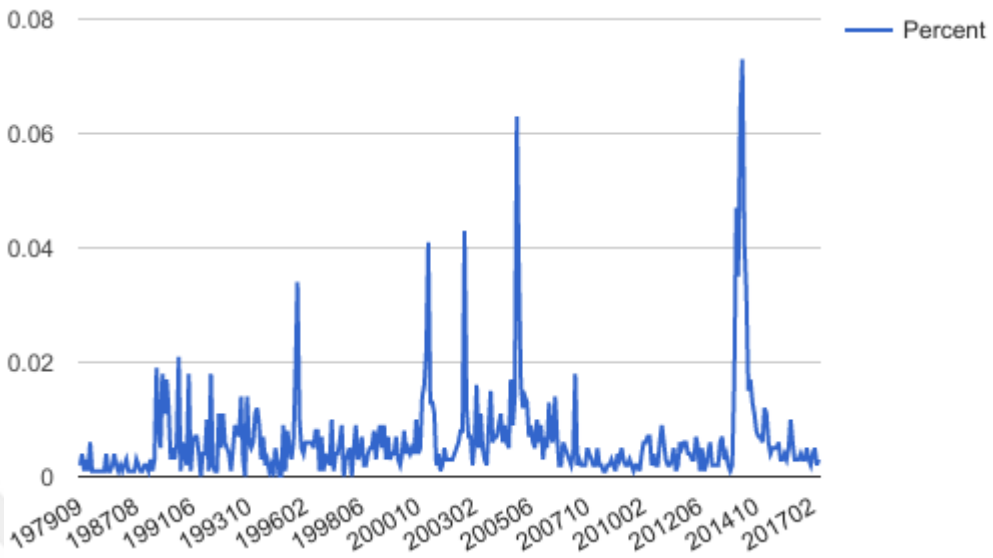
Ukrayna'daki protesto şiddetinin aylık sayısını normalleştirilmiş bir "yoğunluk" ölçüsüne dönüştürmek için SQL de aşağıdaki sorgulama yapılır.

```
SELECT MonthYear MonthYear, INTEGER(norm*100000)/1000 Percent
FROM (
SELECT ActionGeo_CountryCode, EventRootCode, MonthYear, COUNT(1) AS c,
RATIO_TO_REPORT(c) OVER(PARTITION BY MonthYear ORDER BY c DESC) norm
FROM [gdelt-bq:full.events]
GROUP BY ActionGeo_CountryCode, EventRootCode, MonthYear
)
WHERE ActionGeo_CountryCode='UP' and EventRootCode='14'
ORDER BY ActionGeo_CountryCode, EventRootCode, MonthYear;
```

Kaynak: https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html, (28.08.2017).

Analiz sonucunda aylık olarak 1979-2017 yılları arasında Ukrayna da gerçekleşen protesto olaylarını ayrıntılı olarak anlatan 2 milyondan fazla kayıt işlenmiştir. Elde edilen

sonuçlar EK 3'te verilmiştir. EK 3'teki sonuçlardan Şekil 4.12'deki çizgi grafiği elde edilir.

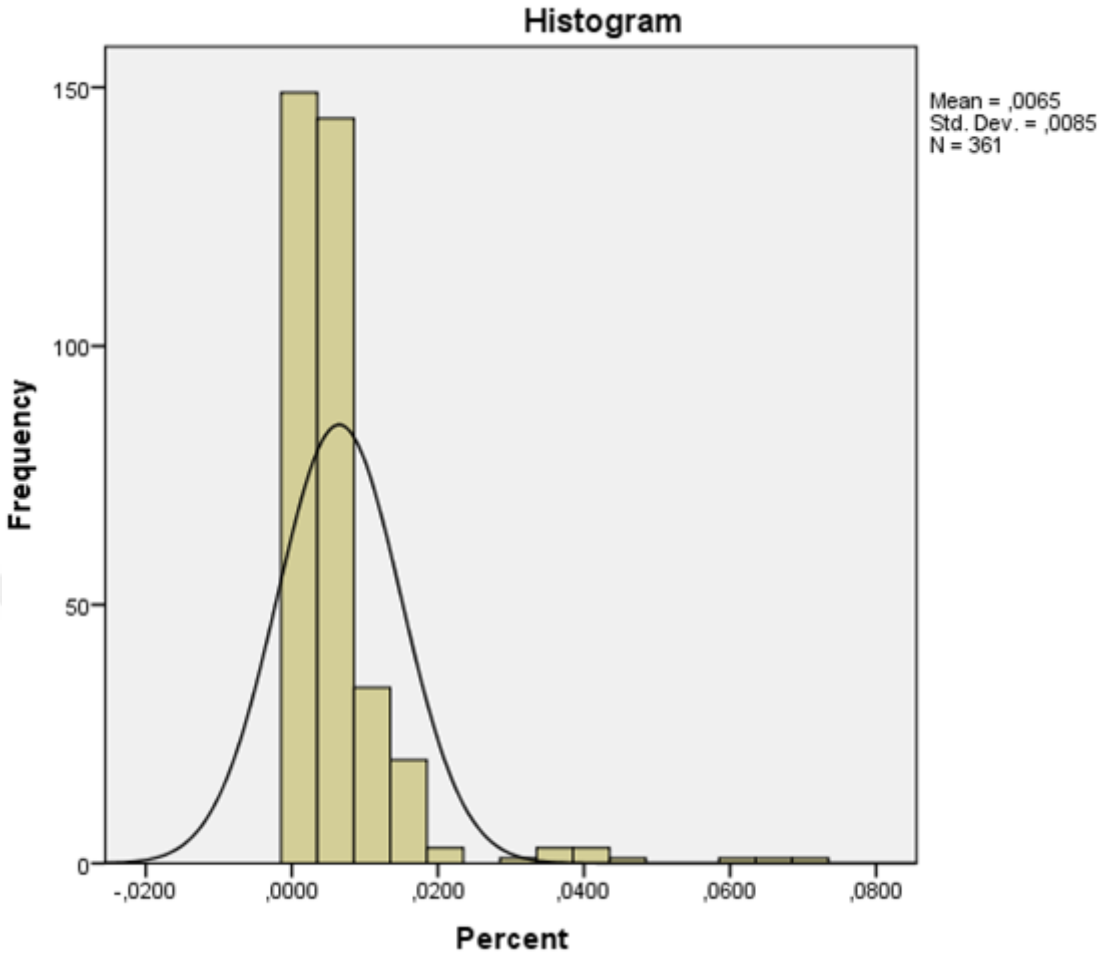


Şekil 4.12: Ukrayna'daki Protesto Yoğunluk Grafiği

Şekil 4.12'ye göre BigQuery, çok daha güçlü bir şekilde, protestoların aylık sayısını normalleştirilmiş bir "yoğunluk" ölçüsüne dönüştürmüştür. Bu sonuçların, 2017'de 1979'a göre daha belirgin olması bugünkü haber medyasının çok daha fazla olması ve makinelerin haberlerin işlenmesinde etkin rol oynamasından kaynaklanmaktadır.

Şekil 4.12'deki aykırı değerlerden "1989 Devrimleri", Mart 2001'deki "Kuçma'sız Ukrayna" protestoları, Kasım 2004'teki "Turuncu Devrimi", ve Kasım 2013'te Ukrayna'nın başkenti Kiev'deki "Euromaidan" protestoları belirgin olarak görülmektedir.

Basit olarak verilerin dağılımını bulmak için verilerin histogram grafiğini çizmek gerekmektedir. EK 3'teki sonuçların histogram grafiği çizilirse, Şekil 4.13 elde edilir.



Şekil 4.13: Ukrayna'daki Protestoların Histogramı

Şekil 4.13'teki histogram grafiği incelendiğin de düşük şiddetteki protestoların sayısının çok fazla, yüksek şiddettekilerin sayısı çok azdır. Ayrıca, grafiğin sol yukarıdan sağ aşağıya doğru bir çizgi ile tanımlandığı görülmektedir. Buradan protesto verisinin dağılımının kuvvet yasası dağılımına uygun olduğu söylenilebilir.

4.4.1. Parametre Tahminleri ve Kuvvet Yasası Uygunluk Analizi

Burada veri setindeki değerler sürekli olduğundan (4.3) denklemindeki normalleştirme sabiti $C = (\alpha - 1)x_{min}^{\alpha-1}$ şeklinde hesaplanır. Bu çerçeve de, sürekli bir veri seti için kuvvet yasası,

$$p(x) = Cx^{-\alpha} \tag{4.3}$$

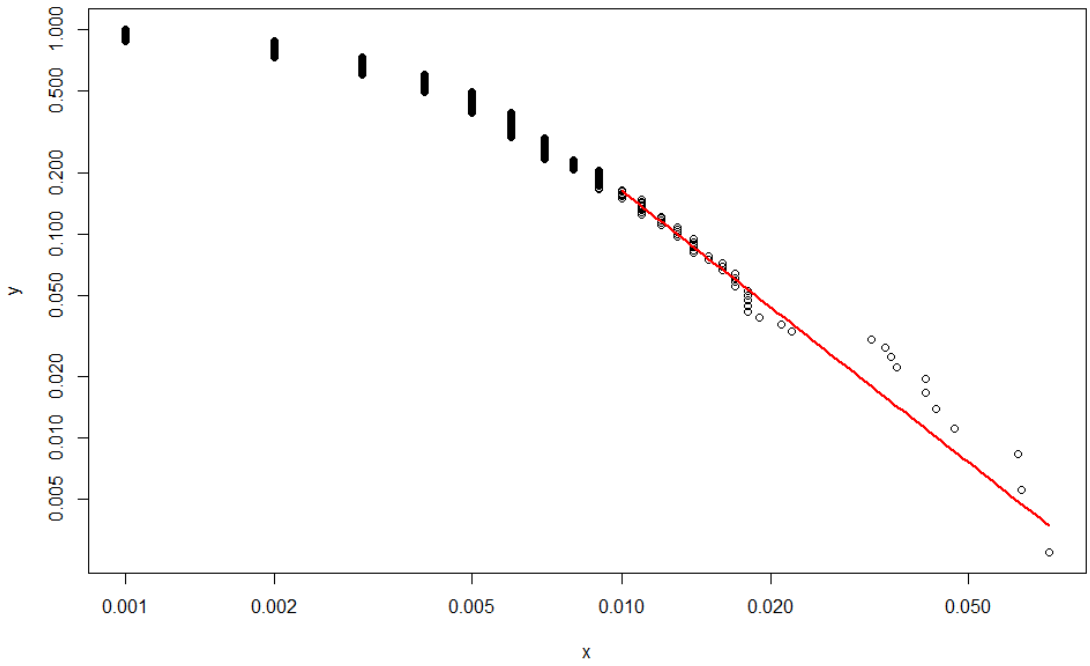
şeklindedir.

Buradaki α ve C değerleri maksimum olabilirlik yöntemi kullanılarak tahmin edilmiştir. Kesme parametresi x_{min} Kolmogorov-Smirnov D istatistiği minimize edilerek tahmin edilmiştir. Analiz sonucunda aşağıdaki değerler bulunmuştur.

Tahminciler	$\hat{C}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{n}_{tail}$	$\widehat{D}$ (KS)	$bs_{\hat{p}}$
Tahmin değerleri	1,222	0,01	2,904	59	0,085	0,26

Tablo 4.3: Maksimum Olabilirlik Tahminleri ve Kolmogorov-Smirnov Test Sonuçları

Tablo 4.3 incelendiğinde, $x_{min} = 0{,}01$ ve α =2,904 olan eşik değer tahminleri bulunmuştur. Model tahminlerinde hesaplanan Kolmogorov-Smirnov D istatistiğinin küçük bir değer (0,085) olması verilerin kuvvet yasası dağılımına uygun olduğuna işaret etmektedir.



Şekil 4.14. Protesto Veri Kümesinin Kümülatif Yoğunluk Fonksiyonu (CDF)

Şekil 4.13 incelendiğinde, bütün grafiğin sol yukarıdan sağ aşağıya doğru bir eğri ile tanımlandığı görülmektedir. Bu görsel biçimden verinin dağılımının kuvvet yasası dağılımına uygun olduğu söylenebilir. Fakat bu hiçbir şekilde yeter koşul değildir.

