



**YILDIZ TEKNİK ÜNİVERSİTESİ
KİMYA-METALÜRJİ FAKÜLTESİ
MATEMATİK MÜHENDİSLİĞİ BÖLÜMÜ**

**MTM4511 MATEMATİK MÜHENDİSLİĞİNDE
TASARIM UYGULAMALARI**

BÜYÜK VERİ VE BÜYÜK VERİ ARAÇLARI

TEZ YÖNETİCİSİ
Prof. Dr. Reşat KÖŞKER

ÖĞRENCİ
Ayşe Nur ÇAPKAN
18052026

İstanbul, 2022

© Bu tezin bütün hakları Yıldız Teknik Üniversitesi Matematik Mühendisliği Bölümü'ne aittir.

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ	iv
ŞEKİL LİSTESİ	v
TABLO LİSTESİ	vi
ÖNSÖZ	vii
ÖZET	viii
ABSTRACT	ix
1. GİRİŞ	1
2. BÜYÜK VERİ	2
3. BÜYÜK VERİ ARAÇLARI	10
3.1 Apache Hadoop	11
3.2 Apache Spark	14
3.3 Apache Hive	16
3.4 MongoDB	17
3.5 ElasticSearch	19
3.6 Apache Kafka ve Apache Zookeeper	21
3.7 Apache Cassandra	23
4. SONUÇLAR	25
KAYNAKLAR	26
ÖZGEÇMİŞ	28

KISALTMA LİSTESİ

BSON	Binary JSON
CQL	Cassandra Query Language (Cassandra Sorgu Dili)
FIFO	First In First Out (İlk giren ilk çıkar)
HDFS	Hadoop Distributed File System (Hadoop Dağıtık Dosya Sistemi)
HiveQL	Hive Query Language (Hive Sorgu Dili)
JSON	JavaScript Object Notation (Javascript Nesne Gösterimi)
MLlib	Machine Learning Library (Makine Öğrenimi Kütüphanesi)
MQL	MongoDB Query Language (MongoDB Sorgu Dili)
NoSQL	Not Only SQL (İlişkisel olmayan)
RDD	Resilient Distributed Dataset (Esnek Dağıtılmış Veri Kümesi)
SQL	Structured Query Language (Yapılandırılmış Sorgu Dili)
XML	Extensible Markup Language (Genişletilebilir İşaretleme Dili)
YARN	Yet Another Resource Negotiator (Kaynak Yöneticisi)

ŞEKİL LİSTESİ

Sayfa

ŞEKİL 2.1.	2021 yılında bir dakikada internette yapılan işlemler	2
ŞEKİL 2.2.	2025 tarihine kadar üretilmesi öngörülen veri miktarları	8
ŞEKİL 3.1.	Büyük veri nerelerden elde edilir temalı görsel	10
ŞEKİL 3.1.1.	Apache Hadoop logosu	11
ŞEKİL 3.1.2.	MapReduce kelime sayma işlemi	13
ŞEKİL 3.2.1.	Apache Spark logosu	14
ŞEKİL 3.3.1.	Apache Hive logosu	16
ŞEKİL 3.4.1.	MongoDB logosu	17
ŞEKİL 3.4.2.	JSON örneği	18
ŞEKİL 3.5.1.	ElasticSearch logosu	19
ŞEKİL 3.6.1.	Apache Kafka logosu	21
ŞEKİL 3.6.2.	Apache ZooKeeper logosu	21
ŞEKİL 3.7.1.	Apache Cassandra logosu	23

TABLO LİSTESİ

Sayfa

TABLO 2.1.	Yapısal, yarı yapısal ve yapısal olmayan veri örnekleri	4
TABLO 2.2.	Bayt birimleri ve değerleri	7
TABLO 3.1.1.	HDFS çalışma prensibi	12
TABLO 3.5.1.	ElasticSearch çalışma prensibi	20

ÖNSÖZ

Bu tezi hazırlama sürecimde destekleriyle ve yardımlarıyla yanımda olan başta sayın Prof. Dr. Reşat Köşker'e ve sevgili aileme teşekkür ederim.