Kuvvet yasası dağılımının gözlenen verilere ne kadar iyi uyduğunu test etmek için seçilen parametrelerle kuvvet yasası dağılımından üretilen verilerin aynı dağılımlardan geldiğini görmek için Kolmogorov-Smirnov testi kullanılır.

$$H_0: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygundur.$$
$$H_1: Verilerin\ dağılımı\ kuvvet\ yasasına\ uygun\ değildir.$$

Tablo 4.3'ten de görüleceği üzere Kolmogorov-Smirnov test sonuçlarından, $p = 0{,}26$ bulunmuştur. Özel olarak anlamlılık düzeyi 0,05 alınırsa, $p > 0{,}05$ olduğundan, verilerin dağılımı kuvvet yasasına uygun olup bu sonuçlar; Richardson ve Cederman (2003), Newman (2005), Clauset, Shalizi ve Newman (2009) ve Biggs (2016) uygundur. Böylece az sayıdaki protestonun büyük bir etkiye ve çok sayıdakinin az bir etkiye sahip olduğunu söylemek mümkündür. Ayrıca, elde edilen bu bulgulardan Türkiye ve Ukrayna'daki fiiliverilerin kuvvet yasası dağılımlarına sahip olduğu e kuvvet yasası ilişkisinin ölçeğe göre değişmediği, yani farklı yerlerden elde edilen sonuçların aynı dağılım özelliklerine sahip olacağını göstermektedir.

SONUÇ

21. yüzyılda insan aklı ve gelişen sofistike makineler dünyada büyük değişimlere sebep oldu. Bilişim Çağı olarak adlandırılan bu dönemde cep telefonu ve internetin hayatımıza girmesiyle birlikte üretilen ve saklanan bilginin miktarında ve hızında büyük bir artış yaşandı. Bilgi miktarındaki bu artış "Büyük Veri" olarak adlandırılan yeni bir terminolojiyi ortaya çıkarmıştır. Büyük Veriler, çoğunlukla bilgisayar veri tabanlarına kaydedilir ve büyük, karmaşık veri setlerini işlemek üzere özel olarak tasarlanmış bir yazılım kullanılarak analiz edilir. Büyük Veri'nin ortaya çıkmasıyla birlikte birçok bilimsel çalışmada "Veri Bilimi" terimi daha fazla kullanılmaya başlandı. Tüm bu değişimler istatistiğin karşısına "Veri Bilimi" olarak adlandırılan yeni bir ilgi alanını ortaya çıkarmıştır. "İstistik" ve "Veri Bilimi" birbiri ile o kadar ilişkilidir ki bunları birbirinden ayırmak artık mümkün değildir. Bugün artık "İstatistik" sözcüğünün kullanıldığı her yerde "Veri Bilimi" de kullanılmaktadır. Bilişim ve Bulut teknolojilerinde yaşanan gelişmeler önümüzdeki süreçte "İstatistik" ve "Veri Bilimi"nin birlikte kullanılmaya devam edeceğini göstermektedir.

Bugün içinde bulunduğumuz çağda bilgi devrimi yaşanmaktadır. Herhangi bir ekonomik devrim gibi toplumun, akademisyenlerin ve işletmelerin bu dönüşümde büyük bir etkisi olmuştur. Günümüzdeki devrim, ağa bağlı iletişim sistemleri ve internet tarafından üretilen veriler sayesinde hepimizi tüketici ve üreticiye dönüştürmesiyle benzersizdir. Bu veriler, yediğimiz gıdalardan, sosyal etkileşimlerimize, çalıştığımız ve oynadığımız yere kadar hayatımızın her alanını giderek daha çok etkilemektedir. Buna karşılık, son derece kişiselleştirilmiş ve vücudumuza, hayatlarımıza ve işletmelerimize ince ayar yapan ürünler ve hizmetler için makul bir beklenti geliştirilmiş ve veri ürünü olarak yeni bir bilgi teknolojisi pazarı yaratılmıştır. Büyük veri kümelerinin makine öğrenme algoritmalarıyla hızlı ve çevik birleşimi, insanların gündelik işleriyle ve birbirleriyle etkileşim biçimini değiştirerek, hızlı ve yeni sonuçlar doğurmuştur. Aslında, büyük veriyi çevreleyen şüphecilik eğilimi, çok sayıda model ve veri kaynağı mevcut olan sürekli yenilikle ilgilidir.

Eskiden, kitaplarda olmayan herhangi bir istatistiksel teknik, istatistiksel olmayan veya çoğu zaman anlaşılmaz olarak kabul edilen tekniklerdi. Bu teknikler: KNN kümeleme, lojistik regresyon, naive Bayes, karar ağaçları, boosted ağaçları, SVM, Bayes istatistiği, merkezi kümeleme ve lineer diskriminantdır. 1980'lerin başlarına gelindiğinde

ise, yeni istatistikler veya sahte veri bilimi ders kitapları her hafta yayınlanır ve bu yöntemler Fisher'ın İris veri seti gibi küçük verilere uygulanırdı. Yazarlar Ensemble yöntemleri, Lasso ve Ridge Regresyonu veya Bayes Ağları gibi yeni tekniklere kitaplarında yer vermezdi. Yeni istatistik kitaplarında, R veya bazen Python'u kullanan küçük Twitter verilerini veya ilişkilendirme kuralları veya öneri motorları hakkındaki bilgileri içeren bazı istatistik kitapları bulunmaktaydı, ancak gerçek uygulanan istatistikler hala veri biliminden uzaktır. Modern istatistiksel veri bilim teknikleri, geleneksel istatistiklerden çok daha iyidir ve bu teknikler büyük veri setleri için tasarlanmıştır. Son 15 yılda Veri Bilimi büyük mesafe kat etmiştir. Bugün Veri Bilimi, keşif amaçlı veri analizi ve istatistiksel bilimin otomatikleştirilmesi konusunda iyi bir noktaya ulaşmıştır. Sonuç olarak, önümüzdeki süreçte veri bilimine olan ilginin daha da artacağı ve veri bilimcilerin iş hayatında önemli bir konuma sahip olacaklarını söylemek mümkündür.

Günümüzde Büyük Veri Yönetimi her zamankinden daha fazla önem kazanmıştır. Bunun temel nedeni işletmelerin geçmişe göre veriye daha çok önem vermesidir. Böyle bir etkiye sebep olan başlıca trendler: ilişkisel olmayan veritabanları, Büyük Veri işleme, Semantik teknolojiler, Nesnelerin İnternet'i gibi bir dizi yeni gelişmelerdir. Tüm bu gelişmeler veri yönetimi sektörünün artan teknik karmaşıklığını eski veriyi yeni veri türleriyle birleştirmeye daha fazla ihtiyaç duymasından kaynaklanmaktadır. Bu gelişmelerle birlikte, kurumsal kullanıcılar tarafından düzenleyici ve uyumluluk sorunları, farklı veri yönetimi, veri kalitesi ve işletmelerdeki diğer girişimler doğrudan veri ile çalışmak için daha büyük bir baskıya maruz kalmıştır. Söz konusu bu baskılar, Büyük Verileri daha doğru, açık ve güvenilirlikte yönetmek için giderek artan bir zorunluluk yaratmıştır. Tüm bu sorunları çözmek için Hadoop gibi Büyük Veri analizi yapan teknolojiler geliştirilmiştir. Hadoop ilk ortaya çıktığında hızla Büyük Veriler ile özdeşleşmeye başladı ancak yine de istenen seviyeye gelmemiş ve Büyük Veri projelerinin birçoğu başarısız olmuştur. Çünkü Hadoop'un yapısı idrak edilen ve düşünülenden çok daha karmaşıktı. Hadoop, EşleİndirgeEşleİndirge ile bu sorunların üstesinden gelmeye çalışmış ama, başarısız olmuştur. Fakat şu an Spark sayesinde Hadoop büyük ölçüde bu sorunların üstesinden gelmiştir. Hadoop sayesinde şirketler bugün hem toplu işleme hem de veri akışı için oldukça başarılı analizler yaparak sorunlara ekonomik ve genel amaçlı çözümler sunmaktadır. Bu bağlamda Hadoop ve Spark gibi yeni teknolojiler

sayesinde işletmeler büyük miktardaki yapısal olmayan veriyi eş zamanlı olarak analiz etmekte, analiz sonuçlarına göre politikalar geliştirmekte ve bu sayede sermaye oranlarını büyük ölçüde arttırmaktadırlar. Sonuç olarak, şirketlerin gelecekte rakipleriyle mücadele edebilmesi ve ayakta kalabilmesi için Büyük Veri Teknolojilerine daha fazla yatırım yapmaları gerekmektedir.

Bu çalışmanın uygulama bölümünde Google'ın altyapısında bulunan BigQuery'deki GDELT veri seti kullanılmıştır. Ocak 1979 ve Haziran 2017 yılları arasında GDELT'te aylık bazda kaydedilen dünyadaki çatışmalar ile Türkiye ve Ukrayna'daki protestolar incelenmiş ve bu çatışma ve protestoların kuvvet yasasına uygun bir dağılım sergileyip sergilemediği test edilmiştir. Uygulamada Clauset, Shalizi ve Newman (2009) tarafından geliştirilen istatistiksel teknikler kullanılmıştır. Ayrıca çalışma da çatışma ve protesto boyutunu tahmin etmek için yeni yaklaşım sağlamak için bu tekniklerden yararlanılmıştır.

Öncelikle Google'ın altyapısında bulunan BigQuery'de bir Google API Console projesi oluşturulmuştur. Daha sonra BigQuery'ye bağlanarak SQL ile buradaki GDELT veri seti üzerinde bazı analizler yapılmıştır. Analiz sonucunda Ocak 1979'dan Haziran 2017'ye kadar aylık bazda aktör ülkeler arasındaki çatışma olayları ile ilgili yapılan en çok haber sayıları bulunmuştur. Bu çatışma sayısı verisinin zaman grafiği çizildiğinde aykırı değerlerden en yoğun çatışma dönemleri: 20 Mart'ta 2003'teki "ABD-Irak Savaşı", Ağustos 2008 de "Güney Osetya Krizi" ile Gürcistan ve Rusya arasındaki savaş, 27 Aralık 2008 ve 18 Ocak 2009 tarihleri arasında İsrail'in başlattığı "Gazze Savaşı", 15 Mart 2011'de başlayan "Suriye Savaşı", Kasım 2013'te Ukrayna-Rusya arasında yaşanan kriz ve 2015'te Suriye de devam eden savaşlar bulunmuştur. Çatışma grafiği incelendiğinde grafiğin artan bir trende sahip olduğu görülmüştür. Buradan, dünyanın giderek daha büyük bir çatışma (kaosa) ortamına doğru sürükleneceği söylenebilir. Ayrıca, Tableau programı yardımıyla çizilen çatışma sayıları grafiği incelendiğinde dünyadaki çatışmaların oluşmasında en büyük aktörün ABD olduğu ve onu ikinci sırada Rusya'nın takip ettiği bulunmuştur. Dolayısıyla buradan dünyadaki çatışmalarda ABD ve Rusya gibi az sayıda ülkenin çok büyük bir etkiye sahip olduğu ve çok sayıda ülkenin az bir etkiye sahip olduğu sonucuna varılmıştır. Son olarak R programı ile çatışma verilerinin kuvvet yasasına uygunluğu test edilerek, modelin ölçekleme parametresi $\alpha = 3{,}789$, normalleştirme sabiti $C = 1{,}089$ ve KS test sonuçlarından $p = 0{,}29$ bulunmuştur.

Anlamlılık düzeyi %5 seçilmiş ve $p > 0{,}05$ olduğundan her iki veri setinin aynı dağılımdan geldiği, yani çatışma verilerinin kuvvet yasasına uygun bir dağılım sergilediği sonucuna varılmıştır. Buda bize az sayıda ülkenin (ABD ve Rusya gibi) çatışmalarda büyük bir etkiye sahip olduğu ve çok sayıda ülkenin ise az bir etkiye sahip olduğunu göstermektedir.

Daha sonra, GDELT veri seti üzerinde Türkiye'deki protesto şiddetinin aylık sayısını normalleştirilmiş bir "yoğunluk" ölçüsüne dönüştürmek için SQL de bazı sorgulamalar yapılmıştır. Elde edilen protesto yoğunluklarının grafiği çizildiğinde aykırı değerlerden: 1996'daki Öğrenci Protestoları, 1998 yılında doğu illerindeki "9 Ekim Komplosu" protestoları, 19 Aralık 2000 "Hayata Dönüş Operasyonları" protestoları ve 31 Mayıs Taksim Gezi Parkı'nda yapılan "Gezi Parkı" protestoları belirgin şekilde görülmüştür. Protesto verisinin histogramından, yüksek şiddetteki protestoların çok az, düşük şiddettekilerin çok fazla olduğu bulunulmuştur. Ayrıca, histogram grafiğinin sol yukarıdan sağ aşağıya doğru bir çizgi ile tanımlanmış olması verinin dağılımının kuvvet yasasına uygun olduğuna işaret etmektedir. Parametre tahminleri yapıldığında modelin ölçekleme parametresi α =3,911, normalleştirme sabiti $C = 1{,}089$ ve $p = 0{,}32$ olarak bulunmuştur. Anlamlılık düzeyi %5 olarak seçilirse, $p > 0{,}05$ olduğundan, verilerin dağılımının kuvvet yasasına uygun olduğu bulunmuştur. Bu sonuç bize protestolardaki katılımcıların çoğunun veya nedenselliklerin çoğunun az sayıda gözlem ile ifade edilebileceğini göstermektedir.

Son olarak GDELT veri seti üzerinde Ukrayna'daki protesto şiddetinin aylık sayısını normalleştirilmiş bir "yoğunluk" ölçüsüne dönüştürmek için SQL de birtakım sorgulamalar yapılmıştır. Sorgulama sonucunda elde edilen protesto yoğunluklarının grafiği çizildiğinde aykırı değerlerden: "1989 Devrimleri", Mart 2001'deki "Kuçma'sız Ukrayna" protestoları, Kasım 2004'teki "Turuncu Devrimi", ve Kasım 2013'te Ukrayna'nın başkenti Kiev'deki Euromaidan protestoları belirgin olarak görülmüştür. Protesto verilerinin histogramından, yüksek şiddetteki protestoların çok az, düşük şiddettekilerin çok fazla olduğu bulunmuştur. Yine histogram grafiğinin sol yukarıdan sağ aşağıya doğru bir çizgi ile tanımlanmış olması verinin dağılımının kuvvet yasasına uygun olduğuna işaret etmektedir. Parametre tahminleri yapıldığında ise modelin ölçekleme parametresi α =2,904, normalleştirme sabiti $C = 1{,}222$ ve $p = 0{,}26$ olarak bulunmuştur. Özel olarak anlamlılık düzeyi %5 seçilirse, $p > 0{,}05$ olduğundan, verilerin

dağılımının kuvvet yasasına uygun olduğu bulunmuştur. Böylece, protestoların büyük kısmının az bir bölümüyle açıklandığı sonucuna varılır.

Çatışma ile ilgili yapılan geçmiş araştırmalar, savaşlar gibi büyük ölçekli olaylara odaklanma eğilimindeydi; bu olaylar, savaşların boyutlarına veya şiddetlerine göre değil, sıklığına veya yokluğuna göre iki ayrı şekilde karakterize edilmiştir. Son zamanlarda çatışma olayları, Cederman (2003) tarafından savaşların ve devlet oluşumunun modellenmesi ve Lacina (2006) tarafından iç savaşların modellenmesi için kullanılmıştır. Ayrıca, bir olayın şiddetinin muhasebeleştirilmesi, politika yapıcılara önemli ölçüde rehberlik sağlayabilir; Örneğin, Cioffi-Revilla (1991), 1991'deki Basra Körfezi Savaşı'nın büyüklüğünü (toplam savaşçı ölümlerini) doğru biçimde öngörmüş ve böylece, savaşın siyasi sonuçlarının tahmin edilmesinde yardımcı olmuştur.

Sonuç olarak, çatışma ve protestoları inceleyen bu araştırma, kuvvet yasası dağılımlarının çatışmaların boyutunu da açıkladığını kaydetmektedir. Richardson (1948, 1960), Roberts ve Turcotte (1998), Cederman (2003), Friedman (2015) ve Clauset, Young ve Gleditsch (2007) gibi araştırmacılara göre çatışma boyutu, savaşın ve terörün nedenselliklerini içerebilir. Biggs (2016) ayaklanma ve gösteri katılımcılarının sayısı gibi daha az şiddetli çatışmaların diğer türlerinin kuvvet yasası dağılımlarıyla birlikte tanımlanabileceğini belirtmiştir. Bir çatışma boyutundaki kuvvet yasası dağılımı oldukça önemlidir çünkü bir kuvvet yasası dağılımı, nedenselliklerin dağılımının çoğunun veya bir çatışmadaki katılımcı sayısının az sayıda gözlem tarafından hesaplanabileceğini ima eder.

KAYNAKÇA

Kitaplar

ALLISON Paul D., "*Missing Data*", Thousand Oaks, CA: Sage University Paper No. 136, 2002.

ARNOLD Barry C., *"Pareto Distributions",* International Cooperative Publishing House, Fairland, Maryland USA, 1983.

AYTAÇ Mustafa, "Matematiksel İstatistik", 2. Baskı, Ezgi Kitabevi Yayınları, Şubat 1999.

CASELLA George – Robert CHRISTIAN, *"Monte Carlo Statistical Methods"*, Second Edition, Springer Verlag, 2004.

CAMEO, *"Conflict and Mediation Event Observations Event and Actor Codebook"*, Event Data Project Department of Political Science Pennsylvania State University Pond Laboratory University Park, PA 16802, Philip A. Schrodt (Project Director), March 2012.

ÇENGEL, Yunus A. - Michael A. BOLES, *"Mühendislik Yaklaşımıyla Termodinamik"*, çev. Taner Derbentli, Literatür Yayıncılık, Đstanbu, 1996.

GÜRSAKAL Necmi, "Büyük Veri", Genişletilmiş 2. Baskı, Dora, Bursa, ISBN:978-605-4798-803, 2014, syf. 26.

GÜRSAKAL, *"R İle Programlama"*, 1. Baskı, Bursa, Dora Yayınevi, 2014, syf. 5.

GÜRSAKAL Necmi, *"R İle Betimsel İstatistik",* 1. Baskı, Bursa, Dora Yayınevi, 2015, syf. 41

GÜRSKAL Necmi, *''R İle Betimsel İstatistik''*, 2. Baskı, Bursa, Dora Yayınevi, 2016, syf. 3.

GREENE Brian, *"Evrenin Dokusu"*, Tübitak Popüler Bilim Kitapları, 2. Baskı, çevirmen: Murat Alev, Num. 342, ISBN: 978-975-403-547-6, 2013.

LARRY Wasserman, *"All of Statistics: A Concise Course in Statistical Inference"*, (SpringerVerlag, Berlin, ISBN 978-0-387-21736-9, 2003.

MANYIKA James – Michael CHUI – Brad BROWN – Jacques BUGHIN – Richard DOBBS – Charles ROXBURGH – Angela Hung BYERS, *"Big Data: The next frontier for innovation, competttition, and productivity"*, Report McKinsey Global Institute, JUNE 2011.

MARTİN Trevor, *''The Undergraduate Guide to R- Abeginner's introduction to the R programming language''*, Princeton University, 2016.

MAYER–S. Viktor – Kenneth CUKIER, *"Big Data Arevolution That will Transform How We Live, Work, and Think"*, 2013, p.20.

MCAFEE Andrew – Erik BRYNJOLFSSON, *"Big Data: The Management Revolution."*, Harvard Business Review, 90, 10 October 2012, pp. 60-66.

NAUR Peter, *"Concise Survey of Computer Methods"*, 397 p., Student litteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974 ISBN/Petrocelli 0-88405-314-8, 1975.

OĞUZLAR Ayşe, "*Veri Madenciliğine Giriş"*, Ekin Kitabevi, Bursa, 2004.

OHLHORST Frank, *''Big data analytics: turning big data into big Money"*, New Jersey 2013, s.19-21.

SOARES Sunil, "*Big Data Governance - An Emerging Imperative*", MC Press Online, LLC, 1st edition edition, 2012.

ŞENTÜRK Aysan, *"Veri Madenciliği Kavram ve Teknikler"*, Ekin Kitabevi, Bursa, 2006.

TIGANI Jordan – Siddartha NAIDU, *"Google® BigQuery Analytics"*, Published by John Wiley & Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, Indiana Published simultaneously in Canada, ISBN: 978-1-118-82482-5, 2014.

TÜZÜNTÜRK Selim, *"Ağ Bilimi",* Dora Yayınevi, Bursa, 2012.

WALLGREN Anders – Britt WALLGREN, *"Register-based Statistics: Administrative Data for Statistical Purposes"*, Wiley Series in Survey Methodology, John Wiley & Sons, New York, 2007, ISBN: 978-0-470-02778-3-0.

Tezler

MAIER Markus, *"Towards a Big Data Reference Architecture"*, Eindhoven University of Technology, Department of Mathematics and Computer Science, Master's Thesis, 13th October 2013.

TÜZÜNTÜRK Selim, *''Firmalarda Organizasyonel Ağ Analizi Ve Bir Uygulama''*, Doktora Tezi, Bursa, 2012, s.75.

VACCARI Carlo, *"Big Data in Official Statistics"*, University of Camerino, Doktora of Philosophy in İnformation Science and Complex System-XXVI Cycle, School of Science and Technology, 2014.

Makaleler

AALST Wil M. P. van der, *"Data Scientist: The Engineer of the Future"*, In book: Enterprise Interoperability VI, Proceeding of the I-ES Conferences 7, DIO:10.1007/978-3-319-04948-9_2, 2014, pp. 13-26.

AKHTAR Nihat–Firoj PARWEJ–Yusuf PERWEJ, *"A Perusal of Big Data Classification and Hadoop Technology"*, International Transaction of Electrical and Computer Engineers System, Vol. 4, No. 1, 2017, 26-38.

BİGGS Michael, *"Strikes as forest fires: Chicago and Paris in the late 19th century"*, American Journal of Sociology Vol. 110, No. 6, 2005, pp. 1684-1714.

CEDERMAN Lars-Erik, *"Modeling the size of wars: From billiard balls to sandpiles"*, The American Political Science Review, Vol. 97, No. 1, (Feb., 2003), pp. 135-150.

CHEN Jinchuan – Yueguo CHEN – Xiaoyong DU – Cuiping LI – Jiaheng LU – Suyun ZHAO – Xuan ZHOU, *"Big data challenge: a data management perspective",* Frontiers of Computer Science, 7 (2):157–164, April 2013. doi: 10.1007/s11704-013-3903-7.

CHEN C.L. Philip – Chun-Yang ZHANG, ''Data-intensive applications, challenges, techniques and technologies: A survey on Big Data'', Information Sciences Volume 275, 10 August 2014, Pages 314-347.

CLAUSET Aaron – Maxwell YOUNG – Kristian Skrede GLEDITSCH, *"On the Frequency of Severe Terrorist Events"*, Journal of Conflict Resolution, Vol. 51, Number 1, 2007, pp. 58-87.

CLAUSET Aaron – Cosma Rohilla SHALIZI – M. E.J. NEWMAN, *"Power-law distributions in empirical data"*, SIAM Review, Vol. 51, No. 4, 2009, pp. 661-703.

CLAUSET Aaron – Kristian Skrede GLEDITSCH, *"The Developmental Dynamics of Terrorist Organizations"*, PLoS One, V.7(11), 21 November 2012.

CIOFFI-REVILLA Claudio, *"On the likely magnitude, extent, and duration of the Iraq-UN war"*, Journal of Conflict Resolution, Vol 35, Issue 3, 1991, ss. 387–411.

CORTES Corinna – Vladimir VAPNIK,"Support-vector networks", Machine Learning, Issue 3, Volume 20, 20 February 1995, pp 273-297.