ÖZET

Sosyal medya, sensörler ve IoT cihazları gibi farklı kaynaklardan elde edilen veri miktarları gün geçtikçe artmaya başlamıştır. Bu durumun sonucunda da veriler önceden kullanılan ilişkisel veritabanları gibi yöntemlerle depolanamaz ve işlem yapılamaz hale gelmiştir. Büyük veri kavramı bu noktada durumu özetleyen bir kavram olarak nitelendirilmiştir. Geleneksel yöntemler ile depolanamayan ve işlem yapılamayan büyük veri kavramı verinin hız, gerçeklik, hacim, çeşitlilik ve değer özelliklerini sağlamasıyla tanıtılmıştır. Saklama ve işleme sorunları sonucunda çalışmalar yapılmış ve büyük veri araçları geliştirilmiştir. Bu çalışmada Büyük Veri ve Büyük Veri Araçları hakkında genel bilgilere yer verilmiştir. Yapılan araştırma sonuçlarından ve bilgi birikiminden yararlanılarak okuyucuya bu konular hakkında bilgiler örnekler ile verilmiştir. Okuyucunun bu konular hakkında bilinçlenmesi amaçlanmıştır.

Büyük veri hakkında ilgili tanımlar, kullanım örnekleri ve avantajları gibi temel başlıklar üzerinde durulduktan sonra Büyük Veri Araçları tanıtılıp çeşitli örnekler ile ilgili araçların çalışma prensipleri şekillerle gösterilmiştir. Yapılan araştırma sonucunda okuyucuya aktarılması gereken araçlar belirlenmiş ve bu araçlar üzerinde durulmuştur.

Anahtar kelimeler: Büyük Veri, 5V, Spark, Hadoop, MapReduce, ElasticSearch, MongoDB, Cassandra, Hive, Kafka, Apache Zookeeper, HDFS

ABSTRACT

The amount of data obtained by different sources like social media, sensors and IoT devices is increasing day by day. As a result of this situation, data can no longer be stored and processed by previously used methods like relational database. The big data is qualified as a concept which summarizes the situation at this point. The big data which cannot be stored and processed by traditional methods has been introduced by providing the velocity, veracity, volume, variety and value features of data. As a result of storing and processing issues, studies have been made and Big Data Tools have been developed. In this study, general information about Big Data and Big Data Tools has been given. The information was given to the reader with examples by using the research results and knowledge in this study. It is aimed to make the reader aware of the subjects.

After emphasizing on topics like definition, usage examples and advantages of Big Data, Big Data Tools has been introduced and working principles of these tools are illustrated with figures. As a result of the research, the tools which should be explained to the reader were determined and these tools were emphasized.

Key Words: Big Data, 5V, Spark, Hadoop, MapReduce, ElasticSearch, MongoDB, Cassandra, Hive, Kafka, Apache Zookeeper, HDFS

1. GİRİŞ

Geçmişten günümüze kadar insanların yaptıkları işler kayıt altında tutulmuştur. Önceleri dosyalarda tutulan kayıtlar bilgisayarın keşfi ve yaygınlaşması ile elektronik ortamlarda tutulmaya başlanmıştır. Eskiden elektronik ortamlarda tutulan veriler güvenilir iken şu an çok farklı kaynaklardan gelen veriler çeşitli işlemler için tutulmaktadır. Bu verilerin depolanmadan önce güvenilirliği ve doğruluğu çok kısa zamanlarda kontrol edilmelidir. Veri miktarı arttıkça bu durum zorlayıcı bir hale gelmektedir. Büyük Veri kavramı verinin miktarı yanında bazı diğer özelliklerin de sağlanması ile elde edilir. Güvenilirlik de bu özelliklerden biridir. Büyük verinin hacminden dolayı elektronik ortamlarda tutulması zorlaştığından Büyük Veri Araçları depolama gibi işlemlerin kolayca yapılmasına olanak sağlamak için geliştirilmiştir. İlgili çalışmada büyük veri ve büyük veri araçları iki başlık altında incelenecektir.

Devasa boyutlarda olmasına ek farklı özelliklerle tanımlanan büyük verinin tanımı 2.Bölüm'de verilmiştir. Büyük veri tanımına ek olarak bu bölümde büyük verinin bileşenleri, yapısal, yarı yapısal ve yapısal olmayan veri örnekleri, büyük verinin avantajları, büyük verinin farklı sektörlerdeki kullanım örnekleri, büyük veriyi etkin kullanan şirketler gibi konulardan bahsedilmiştir.