CRAFT Ralph – Charles LEAKE, *"The Pareto principle in organizationel decision making"* Management Decision 40 (8), 2002, ss. 729-733.

DEIRIC Mccann, *"80-20 vision"* Dairy Industries International, 66 (9), 2001, syf. 25.

DERİNÖZ Cenk, "Google BigQuery Servisi İle Büyük Veri İşlemleri Ve Sorgu Sonuçlarının BIME İş Zekâsı Ürünü İle Görselleştirilip Android Tabanlı Mobil Cihazlar Üzerinden İzlenmesi", Data & Analytics, 22 Nisan 2014.

ERKAN Süleyman, *"2008 Rusya-Gürcistan Savaşı ve Uluslararası Toplum"*, Uluslararası İktisadi ve İdari İncelemeler Dergisi, 06.04.2016, s.41-64.

HALL Peter, *"On some Simple Estimates of an Exponent of Regular Variation"*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 44, No. 1, 1982, pp. 44-37.

HILL Bruce M., *"A Sımple General Approch to Inference about the Tail of a Distribution"*, The Annals of Statistics, Vol. 3, No. 5, 1975, page(s):1163-1174.

IŞIKLI Şevki, *''Büyük Veri, Epistemoloji ve Etik Tartışmalar''*, Online Academic Journal of Information Technology, Fall – Vol: 5 --/Num:17, 2014.

JAGADISH Hosagrahar V. – Johannes GEHRKE– Alexandros LABRINIDIS – Yannis PAPAKONSTANTINOU – Jignesh M. PATEL – Raghu RAMAKRISHNAN – Cyrus SHAHABI, *''Communications of the ACM''*, 2014, 57 (7): 86-94.

KALFA Ceren – Faruk ATAAY, *"Küresel Toplumsal Hareketler"*, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergis (16) 2008/2, ss. 127-149.

LACINA Bethany, "Explaining the severity of civil wars", Journal of Conflict Resolution Vol. 50, No. 2, 2006, ss. 276–289.

LAURA Peters, *"What is happining to the 80/20 rule?",* Semiconductor International, 25 (12): 17, 2002.

LEETRU Kalve – Felipe HOFFA, *"Analayzing the world's news: Exploring the GDELT Project through Google BigQuery"*, 24 November 2015.

MASON David M., *"Law of Large Numbers for Sum of Extreme Values"*, The Annals of Probability Vol. 10, No. 3, Aug. 1982, pp. 754-764.

MELNİK Sergey – Andrey GUBAREV – Jing Jing LONG – Geoffrey Romer – Shiva SHIVAKUMAR – Matt TOLTON – Theo VASSILAKIS, *"Dremel: Interactive Analysis of Web-Scale Datasets"*, Proceedings of the VLDB Endowment, Vol. 3, No. 1, Singapore, 2010.

MUNIRUZZAMAN A. N. M., *"On Measures of Location and Dispersion and Test of Hypotheses in a Paretto Population"*, Vol. 7, İssue:3, Page(s):115_123, Issue published: 1 July 1957, page(s):115-123.

ÖZEN Hayriye, *"Meydan Hareketleri ve Eski ve Yeni Toplumsal Hareketler"*, Mülkiye Dergisi, 2015, 39(2), 11-40.

RICHARDSON Lewis F., *"Variation of the frequency of fatal quarrels with magnitude"*, Journal of the American Statistical Association, Vol. 43, Issue, 244, 1948, pp. 523-546.

SEAL H. L., *"Maximum Likelihood Fitting of the Discrete Pareto Law"*, Journal of the Institute of Actuaries, Vol. 78, Issue 1, June 1952, pp. 115-121.

SHEKHAR Shashi – Viswanath GUNTURI – Michael R. EVANS – KwangSoo YANG, *"Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing"*, In Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, MobiDE '12, pages 1–6, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1442-8.

TATIKONDA L. U. – D. O'BRIEN – R. J. TATIKNDA, "*Succeeding with 80/20 rule. Management Accounting"* (February) 1999, ss. 40-44.

TUKEY John W., *"The Future of Data Analysis"*, The Annals of Mathematical Statistics, Vol. 33, No. 1, Mar. 1962, pp. 1-67.

WANG Lidong – Guanghui WANG – Cheryl Ann ALEXANDER, *"Big Data and Visualization: Methods, Challenges and Technology Progress"*, Digital Technologies, Vol. 1, No. 1, 27 June 2015, pp. 33-38.

Diğer Kaynaklar

AKÇAY Mustafa, *"Veri Bilimi"*, 2016, http://mustafaakca.com/veri-bilimi/, (02.10.2016).

ANDRIOTIS Nikos, *"The power-law distribution (and you)"*, 2014, https://www.efrontlearning.com/blog/2014/11/the-power-law-distribution.html, (20.06.2017).

Apache Spark – Tutorial, *"Apache Spark – Introduction"*, 2016, https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm, (01.02.2016).

BAYRAKCI Serkan, *"Büyük Veri Nedir?"*, 19 NİSAN 2015, https://serkanbayrakci.wordpress.com/tag/big-data/, (10.11.2016).

BETTİNO Larry, *"Transforming big data challenges into opportunities"*, *HealthData Management, 18* April 2012, http://www.healthdatamanagement.com/news/big-data-Starvest-IBM-Walmart-44338-1.html?zkPrintable=true, (10.12.2016).

BilgiUstam, *''Bilim ve Teknoloji Tarihi – Son Yüzyıldaki Gelişmeler''*, 2017, http://www.bilgiustam.com/bilim-ve-teknoloji-tarihi-son-yuzyildaki-gelismeler, (10.01.2017).

BHALLA Deepanshu, *"Companies Using R"*, Aralık 2016, http://www.listendata.com/2016/12/companies-using-r.html, (13.12.2017).

BROWN Brad – Michael CHUI – James MANYIKA, *"Are you ready for the era of 'Big Data'?"*, Article McKinsey Quarterly, October 2011: 24-35, http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data, (10.12.2016).

Capital.com, *"Gelecek Trendler: 80-20 Kuralı Nasıl Şirket Kurtardı?"*, http://www.capital.com.tr/gelecek-trendler/8020-kurali--nasil-sirket-kurtardi-haberdetay, (20.06.2017).

Camille Mendler. M2M and big data. Website, *"A report the Economist: Intelligence Unit"*, 2013.

CLEVELAND William *S., "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics"*, Bell Labs, 2001, http://www.stat.purdue.edu/~wsc/papers/datascience.pdf. , (10.09.2016).

CONWAY Drew, *"The Data Science Venn Diagram"*, 30 September 2010, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, (10.11.2016).

CORTES Corinna – Vladimir VAPNIK, *"Support-vector networks",* Machine Learning 20(3), September 1995, www.springerlink.com/content/k238jx04hm87j80g/, (10.12.2016).

Databricks, "What is Apache Spark™?", 2016, https://databricks.com/spark/about/, (01.02.2016).

DATAFLOQ, *"How Unlimited Computing Power, Swarms of Sensors and Algorithms Will Rock our World"*, 20 JUNE, 2016, https://datafloq.com/read/unlimited-computing-swarm-sensors-algorithms-world/2138, (10.01.2017).

DATAFLOQ, *"Public Data Marketplaces and Initiatives"*, 2017, https://datafloq.com/public-data/, (10.06.2017).

DATAFLOQ, *"How Cloud Computing Affects Individuals and Organizations"*, 30 DECEMBER 2016,https://datafloq.com/read/cloud-computing-affects-individuals-organizations/2559, (10.05.2017).

DAVENPORT Thomas H. – D.J. PATIL, *"Data Scientist: The Sexiest Job of the 21st Century"*, October 2012, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, (10.05.2016).

DEMİR Timur, *"Bulut Bilişim (Cloud Computing) Nedir?"*, 2016, http://www.timurdemir.com.tr/bulut-bilisim-cloud-computing-nedir, (07.01.2016).

DEVLIN Barry – Shawn ROGERS – John MYERS, "*Big DataComes of Age"*, EMA and 9sight Consulting Research Report, 11.01.2012, http://www.enterprisemanagement.com/research/asset.php/2409/Big-Data-Comes-of-Age, (17.03.2017).

DeZyre, *"Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for Hadoop Users"*, 25 November 2014, https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84, (10.02.2017).

DeZyre, *"Data Science Programming: Python vs R"*, 20 JUNE 2015, https://www.dezyre.com/article/data-science-programming-python-vs-r/128, (20.01.2017).

DeZyre, *"Hadoop Components and Architecture: Big Data and Hadoop Training"*, November 24, 2016, https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114, (20.01.2017).

EL Siraceddin, *"Nedir Bu Web 2.0 Teknolojisi?",* 21 Haziran 2008, http://sanalkurs.net/nedir-bu-web-2-0-teknolojisi-2212.html, (02.09.2016).

EticaretMag, *"E-Ticarette Veri Bilimcilerin Önemi Artıyor"*, 06 Ağustos 2014, http://eticaretmag.com/e-ticarette-veri-bilimcilerin-onemi-artiyor, (29.06.2016).

FENG Jeff – Marc LOBREE – Tino TERESHKO – Mike GRABOSKI, *"Google BigQuery & Tableau: Best Practices"*, 2017, http://www.tableau.com/sites/default/files/media/whitepaper_googlebigquerytableaubestpractices_eng.pdf/, (06.01.2017).

GANTZ John – David REINSEL, *"The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.",* Study report, IDC,

GDELT, *"DATA FORMAT CODEBOOK"*, V 1.03, 8/25/2013 http://data.gdeltproject.org/documentation/GDELT-DataFormatCodebook.pdf, (12.10.2016).

GE TÜRKİYE BLOG, *"Endüstriyel İnternet, Büyük Veri ve Operasyon Optimizasyonu"*, 24 Kasım 2015, https://geturkiyeblog.com/endustriyel-internet-buyuk-veri-operasyon-optimizasyonu/, (19.09.2016).

GE TÜRKİYE BLOG, *"Büyük Veri Trafiği Nasıl Alt Eder?"*, 13 Mayıs 2016, https://geturkiyeblog.com/buyuk-veri-trafigi-nasil-alt-eder, (29.07.2016).

GHEMAWAT Sanjay – Howard GOBIOFF – Shun-Tak LEUNG, *"The Google file system",* 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003.

GIN Jasper, *"Distilling Event Data from News Articles"*, 18 JANUARY 2015. (https://www.jasperginn.nl/distilling-event-data-from-news-articles/#fnref, 06/09/2017).

Google BigQuery, *"Developers Google BigQuery"*, October 2012, http://developers.google.com/bigquery/, (05.02.2017).

Google BigQuery Export, "BigQuery veri yönetimi", 2017, https://support.google.com/analytics/answer/3416092?hl=tr&ref_topic=3416089, (12.05.2017).

Google Cloud Platform, *"Using Tables",* 2017, https://cloud.google.com/bigquery/docs/tables, (25.05.2017).

Google Cloud Platform, *"Slots"*, 2017, https://cloud.google.com/bigquery/docs/slots, (29.08.2017).

Google Cloud Platform, *"GDELT HathiTrust and Internet Archive Book Data"*, 2017, https://cloud.google.com/bigquery/public-data/gdelt-books, (10.08.2017).

GUTIERREZ Daniel, *"Evolution of the Data Scientist: How Number Crunching Became the Number One Job in America"*, March24, 2014, https://insidebigdata.com/2016/03/24/evolution-of-the-data-scientist-how-number-crunching-became-the number-one-job-in-america/ (28.12.2016).

HABERMAG, *"Nano Boyuttta Veri Depolamak"*, 2016, http://habermag.net/nano-boyutlara-veri-depolamak/ (21/12/2016).

HINCHCLIFFE Dion, *"Twenty-two power laws of the emerging social economy"*, 5 October 2009, http://www.zdnet.com/article/twenty-two-power-laws-of-the-emerging-social-economy/, (20.06.2017).

HEDLUND Brad, *"Understanding Hadoop Cluster and the Network"*, September 10, 2011, http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/, (10.03.2017).IDC's Digital Universe Study, sponsored by EMC, December 2012.

HOWE Jeff Howe, *"The Rise of Crowdsourcing",* Wired, Issue 14, 06.01.2006, https://www.wired.com/2006/06/crowds/, (02.09.2016).

INVESTOPEDIA, *"80 – 20 Rule",* 2017, http://www.investopedia.com/terms/1/80-20-rule.asp, (20.06.2017).

İLTER Hakan – Hakan SARIBIYIK – Emre YAZICI – Erkan ÜLGEY – Yasemin Kaya – İmran KOCABIYIK – Emre Ekış – Hüseyin Babal – Harun YARDIMCI – Cevat UZUN –Ahmet ARSLAN – Ayhan DEMİRCİ – Erdem EĞAOĞLU – Ümit ÜNAL – Arif BALIK, *"NoSQL", 2012. http://devveri.com/nosql-nedir (25.09.2016).*

KDnuggets, *"Python overtakes R, becames the leader in Data Science, Machine Learning platforms"*, Aug 2017, http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html, (09.09.2017).

KOTTKE Jason, *"Weblogs and power laws"*, 9 February 2003, http://kottke.org/03/02/weblogs-and-power-laws, (20.06.2017).

MastersInDataScience, *"using Tableau for Data Science"*, 12 sep 2014, http://www.mastersindatascience.org/data-scientist-skills/tableau/, (04.02.2017).

MARR Bernard, *"Big Data Terminology: 16 Key Concepts Everyone Should Understand (Part II)"*, 17 May 2017, http://data-informed.com/big-data-terminology-16-key-concepts-everyone-should-understand-part-ii/, (10.06.2017).

MILLS Steve –Steve LUCAS – Leo IRAKLIOTIS – Michael RAPPA – Teresa CARLSON – Bill PERLOWITZ, "*Demystifying Big Data–A Practical Guide to Transforming The Business of Government"*, prepared by TechAmerica Foundation's Federal Big Data Commission, 2016, https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095pdf, (10.12.2016).

MOLTKE helmuth von, *"İsrail'in Yaptığı 10 Operasyon"*, 18 Temmuz 2014, https://onedio.com/haber/israil-in-gazze-ye-yaptigi-10-operasyon-337586, (27.08.2017).

NARİN Bilge, *"Big Data"*, 24 Mart 2015, http://es.slideshare.net/BilgeNarin1/big-data-24-mart-2015, (29.09.2016).

NTV Haber, *"Suriye iç savaşında 6 yılda neler yaşandı"*, 15 Mart Çarşamba 2017, http://www.ntv.com.tr/dunya/suriye-ic-savasinda-6-yilda-neler-yasandi,Ik0VGc0sPUqbvD, (27.08.2017).

ÖZSOY Canan M., *"Endüstrinin Geleceği ve Endüstriyel İnternet Devrimi"*, 25 Kaasım 2014, http://geturkiyeblog.com/endustrinin-gelecegi-ve-endustriyel-internet-devrimi, (19.09.2016).

ÖZŞİMŞİR Şafak, *"2003 ABD-Irak Savaşı ve Nedenleri"*, 30 Ocak 2011. http://www.tuicakademi.org/2003-abd-irak-savasi-ve-nedenleri/, (27.08.2017).

PENCHİKALA Srini, *"Big Data Processing with Apache Spark – Part 1: Introduction"*, Jan 30, 2015, https://www.infoq.com/articles/apache-spark-introduction, (22.10.2017).

Popular Science Türkiye, *''Karanlık Web'i Aydınlatan Adam''*, 03.11.2016, http://www.pressreader.com/turkey/popular-scence-turkey, (20.11.2016).

Ratheesh's- Tech Blog, *"Data Science"*, 2016, http://rathishnair.com/techblog/data-science-machine-learning/, (10.11.2016).

ROBB Drew, *"Semi-Structured Data"*, July 3, 2017, https://www.datamation.com/big-data/semi-structured-data.html, (20.11.2017).

ROUSE Margaret, "business intelligence (BI)", August 2017, http://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI, (20.11.2017)

PRESS Gil, *"A Very Short History Of Data Science"*, 9 May 2013, https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#14fab5ea65a1, (10.10.2016).

SARKAR Deepayan, *"An Introduction to R"*, May 2012, http://www.isid.ac.in/~deepayan/R-tutorials/labs/01_introduction_lab.pdf, (10.01.2017).

SOSYAL BİLİMLER, *"Sosyal Hareketler"*, 2017, http://www.enfal.de/sosyalbilimler/s/055.htm, (30.08.2017).

STARAK Yaro, *"What Is The 80/20 Rule And Why It Will Change Your Life"*, 2017, https://www.entrepreneurs-journey.com/397/80-20-rule-pareto-principle/, (20.06.2017).

STROUD Forrest, *"NoSQL Database"*, 2017, Nosql Database: http://www.webopedia.com/TERM/N/nosql_databese.html, (10.03.2017).

SZKOLAR Dorotea, *"Data Mining in Obama's 2012 Victory"*, 24 January 2013, http://infospace.ischool.syr.edu/2013/01/24/data-mining-in-obamas-2012-victory, (29.07.2016).

The GDELT Project, *"The GDELT Story"*, 2017, http://gdeltproject.org/about.html#intro, (10.08.2017).

The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html, (15.08.2017).

The GDELT Project, *"Google BigQuery"*, 2017, http://www.gdeltproject.org/data.html#googlebigquery, (20.08.2017).

The GDELT Project, *"What Turkey's Attempted Coup Looked Like Through GDELT's Eyes"*, 16 July 2016, http://blog.gdeltproject.org/turkeys-attempted-coup-looked-like-gdelts-eyes/, (20.08.2017).

The GDELT Project, *"GDELT Analysis Service"*, 2017, http://gdeltproject.org/data.html#gdeltanalysisservice, (25.08.2017).

The GDELT Project, *"GDELT 2.0: Our Global World in Realtime"*, 19 February 2015. http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/, (25.08.2017).

Tuturialspoint, *"Apache Kafka Tutorials"*, 2017, https://www.tutorialspoint.com/apache_kafka/index.htm, (12.01.2017).

TÜFENKÇİ Zeynep, *'' In Defense of Nate Silver, Election Pollsters, and Statistical Predictions''*, 11.02.2012, https://www.wired.com/2012/11/why-predictions-and-statistical-models-are-necessary-and-good-for-democracy, (23/09/2016).

Wikipedia, *"Bilişim Çağı",* 2016, https://tr.wikipedia.org/wiki/Bilinote-1, (10.05.2016).

Wikipedia, *"Extensible Metadata Platform"*, 2016, https://en.wikipedia.org/wiki/Exif (20.09.2016).

ZORLU Emre, *"Big Data"*, 12 TEMMUZ 2012, https://emrezorlu.com/2014/07/12/big-data/, (25.12.2016).

EKLER

EK 1 1979-2017 Yılları Arasında Aylık Bazda Aktör Ülkeler Arasındaki Çatışma Sayıları (SQL ÇIKTILARI)

MonthYear	Actor1Name	Actor2Name	Count	MonthYear	Actor1Name	Actor2Name	Count
197901	CHINA	UNITED STATES	200	199804	ISRAEL	PALESTINIAN	1028
197902	CHINA	VIETNAM	535	199805	ISRAEL	UNITED STATES	1207
197903	CHINA	VIETNAM	465	199806	CHINA	UNITED STATES	1446
197904	CHINA	VIETNAM	221	199807	ISRAEL	PALESTINIAN	877
197905	CHINA	VIETNAM	307	199808	ISRAEL	PALESTINIAN	635
197906	RUSSIA	UNITED STATES	242	199809	ISRAEL	PALESTINIAN	1176
197907	CHINA	VIETNAM	256	199810	SYRIA	TURKEY	1879
197908	CHINA	VIETNAM	188	199811	ISRAEL	PALESTINIAN	1478
197909	EGYPT	ISRAEL	194	199812	ISRAEL	PALESTINIAN	1620
197910	JAPAN	RUSSIA	142	199901	IRAQ	UNITED STATES	653
197911	IRAN	UNITED STATES	376	199902	ERITREA	ETHIOPIA	819
197912	IRAN	UNITED STATES	370	199903	RUSSIA	UNITED STATES	859
198001	AFGHANISTAN	RUSSIA	847	199904	CHINA	UNITED STATES	1115
198002	AFGHANISTAN	RUSSIA	481	199905	CHINESE	NATO	1082
198003	AFGHANISTAN	RUSSIA	403	199906	RUSSIA	UNITED STATES	902
198004	IRAN	UNITED STATES	359	199907	ISRAEL	PALESTINIAN	917
198005	AFGHANISTAN	RUSSIA	257	199908	ISRAEL	PALESTINIAN	1096
198006	AFGHANISTAN	RUSSIA	389	199909	EAST TIMOR	INDONESIA	1566
198007	IRAN	UNITED STATES	211	199910	ISRAEL	PALESTINIAN	1048
198008	EGYPT	ISRAEL	228	199911	ISRAEL	PALESTINIAN	1113
198009	IRAN	IRAQ	773	199912	ISRAEL	SYRIA	2119
198010	IRAN	IRAQ	1193	200001	ISRAEL	SYRIA	2533
198011	IRAN	IRAQ	707	200002	ISRAEL	LEBANON	1191
198012	IRAN	IRAQ	417	200003	ISRAEL	SYRIA	1477
198101	IRAN	UNITED STATES	640	200004	ISRAEL	LEBANON	1224
198102	IRAN	IRAQ	338	200005	ISRAEL	LEBANON	1813
198103	IRAN	IRAQ	367	200006	ERITREA	ETHIOPIA	1133
198104	POLAND	RUSSIA	345	200007	ISRAEL	PALESTINIAN	1108
198105	ISRAEL	SYRIA	391	200008	ISRAEL	PALESTINIAN	1086

198106	POLAND	RUSSIA	286	200009	ISRAEL	PALESTINIAN	1194
198107	ISRAEL	LEBANON	291	200010	ISRAELI	PALESTINIAN	3821
198108	FRANCE	IRAN	326	200011	ISRAELI	PALESTINIAN	2547
198109	RUSSIA	UNITED STATES	329	200012	ISRAELI	PALESTINIAN	1848
198110	EGYPT	ISRAEL	404	200101	ISRAELI	PALESTINIAN	2271
198111	RUSSIA	UNITED STATES	378	200102	ISRAELI	PALESTINIAN	1461
198112	RUSSIA	UNITED STATES	454	200103	ISRAELI	PALESTINIAN	1689
198201	EGYPT	ISRAEL	347	200104	ISRAELI	PALESTINIAN	2712
198202	EGYPT	ISRAEL	231	200105	ISRAELI	PALESTINIAN	2270
198203	NICARAGUA	UNITED STATES	271	200106	ISRAELI	PALESTINIAN	2026
198204	ARGENTINE	BRITISH	441	200107	ISRAELI	PALESTINIAN	2094
198205	ARGENTINE	BRITISH	1220	200108	ISRAELI	PALESTINIAN	2305
198206	ISRAEL	LEBANON	745	200109	ISRAELI	PALESTINIAN	1687
198207	ISRAEL	LEBANON	498	200110	AFGHANISTAN	UNITED STATES	2067
198208	ISRAEL	LEBANON	585	200111	ISRAELI	PALESTINIAN	1253
198209	ISRAEL	LEBANON	633	200112	ISRAELI	PALESTINIAN	1913
198210	ISRAEL	LEBANON	440	200201	ISRAELI	PALESTINIAN	1851
198211	RUSSIA	UNITED STATES	392	200202	ISRAELI	PALESTINIAN	2139
198212	ISRAEL	LEBANON	703	200203	ISRAELI	PALESTINIAN	2894
198301	ISRAEL	LEBANON	677	200204	ISRAELI	PALESTINIAN	3030
198302	ISRAEL	LEBANON	486	200205	ISRAELI	PALESTINIAN	1483
198303	ISRAEL	LEBANON	449	200206	ISRAELI	PALESTINIAN	1459
198304	ISRAEL	LEBANON	557	200207	ISRAELI	PALESTINIAN	1667
198305	ISRAEL	LEBANON	943	200208	ISRAELI	PALESTINIAN	1559
198306	RUSSIA	UNITED STATES	436	200209	ISRAELI	PALESTINIAN	1401
198307	RUSSIA	UNITED STATES	431	200210	ISRAELI	PALESTINIAN	1014
198308	CHAD	LIBYA	293	200211	ISRAELI	PALESTINIAN	1037
198309	RUSSIA	UNITED STATES	1012	200212	IRAQ	UNITED STATES	1100
198310	RUSSIA	UNITED STATES	446	200301	IRAQ	UNITED STATES	1879
198311	RUSSIA	UNITED STATES	649	200302	IRAQ	UNITED STATES	2264
198312	RUSSIA	UNITED STATES	308	200303	IRAQ	UNITED STATES	3804
198401	RUSSIA	UNITED STATES	428	200304	IRAQ	UNITED STATES	3583
198402	ISRAEL	LEBANON	394	200305	IRAQ	UNITED STATES	1698
198403	IRAN	IRAQ	443	200306	ISRAELI	PALESTINIAN	1748