Bir sonraki bölüm olan Büyük Veri Araçları Bölümü'nde büyük veri araçlarına neden ihtiyaç duyulduğuna değinilip çeşitli büyük veri araçlarından bahsedilmiştir. Bu büyük veri araçları sırasıyla Apache Hadoop, Apache Spark, Apache Hive, MongoDB, Elasticsearch, Apache Kafka ve Apache ZooKeeper, Apache Cassandra olarak belirlenmiştir. Bu araçların incelenmesi sırasında konuyu açıklayacak örnekler ve şekiller kullanılmıştır.

2. BÜYÜK VERİ

Son birkaç yılda insanlığın bugüne kadar oluşturduğundan daha fazla veri üretildi. Bunun en büyük sebebi ise teknolojinin gelişmesi ve internetin yaygınlaşmasıdır. İnternetin yaygınlaşması sayesinde insanların anlık olarak neler yaptıkları gerek sosyal medya ile gerek ise arama geçmişlerinde bıraktıkları izler sayesinde bilinir. 2021 yılında bir dakikada internette yapılan işlemler Şekil 2.1’de gösterilmektedir. Şekilde 1 dakikada App Store ve Google Play’den 414.764 uygulama indirme, 21.1 milyon mesaj gönderme, Snapchat’te 3.4 milyon snap oluşturma ve Instagram’da 695.000 hikaye paylaşma gibi işlemlerin yapıldığı görülmektedir.

2021 *This Is What Happens In An Internet Minute*



Şekil 2.1. 2021 yılında bir dakikada internette yapılan işlemler

Yapılan bu işlemler sonucunda ortaya çıkan veri ise günümüz dünyasında büyük bir öneme sahiptir. Çünkü çeşitli kurum ve kuruluşlar bu verileri analiz ederek kendi çıkarları doğrultusunda kullanmaktadır.

Veriler tek başlarına kullanılabilir durumda değildirler. Verilerin kullanılabilir hale gelebilmeleri için anlamlı hale getirilmeleri gerekir. Verilerin anlamlandırılmış haline ise bilgi denir. Bu tanım aklımıza veri nedir sorusunu getirebilir. Veri İngilizcedeki data kelimesinden veri olarak Türkçeye çevrilmiştir. Anlamı ise deney, gözlem, ölçüm ve araştırma gibi yöntemler sonucu elde edilmiş her türlü gerçektir[1]. Deney, gözlem, ölçüm ve araştırma gibi yöntemler sonucu farklı türde veriler üretilir ve bu farklı tipteki veriler eski yöntemler ile depolanamaz.

İnternetin yaygınlaşması ve teknolojinin gelişmesi sonucu oluşan veriler büyük veri problemini gündeme getirmiştir. Büyük veri; geleneksel yöntemler ile depolanması, işlenmesi, analiz edilmesi güç olan verilerdir. Her ne kadar büyük veri dediğimizde aklımıza verinin hacimsel boyutu gelse de bir verinin büyük veri olarak isimlendirilmesinde hacim yeterli bir kriter değildir. Bir verinin büyük veri olarak nitelendirilebilmesi için Gartner 3V ana bileşeni tanımlamıştır. Bu bileşenler; hacim(volume), çeşitlilik(variety) ve hız(velocity) olarak belirtilmiştir. İlerleyen zamanlarda bu ana bileşenler yetersiz kalmıştır. Bu bileşenlere değer(value) ve gerçeklik(veracity) de eklenerek büyük verinin 5V bileşenlerini oluşturmuşlardır. Bu bileşenler ve özellikleri ise şöyledir[2]:

- Hacim (Volume): Büyük veri denilince akla gelen ilk özelliktir ve verinin miktarını ifade eder. Büyük verinin birçok farklı kaynaklardan üretilmesi exabayt, zetabayt gibi birimlerle ölçülmesi ile sonuçlanmıştır. Büyük verinin hacminin bu kadar büyük olması ise depolama sorununu doğurmaktadır. Bu dezavantajın yanında veri miktarının büyük olması istatistiksel analizlerden daha az hata oranı ile doğru sonuçlar alma olasılığını artırır.
- Çeşitlilik (Variety): Çeşitlilik özelliği büyük verinin farklı kaynaklardan elde edilmesini tanımlar. Bu kaynaklardan bazıları örnek olması için şöyle sıralanabilir[3]:
 - Sosyal Medya Hesaplarında Yapılan Hareketler
 - Arama Geçmişleri
 - Banka Hesaplarında Yapılan Hareketler
 - E-mailler
 - Güvenlik Sistemleri
 - Akıllı Saatler
 - Alışveriş Sitelerinde Yapılan Hareketler
 - GPS Verileri
 - Video ve Kameralardan Edinilen Görüntüler
 - Bloglar
 - Sensörler

Bu farklı kaynaklardan elde edilen veriler yapısal, yarı yapısal ve yapısal olmayan olarak tanımlanır. Bu tanımlamalara Tablo 2.1'deki gibi çeşitli örnekler verilebilir.

TABLO 2.1. Yapısal, yarı yapısal ve yapısal olmayan veri örnekleri

Yapısal Veri	Yarı Yapısal Veri	Yapısal Olmayan Veri
<ul style="list-style-type: none">• İlişkisel Veritabanları• İstatiksel Veriler• Excel Tabloları	<ul style="list-style-type: none">• Sensörlerden Alınan Veriler• XML• JSON	<ul style="list-style-type: none">• Video• Görüntü• Ses• Konum Bilgileri• Tweetler• Makaleler• Blog Yazıları• E-Mailler

Geleneksel analiz yöntemleri yapısal verileri analiz etmeye ve işlemeye müsaittir. Bu nedenle verinin yapısal, yarı yapısal ve yapısal olmayan olarak tanımlanması geleneksel analiz yöntemleri ile analiz edilememeleri ve işlenememeleri sorununu doğurmaktadır. Bu sorun Büyük Veri Araçları ile aşılmaktadır.

- **Hız (Velocity):** Büyük verinin değişken hızlarla üretilmesini tanımlayan bir özelliktir. Büyük verinin üretilme hızı yüksektir. Veri üretim hızına örnek bir şekil ve açıklaması Giriş Bölümü’nde gösterilen Şekil 1.1’de verilmişti. Üretilen bu hızlı verinin yine aynı hızda işlenip analiz edilmesi gerekmektedir.
- **Değer (Value):** Üretilen veriden değer elde etmeyi amaçlayan bir özelliktir. Büyük veride önemli olan üretilen verinin miktarından ziyade bu verilerden nasıl yararlanıldığıdır. Uygun ve anlık kararlar verebilmek için veriden değer üretmek gerekmektedir.
- **Gerçeklik (Veracity):** Verinin doğru olmasını tanımlayan bir özelliktir. Veriler farklı kaynaklardan üretildiği için doğruluğu üzerinde iyi tespitler yapılmalıdır çünkü büyük veriden yararlanıp analiz yapacak kurum, kuruluş ve şahısların doğru, güvenilir ve tarafsız bilgilere ihtiyaçları vardır.

Büyük veri, ilk olarak Michael Cox ve David Ellsworth tarafından 1997 yılında düzenlenen 8. IEEE Görüntüleme Konferansı’nda (Proceedings of the 8th Conference on Visualization), “Application-Controlled Demand Paging for Out-of-Core Visualization” adlı makalede kullanılmıştır[2]. İlk kullanılışından yıllar sonra ise büyük veri ivme kazanmış, çoğu bilişim şirketi tarafından dikkate alınmış ve üzerinde çalışılmıştır.

Farklı kaynaklardan elde edilen verilerin hepsi anlamlı, güvenilir ve doğru değildir. Bu nedenle bu verilerin öncelikle depolanması daha sonra ise işlenmesi gerekir. Büyük veri geleneksel yöntemler ile depolanamaz ve işlenemez. Çünkü bu yöntemler büyük veri için

yetersiz kalmaktadır[4]. Bu nedenle büyük veri üzerinde işlemler yapabilmek için çeşitli araçlar geliştirilmiştir. Büyük veri için geliştirilen bu araçlar sayesinde veriler depolandı, işlendi ve ilgili kurum ya da kuruluşların işine yarayacak bilgiler haline geldi. Verilerin depolanması ile veri artışı daha da hızlandı. Şirketler, büyük verinin yol göstericiliğinden yararlanmak için büyük veri araçlarına yatırım yapar hale geldi.