198404	CHINA	UNITED STATES	363	200307	ISRAELI	PALESTINIAN	1511
198405	IRAN	IRAQ	278	200308	ISRAEL	PALESTINIAN	1286
198405	RUSSIA	UNITED STATES	278	200309	IRAQ	UNITED STATES	1643
198406	IRAN	IRAQ	654	200310	IRAQ	UNITED STATES	1401
198407	RUSSIA	UNITED STATES	519	200311	IRAQ	UNITED STATES	1306
198408	RUSSIA	UNITED STATES	384	200312	ISRAELI	PALESTINIAN	1330
198409	RUSSIA	UNITED STATES	450	200401	IRAQ	UNITED STATES	1096
198410	RUSSIA	UNITED STATES	312	200402	RUSSIA	UNITED STATES	1047
198411	ISRAEL	LEBANON	424	200403	ISRAELI	PALESTINIAN	1447
198412	ISRAEL	LEBANON	273	200404	IRAQ	UNITED STATES	1837
198501	RUSSIA	UNITED STATES	503	200405	IRAQ	UNITED STATES	1994
198502	RUSSIA	UNITED STATES	358	200406	IRAQ	UNITED STATES	1377
198503	RUSSIA	UNITED STATES	691	200407	IRAQ	UNITED STATES	871
198504	RUSSIA	UNITED STATES	585	200408	ISRAELI	PALESTINIAN	799
198505	RUSSIA	UNITED STATES	423	200409	IRAQ	UNITED STATES	1055
198506	RUSSIA	UNITED STATES	506	200410	IRAQ	UNITED STATES	1147
198507	RUSSIA	UNITED STATES	437	200411	ISRAEL	PALESTINIAN	1115
198508	RUSSIA	UNITED STATES	477	200412	ISRAELI	PALESTINIAN	887
198509	RUSSIA	UNITED STATES	377	200501	ISRAEL	PALESTINIAN	1261
198510	ITALY	UNITED STATES	551	200502	ISRAEL	PALESTINIAN	1037
198511	RUSSIA	UNITED STATES	991	200503	LEBANON	SYRIA	1573
198512	RUSSIA	UNITED STATES	521	200504	CHINA	JAPAN	1181
198601	LIBYA	UNITED STATES	589	200505	CHINA	JAPAN	719
198602	PHILIPPINE	UNITED STATES	510	200506	ISRAELI	PALESTINIAN	668
198603	LIBYA	UNITED STATES	414	200507	ISRAELI	PALESTINIAN	667
198604	LIBYA	UNITED STATES	1165	200508	GAZA	ISRAEL	731
198605	LIBYA	UNITED STATES	335	200509	ISRAEL	PALESTINIAN	678
198606	RUSSIA	UNITED STATES	254	200510	ISRAELI	PALESTINIAN	704
198607	RUSSIA	UNITED STATES	319	200511	IRAQ	UNITED STATES	630
198608	RUSSIA	UNITED STATES	472	200512	IRAQ	UNITED STATES	612
198609	RUSSIA	UNITED STATES	935	200601	RUSSIA	UKRAINE	811
198610	RUSSIA	UNITED STATES	908	200602	IRAN	RUSSIA	1018
198611	IRAN	UNITED STATES	493	200603	IRAQ	UNITED STATES	1936
198612	IRAN	IRAQ	364	200604	IRAQ	UNITED STATES	1273

198701	IRAN	IRAQ	691	200605	IRAN	UNITED STATES	1358
198702	IRAN	IRAQ	371	200606	IRAQ	UNITED STATES	2441
198703	ISRAEL	UNITED STATES	716	200607	ISRAEL	LEBANON	5580
198704	RUSSIA	UNITED STATES	474	200608	ISRAEL	LEBANON	5943
198705	IRAN	IRAQ	300	200609	ISRAEL	LEBANON	1891
198706	IRAN	IRAQ	274	200610	IRAQ	UNITED STATES	1729
198707	IRAN	IRAQ	561	200611	IRAQ	UNITED STATES	3403
198708	IRAN	IRAQ	498	200612	IRAQ	UNITED STATES	3486
198709	IRAN	IRAQ	1024	200701	IRAQ	UNITED STATES	4260
198710	IRAN	IRAQ	515	200702	IRAQ	UNITED STATES	3235
198711	IRAN	IRAQ	547	200703	IRAQ	UNITED STATES	3666
198712	RUSSIA	UNITED STATES	595	200704	IRAQ	UNITED STATES	4725
198801	ISRAELI	PALESTINIAN	376	200705	IRAQ	UNITED STATES	6066
198802	PANAMA	UNITED STATES	435	200706	IRAQ	UNITED STATES	3435
198803	IRAN	IRAQ	1032	200707	IRAQ	UNITED STATES	4528
198804	IRAN	IRAQ	617	200708	IRAQ	UNITED STATES	3409
198805	RUSSIA	UNITED STATES	894	200709	IRAQ	UNITED STATES	4989
198806	RUSSIA	UNITED STATES	383	200710	IRAQ	UNITED STATES	3810
198807	IRAN	IRAQ	1176	200711	IRAQ	UNITED STATES	2960
198808	IRAN	IRAQ	1407	200712	IRAQ	UNITED STATES	2273
198809	IRAN	IRAQ	767	200801	IRAQ	UNITED STATES	3296
198810	IRAN	IRAQ	537	200802	IRAQ	UNITED STATES	3230
198811	IRAN	IRAQ	504	200803	IRAQ	UNITED STATES	4258
198812	ISRAEL	UNITED STATES	504	200804	IRAQ	UNITED STATES	3873
198901	AFGHANISTAN	RUSSIA	573	200805	IRAQ	UNITED STATES	3227
198902	AFGHANISTAN	RUSSIA	650	200806	IRAQ	UNITED STATES	3355
198903	ISRAELI	PALESTINIAN	373	200807	IRAQ	UNITED STATES	4311
198904	AFRICA	NAMIBIA	252	200808	RUSSIA	UNITED STATES	9187
198905	PANAMA	UNITED STATES	376	200809	RUSSIA	UNITED STATES	4038
198906	RUSSIA	UNITED STATES	286	200810	IRAQ	UNITED STATES	3120
198907	RUSSIA	UNITED STATES	295	200811	IRAQ	UNITED STATES	3152
198908	ISRAEL	LEBANON	367	200812	GAZA	ISRAEL	4521
198909	ISRAELI	PALESTINIAN	341	200901	GAZA	ISRAEL	12611
198910	ISRAEL	PALESTINIAN	295	200902	GAZA	ISRAEL	3260

198911	IRAN	IRAQ	243	200903	IRAN	UNITED STATES	2636
198912	PANAMA	UNITED STATES	725	200904	IRAN	UNITED STATES	3353
199001	PANAMA	UNITED STATES	526	200905	RUSSIA	UNITED STATES	2974
199002	RUSSIA	UNITED STATES	278	200906	ISRAEL	PALESTINIAN	2795
199003	LITHUANIA	RUSSIA	435	200907	RUSSIA	UNITED STATES	3189
199004	LITHUANIA	RUSSIA	509	200908	AFGHANISTAN	UNITED STATES	4110
199005	ISRAELI	PALESTINIAN	352	200909	CHINA	UNITED STATES	4367
199006	RUSSIA	UNITED STATES	274	200910	AFGHANISTAN	UNITED STATES	5726
199007	IRAQ	KUWAIT	523	200911	AFGHANISTAN	UNITED STATES	3903
199008	IRAQ	KUWAIT	1619	200912	AFGHANISTAN	UNITED STATES	5286
199009	IRAQ	KUWAIT	703	201001	HAITI	UNITED STATES	5155
199010	IRAQ	KUWAIT	539	201002	CHINA	UNITED STATES	4724
199011	IRAQ	KUWAIT	489	201003	ISRAEL	PALESTINIAN	1727
199012	IRAQ	KUWAIT	577	201004	RUSSIA	UNITED STATES	2209
199101	IRAQ	UNITED STATES	1007	201005	PAKISTAN	UNITED STATES	1965
199102	IRAQ	KUWAIT	1494	201006	GAZA	ISRAEL	2193
199103	IRAQ	KUWAIT	555	201007	RUSSIA	UNITED STATES	2990
199104	IRAQ	UNITED STATES	525	201008	IRAQ	UNITED STATES	5247
199105	ISRAEL	RUSSIA	366	201009	CHINA	JAPAN	6222
199106	RUSSIA	UNITED STATES	326	201010	CHINA	JAPAN	3839
199107	RUSSIA	UNITED STATES	425	201011	CHINA	UNITED STATES	4076
199108	ISRAEL	PALESTINIAN	381	201012	CHINA	UNITED STATES	4052
199109	RUSSIA	UNITED STATES	398	201101	CHINA	UNITED STATES	5634
199110	ISRAEL	PALESTINIAN	673	201102	EGYPT	UNITED STATES	5565
199111	ISRAEL	PALESTINIAN	525	201103	LIBYA	UNITED STATES	6464
199112	RUSSIA	UKRAINE	716	201104	LIBYA	UNITED STATES	3189
199201	ISRAELI	PALESTINIAN	544	201105	PAKISTAN	UNITED STATES	8824
199202	ARMENIA	AZERBAIJAN	213	201106	AFGHANISTAN	UNITED STATES	4278
199203	ARMENIA	AZERBAIJAN	329	201107	CHINA	UNITED STATES	3950
199204	RUSSIA	UKRAINE	469	201107	AFGHANISTAN	UNITED STATES	3950
199205	ARMENIA	AZERBAIJAN	388	201108	AFGHANISTAN	UNITED STATES	3807
199206	RUSSIA	UNITED STATES	366	201109	PAKISTAN	UNITED STATES	4496
199207	ISRAEL	PALESTINIAN	208	201110	ISRAEL	PALESTINIAN	3460
199208	ISRAEL	PALESTINIAN	254	201111	CHINA	UNITED STATES	3483

199209	ISRAEL	SYRIA	346	201112	IRAQ	UNITED STATES	5112
199210	ISRAEL	PALESTINIAN	189	201201	IRAN	UNITED STATES	5010
199211	ISRAEL	PALESTINIAN	237	201202	IRAN	ISRAEL	5660
199212	ISRAEL	PALESTINIAN	523	201203	AFGHANISTAN	UNITED STATES	6417
199301	ISRAEL	PALESTINIAN	425	201204	AFGHANISTAN	UNITED STATES	3929
199302	ISRAEL	PALESTINIAN	363	201205	CHINA	UNITED STATES	6388
199303	ISRAEL	PALESTINIAN	363	201206	CHINA	UNITED STATES	3931
199304	ISRAEL	PALESTINIAN	494	201207	RUSSIA	SYRIA	3560
199305	ISRAEL	PALESTINIAN	406	201208	IRAN	SYRIA	3407
199306	ISRAEL	PALESTINIAN	347	201209	CHINA	APAN	5636
199307	ISRAEL	LEBANON	595	201210	SYRIA	TURKEY	5041
199308	ISRAEL	PALESTINIAN	440	201211	GAZA	ISRAEL	7789
199309	ISRAEL	PALESTINIAN	674	201212	RUSSIA	UNITED STATES	3458
199310	ISRAEL	PALESTINIAN	518	201301	FRANCE	MALI	3224
199311	ISRAEL	JORDAN	563	201302	RUSSIA	UNITED STATES	3335
199312	ISRAELI	PALESTINIAN	572	201303	ISRAEL	OBAMA	4018
199401	ISRAEL	SYRIA	557	201304	NORTH KOREA	UNITED STATES	2123
199402	RUSSIA	UNITED STATES	544	201305	ISRAEL	UNITED STATES	2067
199403	ISRAELI	PALESTINIAN	600	201306	CHINA	UNITED STATES	5287
199404	ISRAEL	PALESTINIAN	460	201307	RUSSIA	UNITED STATES	8994
199405	ISRAEL	PALESTINIAN	502	201308	RUSSIA	UNITED STATES	8788
199406	NORTH KOREA	UNITED STATES	397	201309	SYRIA	UNITED STATES	14407
199407	ISRAEL	JORDAN	1649	201310	IRAN	UNITED STATES	6369
199408	ISRAEL	JORDAN	767	201311	IRAN	UNITED STATES	9855
199409	HAITI	UNITED STATES	653	201312	CHINA	UNITED STATES	5682
199410	ISRAEL	JORDAN	1719	201401	ISRAEL	PALESTINIAN	5010
199411	ISRAEL	JORDAN	736	201402	RUSSIA	UKRAINE	5291
199412	ISRAEL	SYRIA	694	201403	RUSSIA	UKRAINE	23135
199501	ISRAEL	PALESTINIAN	416	201404	RUSSIA	UKRAINE	17925
199502	ECUADOR	PERU	992	201405	RUSSIA	UKRAINE	12083
199503	ISRAEL	SYRIA	739	201406	IRAQ	UNITED STATES	10390
199504	ISRAEL	SYRIA	446	201407	GAZA	ISRAEL	16774
199505	ISRAEL	SYRIA	425	201408	RUSSIA	UKRAINE	13600
199506	ISRAEL	SYRIA	847	201409	RUSSIA	UKRAINE	10074

199507	CHINA	UNITED STATES	531	201410	RUSSIA	UKRAINE	6506
199508	CHINA	UNITED STATES	628	201411	CHINA	UNITED STATES	7148
199509	ISRAEL	PALESTINIAN	436	201412	CUBA	UNITED STATES	11808
199510	ISRAEL	PALESTINIAN	261	201501	CUBA	UNITED STATES	7177
199511	ISRAEL	SYRIA	392	201502	RUSSIA	UKRAINE	11092
199512	ISRAEL	SYRIA	876	201503	IRAN	UNITED STATES	8335
199601	ISRAEL	SYRIA	847	201504	IRAN	UNITED STATES	7019
199602	ISRAEL	PALESTINIAN	493	201505	CHINA	UNITED STATES	5216
199603	ISRAEL	PALESTINIAN	868	201506	CHINA	UNITED STATES	5505
199604	ISRAEL	LEBANON	1043	201507	IRAN	UNITED STATES	10124
199605	ISRAEL	PALESTINIAN	500	201508	IRAN	UNITED STATES	7269
199606	ISRAEL	PALESTINIAN	671	201509	CHINA	UNITED STATES	8542
199607	ISRAEL	PALESTINIAN	815	201510	RUSSIA	SYRIA	13039
199608	ISRAEL	SYRIA	919	201511	RUSSIA	TURKEY	8732
199609	ISRAEL	PALESTINIAN	1482	201512	RUSSIA	TURKEY	8718
199610	ISRAEL	PALESTINIAN	1593	201601	IRAN	SAUDI ARABIA	12807
199611	ISRAEL	PALESTINIAN	1021	201602	RUSSIA	SYRIA	10825
199612	ISRAEL	PALESTINIAN	1209	201603	RUSSIA	SYRIA	6884
199701	ISRAEL	PALESTINIAN	1136	201604	ARMENIA	AZERBAIJAN	9663
199702	ISRAEL	PALESTINIAN	708	201605	CHINA	UNITED STATES	5909
199703	ISRAEL	PALESTINIAN	1444	201606	MEXICO	UNITED STATES	7880
199704	ISRAELI	PALESTINIAN	881	201607	CHINA	PHILIPPINE	9209
199705	NATO	RUSSIA	949	201608	RUSSIA	UNITED STATES	6278
199706	ISRAELI	PALESTINIAN	917	201609	RUSSIA	UNITED STATES	9537
199707	ISRAELI	PALESTINIAN	909	201610	RUSSIA	UNITED STATES	10417
199708	ISRAEL	PALESTINIAN	1348	201611	RUSSIA	UNITED STATES	7478
199709	ISRAEL	PALESTINIAN	1050	201612	RUSSIA	UNITED STATES	10726
199710	CHINA	UNITED STATES	1093	201701	RUSSIA	UNITED STATES	12995
199711	IRAQ	UNITED STATES	1134	201702	RUSSIA	UNITED STATES	14066
199712	ISRAEL	PALESTINIAN	855	201703	RUSSIA	UNITED STATES	14864
199801	ISRAEL	PALESTINIAN	978	201704	RUSSIA	SYRIA	11362
199802	IRAQ	UNITED STATES	1716	201705	RUSSIA	UNITED STATES	16434
199803	ISRAEL	PALESTINIAN	1268	201706	RUSSIA	UNITED STATES	13527

EK 2 1979-2017 Yılları Arasında Türkiye'deki Protesto Yoğunlukları (SQL ÇIKTILARI)

MonthYear	Percent	MonthYear	Percent	MonthYear	Percent	MonthYear	Percent	MonthYear	Percent
197902	0,017	198803	0,003	199512	0,007	200306	0,001	201012	0,008
197903	0,014	198804	0,004	199601	0,032	200307	0,006	201101	0,007
197904	0,023	198805	0,004	199602	0,002	200308	0,014	201102	0,017
197906	0,002	198806	0,008	199603	0,029	200309	0,001	201103	0,012
197908	0,014	198807	0,005	199604	0,001	200310	0,023	201104	0,016
197909	0,013	198808	0	199605	0,001	200311	0,014	201105	0,014
197910	0,005	198811	0,048	199606	0,043	200312	0,009	201106	0,027
197911	0,019	198812	0,009	199607	0,096	200401	0,006	201107	0,012
198001	0,037	198901	0,004	199608	0,004	200402	0,007	201108	0,014
198002	0,023	198902	0,001	199609	0,001	200403	0,012	201109	0,015
198003	0,011	198903	0,002	199610	0,014	200404	0,012	201110	0,015
198004	0,012	198904	0,001	199611	0,001	200405	0,015	201111	0,017
198006	0,002	198905	0,028	199612	0,005	200406	0,042	201112	0,021
198007	0,008	198906	0,017	199701	0,016	200407	0,004	201201	0,015
198008	0,002	198907	0,034	199702	0,018	200408	0,005	201202	0,008
198009	0,023	198908	0,011	199703	0,004	200409	0,014	201203	0,013
198010	0,009	198909	0,001	199704	0,001	200410	0,001	201204	0,009
198011	0,001	198910	0,017	199705	0,025	200411	0,016	201205	0,007
198101	0,003	198911	0,014	199706	0,012	200412	0,016	201206	0,006
198103	0,001	198912	0,017	199707	0,018	200501	0,007	201207	0,007
198104	0,005	199001	0,038	199708	0,043	200502	0,014	201208	0,005
198105	0,015	199002	0,016	199709	0,027	200503	0,028	201209	0,011
198107	0,015	199003	0,021	199710	0,015	200504	0,007	201210	0,015
198109	0,003	199004	0,004	199711	0,016	200505	0,001	201211	0,021
198110	0,004	199005	0,002	199712	0,001	200506	0,009	201212	0,008
198111	0,001	199006	0,009	199801	0,001	200507	0,004	201301	0,012
198112	0,001	199007	0,008	199802	0,019	200508	0,004	201302	0,009
198201	0,022	199008	0,007	199803	0,032	200509	0,003	201303	0,012
198202	0,001	199009	0,002	199804	0,001	200510	0,009	201304	0,011
198203	0,016	199011	0,011	199805	0,013	200511	0,032	201305	0,035
198204	0,004	199012	0,006	199806	0,002	200512	0,006	201306	0,177
198205	0,011	199101	0,029	199807	0,016	200601	0,009	201307	0,039
198206	0,002	199102	0,012	199808	0,007	200602	0,032	201308	0,018
198208	0,009	199103	0,028	199809	0,013	200603	0,031	201309	0,002
198209	0,002	199104	0,025	199810	0,027	200604	0,042	201310	0,011
198211	0,001	199105	0,002	199811	0,054	200605	0,014	201311	0,013
198301	0,003	199106	0,009	199812	0,017	200606	0,003	201312	0,028
198302	0,007	199107	0,032	199901	0,001	200607	0,007	201401	0,013
198303	0,001	199108	0,001	199902	0,115	200608	0,006	201402	0,019
198304	0,001	199109	0,003	199903	0,026	200609	0,013	201403	0,028
198305	0,006	199110	0,018	199904	0,015	200610	0,007	201404	0,014
198306	0,005	199111	0,009	199905	0,031	200611	0,025	201405	0,044
198307	0,011	199112	0,007	199906	0,022	200612	0,005	201406	0,002

198308	0,002	199201	0,005	199907	0,032	200701	0,014	201407	0,013
198309	0,006	199202	0,009	199908	0,011	200702	0,004	201408	0,001
198311	0,004	199203	0,032	199909	0,011	200703	0,015	201409	0,012
198312	0,001	199204	0,013	199910	0,011	200704	0,039	201410	0,004
198401	0,005	199205	0	199911	0,028	200705	0,046	201411	0,001
198402	0,012	199206	0,007	199912	0,003	200706	0,007	201412	0,013
198403	0,016	199207	0,011	200001	0,011	200707	0,011	201501	0,012
198404	0,026	199208	0,009	200002	0,028	200708	0,023	201502	0,016
198405	0,007	199209	0,003	200003	0,012	200709	0,006	201503	0,011
198406	0,036	199210	0,005	200004	0,021	200710	0,035	201504	0,001
198407	0,005	199211	0,015	200005	0,013	200711	0,014	201505	0,018
198408	0,012	199212	0,002	200006	0,022	200712	0,001	201506	0,014
198409	0,009	199301	0,005	200007	0,031	200801	0,006	201507	0,021
198410	0,007	199302	0,001	200008	0,029	200802	0,002	201508	0,001
198411	0,015	199303	0,005	200009	0,002	200803	0,002	201509	0,011
198501	0,004	199304	0,014	200010	0,021	200804	0,008	201510	0,018
198502	0,002	199305	0,026	200011	0,029	200805	0,001	201511	0,016
198503	0,002	199306	0,034	200012	0,125	200806	0,005	201512	0,016
198504	0,007	199307	0,016	200101	0,028	200807	0,006	201601	0,001
198505	0,007	199308	0,016	200102	0,035	200808	0,006	201602	0,012
198506	0,001	199309	0,005	200103	0,013	200809	0,008	201603	0,017
198507	0,005	199310	0,008	200104	0,067	200810	0,012	201604	0,019
198508	0,002	199311	0,004	200105	0,017	200811	0,012	201605	0,009
198510	0,007	199312	0,007	200106	0,026	200812	0,013	201606	0,011
198511	0,001	199401	0,015	200107	0,014	200901	0,016	201607	0,034
198512	0,001	199402	0,001	200108	0,039	200902	0,011	201608	0,015
198601	0,003	199403	0,024	200109	0,021	200903	0,008	201609	0,011
198602	0,006	199404	0,001	200110	0,018	200904	0,012	201610	0,001
198603	0,003	199405	0,011	200111	0,024	200905	0,011	201611	0,017
198605	0,005	199406	0,005	200112	0,011	200906	0,002	201612	0,011
198606	0,002	199407	0,011	200201	0,018	200907	0,013	201701	0,007
198607	0,007	199408	0,007	200202	0,007	200908	0,004	201702	0,008
198609	0,001	199409	0,014	200203	0,024	200909	0,006	201703	0,035
198610	0	199410	0,004	200204	0,022	200910	0,015	201704	0,016
198612	0,019	199411	0,018	200205	0,016	200911	0,006	201705	0,019
198701	0,027	199412	0,001	200206	0,004	200912	0,017	201706	0,011
198703	0,024	199501	0,004	200207	0,004	201001	0,012	201707	0,017
198704	0,001	199502	0,006	200208	0,017	201002	0,007		
198705	0,006	199503	0,052	200209	0,004	201003	0,006		
198706	0,002	199504	0,033	200210	0,011	201004	0,001		
198707	0,008	199505	0,014	200211	0,012	201005	0,001		
198708	0,008	199506	0,014	200212	0,002	201006	0,035		
198709	0,003	199507	0,018	200301	0,024	201007	0,007		
198710	0,001	199508	0,015	200302	0,024	201008	0,004		
198711	0,001	199509	0,003	200303	0,041	201009	0,006		
198712	0,012	199510	0,013	200304	0,001	201010	0,003		

198802	0,001	199511	0,011	200305	0,008	201011	0,005		

EK 3 1979-2017 Yılları Arasında Ukrayna'daki Aylık Protesto Yoğunlukları (SQL ÇIKTILARI)

MonthYear	Percent	MonthYear	Percent	MonthYear	Percent	MonthYear	Percent	MonthYear	Percent
197909	0,002	199212	0,004	199902	0,007	200504	0,013	201106	0,001
198004	0,004	199301	0,001	199903	0,009	200505	0,007	201107	0,002
198102	0,001	199302	0,005	199904	0,005	200506	0,009	201108	0,006
198103	0,003	199303	0,009	199905	0,009	200507	0,006	201109	0,005
198104	0,001	199304	0,008	199906	0,003	200508	0,005	201110	0,006
198105	0,006	199305	0,007	199907	0,007	200509	0,01	201111	0,006
198106	0,001	199306	0,014	199908	0,003	200510	0,006	201112	0,004
198107	0,001	199307	0,003	199909	0,004	200511	0,009	201201	0,004
198110	0,001	199308	0	199910	0,004	200512	0,003	201202	0,003
198202	0,001	199309	0,014	199911	0,007	200601	0,007	201203	0,003
198303	0,001	199310	0,006	199912	0,003	200602	0,005	201204	0,007
198307	0,001	199311	0,005	200001	0,002	200603	0,013	201205	0,005
198311	0,001	199312	0,006	200002	0,004	200604	0,007	201206	0,001
198401	0,004	199401	0,011	200003	0,008	200605	0,006	201207	0,005
198403	0,001	199402	0,012	200004	0,004	200606	0,014	201208	0,001
198404	0,001	199403	0,009	200005	0,005	200607	0,007	201209	0,002
198405	0,002	199404	0,003	200006	0,004	200608	0,002	201210	0,004
198409	0,004	199405	0,007	200007	0,005	200609	0,002	201211	0,006
198511	0,002	199406	0,002	200008	0,004	200610	0,006	201212	0,002
198604	0,001	199407	0,003	200009	0,01	200611	0,005	201301	0,002
198605	0,002	199408	0,001	200010	0,004	200612	0,004	201302	0,002
198606	0,001	199409	0,002	200011	0,005	200701	0,003	201303	0,002
198608	0,002	199410	0	200012	0,014	200702	0,002	201304	0,006
198609	0,003	199411	0,005	200101	0,016	200703	0,004	201305	0,007
198702	0,001	199412	0,003	200102	0,022	200704	0,018	201306	0,003
198704	0,001	199501	0	200103	0,041	200705	0,002	201307	0,004
198706	0,001	199502	0	200104	0,013	200706	0,003	201308	0,002
198707	0,001	199503	0,009	200105	0,013	200707	0,002	201309	0,001
198708	0,003	199504	0,001	200106	0,011	200708	0,002	201310	0,002

198709	0,002	199505	0,008	200107	0,002	200709	0,002	201311	0,017
198711	0,001	199506	0,005	200108	0,004	200710	0,005	201312	0,047
198712	0,001	199507	0,003	200109	0,001	200711	0,004	201401	0,035
198803	0,002	199508	0,006	200110	0,002	200712	0,003	201402	0,064
198806	0,002	199509	0,018	200111	0,005	200801	0,002	201403	0,073
198810	0,001	199510	0,034	200112	0,003	200802	0,002	201404	0,041
198811	0,003	199511	0,01	200201	0,003	200803	0,005	201405	0,032
198902	0,001	199512	0,005	200202	0,003	200804	0,002	201406	0,015
198903	0,003	199601	0,004	200203	0,003	200805	0,002	201407	0,017
198906	0,019	199602	0,006	200204	0,004	200806	0,001	201408	0,013
198907	0,009	199603	0,006	200205	0,005	200807	0,001	201409	0,011
198908	0,005	199604	0,006	200206	0,006	200808	0,002	201410	0,008
198909	0,018	199605	0,006	200207	0,008	200809	0,002	201411	0,007
198910	0,011	199606	0,005	200208	0,008	200810	0,003	201412	0,007
198911	0,017	199607	0,008	200209	0,043	200811	0,002	201501	0,006
198912	0,012	199608	0,008	200210	0,012	200812	0,001	201502	0,012
199001	0,003	199609	0,001	200211	0,007	200901	0,004	201503	0,011
199003	0,005	199610	0,007	200212	0,007	200902	0,003	201504	0,006
199004	0,003	199611	0,001	200301	0,002	200903	0,005	201505	0,004
199007	0,006	199612	0,004	200302	0,006	200904	0,003	201506	0,005
199010	0,021	199701	0,003	200303	0,016	200905	0,002	201507	0,005
199011	0,001	199702	0,002	200304	0,005	200906	0,002	201508	0,005
199012	0,006	199703	0,01	200305	0,011	200907	0,003	201509	0,006
199101	0,004	199704	0,001	200306	0,005	200908	0,002	201510	0,003
199103	0,002	199705	0,004	200307	0,003	200909	0,001	201511	0,003
199104	0,018	199706	0,004	200308	0,002	200910	0,002	201512	0,004
199105	0,001	199707	0,006	200309	0,009	200911	0,002	201601	0,003
199106	0,006	199708	0,009	200310	0,015	200912	0,001	201602	0,005
199107	0,007	199709	0	200311	0,006	201001	0,004	201603	0,01
199108	0,007	199710	0,003	200312	0,007	201002	0,006	201604	0,006
199109	0,004	199711	0,004	200401	0,007	201003	0,006	201605	0,003
199110	0	199712	0,005	200402	0,009	201004	0,007	201606	0,003
199111	0,004	199801	0	200403	0,011	201005	0,007	201607	0,003
199112	0,004	199802	0,005	200404	0,006	201006	0,002	201608	0,004

199201	0,01	199803	0,009	200405	0,009	201007	0,004	201609	0,003
199202	0,001	199804	0,003	200406	0,006	201008	0,002	201610	0,003
199203	0,018	199805	0,004	200407	0,005	201009	0,002	201611	0,005
199204	0,002	199806	0,007	200408	0,017	201010	0,006	201612	0,003
199205	0,001	199807	0,002	200409	0,009	201011	0,009	201701	0,002
199206	0,001	199808	0,002	200410	0,014	201012	0,006	201702	0,004
199207	0,011	199809	0,004	200411	0,063	201101	0,003	201703	0,005
199208	0,005	199810	0,005	200412	0,036	201102	0,002	201704	0,002
199209	0,011	199811	0,005	200501	0,016	201103	0,002	201705	0,003
199210	0,006	199812	0,008	200502	0,012	201104	0,003	201706	0,003
199211	0,005	199901	0,003	200503	0,014	201105	0,005	201707	

EK 4 Çatışmalar için R Kodları

```
> install.packages("poweRlaw")
> library("poweRlaw")
> data("çaltışma verisi", package="poweRlaw")
> m1_pl = displ$new(çatışma)
> est = estimate_xmin(m1_pl)
> m1_pl$setXmin(est)
> get_distance_statistic(m1_pl, xmax = 1e+05, distance = "ks")
> plot(m1_pl)
> lines(m1_pl, col=2)
> bstrp <- bootstrap(m1_pl, no_of_sims = 5000, threads = 4, seed = 1)
> sd(bs$bootstraps[,2])
> sd(bs$bootstraps[,3])
> bs3_p = bootstrap_p(m1_pl)
> bs3_p$p
> install.packages('pracma')
> library(pracma)
> zeta(α)
```

EK 5 Protestolar için R Kodları

```
> install.packages("poweRlaw")
> library("poweRlaw")
> data("veri", package="poweRlaw")
> m2_bl  = conpl$new(protesto)
> est = estimate_xmin(m2_bl )
> m2_bl $setXmin(est)
> plot(m2_bl )
> lines(m2_bl , col=2, lwd=2)
```

```
> get_KS_statistic(m2_bl , xmax = 1e+05, distance = "ks")
> bstrp <- bootstrap(m2_bl ,no_of_sims = 5000, threads = 4, seed = 1)
> sd(bs$bootstraps[, 2])
> sd(bs$bootstraps[, 3])
> bs3_p = bootstrap_p(m2_bl )
> bs3_p$p
> install.packages('pracma')
> library(pracma)
> zeta(α)
```

ÖZGEÇMİŞ			
Ad, Soyadı	Sadullah		ÇELİK
Bildiği Yabancı Diller	İngilizce		
Eğitim Durumu	**Başlama-Bitirme Yılı**		**Kurum Adı**
Lise	2000	2004	Turgutlu Niyazi Üzmez Lisesi (Yabancı Dil Ağırlıklı Lise)
Lisans	2007	2011	Celal Bayar Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü
Yüksek Lisans	2011	2013	Ege Üniversitesi Fen Bilimleri Enstitüsü Matematik
Doktora	2013	2017	Uludağ Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri
Çalıştığı Kurum	**Başlama-Ayrılma Yılı**		**Çalıştığı Kurumun Adı**
1.	2014	-	Adnan Menderes Üniversitesi Nazilli İİBF Ekonometri Bölümü
Üye Olduğu Bilimsel Mesleki Kuruluşlar	-		
Katıldığı Proje ve Toplantılar	**Katıldığı Bilimsel Toplantılar** **1.** 15 th International Symposium on Econometrics, Operations Research and Statistics, Süleyman Demirel University, Isparta, 22-25 May 2014 (Katılımcı) **2.** Sakarya Üniversitesi, Sakarya Ekonometri Seminerleri, "Eviews ile Uygulamalı Ekonometri", Sakarya, 18-22 Ağustos 2014. **3.** Sakarya Üniversitesi, Sakarya Ekonometri Seminerleri, "SPSS ile Veri Analizi", Sakarya, 25-29 Ağustos 2014. **4.** Sakarya Üniversitesi, Sakarya Ekonometri Seminerleri, "Zaman Serileri Analizi", Sakarya, 8-12 Eylül 2014.		

	5. İstanbul Teknik Üniversitesi İşletme Fakültesi Ekonomi Bölümü, Uluslararası İTÜ Ekonomi Yaz Okulu Programı, "Yapısal Eşitlik Modelleri ve Çok Değişkenli İstatistikn I", İstanbul, 3-21 Ağustos 2015. **6.** Sakarya Üniversitesi, Sakarya Ekonometri Seminerleri, "Zaman Serileri Analizi", Sakarya, 17-21 Ağustos 2015. **7.** I. Uluslararası Ekonomi, Finans ve Ekonometri Öğrenci Sempozyumu (EFEOS), Sakarya, 17-18 Mayıs 2017 (1 Bildirili). **8.**
Yayınlar:	**Uluslararası Makale** **1.** ÇELİK, S., "GDELT Verileri Kullanılarak Kuvvet Yasası Dağılımı İçin Uyum İyiliği Testi", Sosyal Bilimler Dergisi, The Journal of Social Science, Sayı:14, Eylül 2017, s. 452-464. **2.** ÇELİK, S., "Büyük Veri Teknolojilerinin İşletmeler İçin Önemi", Social Sciences Studies Journal, Sayı:9, Kasım 2017, s. 873-883.
Diğer:	
İletişim (e-posta):	ssadullah.celik@gmail.com

Tarih	**30.12.2017**
İmza	
Adı Soyadı	**Sadullah ÇELİK**