Büyük veri uygun depolama ve analiz yöntemleri ile kurum ve kuruluşlara çeşitli avantajlar sağlayabilir. Bu avantajları doğru yöneten kurum ve kuruluşlar yatırımlarını farklı alanlarda değerlendirerek kâr elde edebilir. Bu avantajlarından bazılarını ise şöyle sıralayabiliriz[5, 6, 7]:

- Daha iyi müşteri deneyimi sunmak.
- Değişen müşteri ihtiyaçlarına göre yeni pazarlama stratejileri belirleyerek ihtiyaçlara hızlı cevap vermek.
- Ürün ve satış çeşitliliği sağlamak.
- Analizler yaparak öngörülerde bulunmak.
- Zamandan ve maliyetten tasarruf etmek.
- Gerçek zamanlı doğru verilere ulaşılmasını mümkün kılmak.
- Yapılan hata ve kusurların kolaylıkla görünmesini ve hızlı müdahale edilmesini sağlamak.
- Müşteri kaybı sebeplerini belirleyerek uygun önleyici politika izlemek.
- Dolandırıcılık tespitinde bulunmak.
- İşlem süresini azaltmak.
- Kriz dönemlerinde hızlı cevap üretmek.

Bu avantajları kullanan şirketler isimlerini yükseltmişlerdir ve oldukça kâr elde etmişlerdir. Şirketlerin büyük veri kullanımlarına günlük hayatta şahit olunmaktadır. Netflix'in kullanıcının izlediği film ve dizi verilerine göre kişiye özel film önermesi, sosyal medya uygulamalarında kişinin karşısına arama geçmişinde bıraktığı izlere özel reklam çıkması, bir alışveriş sitesinde aranılan kelime için en yakın önerilerin çok kısa sürede çıkması büyük verinin ve büyük veri araçlarının kullanımına dair en sık karşılaşılan örnekleri olarak verilmektedir.

Bu örnekler hem Dünya hem de Türkiye bazında çoğaltılabilir. Büyük veri birçok farklı sektörü etkileyerek çeşitli kolaylıklar sağlamıştır. Günlük hayattaki büyük veri kullanımı örneklerini farklı sektörler üzerinden şöyle çoğaltabiliriz:

- Sağlık Sektörü: Sağlık sektöründe büyük veri sayesinde hastalık teşhisinde önceden DNA analizi yapılan hastaların verileri sayesinde bir başka bireye daha az işlem ile doğru tanı konulabilir[8]. Yeni ilaçların ve tedavilerin geliştirilmesinde de bu verilerden yararlanılır. Türkiye'de büyük verinin sağlık alanındaki uygulamalarına e-Nabız, Sağlık.NET, MHRS, Aşı Takip örnek olarak verilebilir[9].
- Bankacılık Sektörü: Bankacılık sektöründe büyük veri sayesinde müşterilerin hesap hareketleri analiz edilerek onlara uygun kredi tekliflerinde bulunulabilir. Aynı zamanda büyük veri sayesinde dolandırıcılıkların önüne geçilebilir.

- Ulaşım Sektörü: Ulaştırma sektöründe büyük veri sayesinde trafik kalıpları analiz edilerek büyük şehirlerde trafiğin önüne geçilebilir[10]. Aynı zamanda sensörler, GPS verileri ve sosyal medya paylaşımları ile trafik kazalarından da kaçınılabılır[11].
- Tarım Sektörü: Tarım sektöründe büyük veri sayesinde önceki yılların iklim koşullarından, toprağın besin yapısından ve yağışlardan tahminler çıkartılarak sonraki yıllara uygun bir planlama ile toprak ekilip biçilebilir ve daha verimli tarım yapılabilir.[12]. Bu sayede giderek artan nüfus yoğunluğundan doğabilecek gıda kıtlığı önlenabilir.
- Eğitim Sektörü: Eğitim sektöründe büyük veri sayesinde eğitim kalitesi artırılır, öğrenciye özgü öğrenme modeli sağlanır, öğrenci performansı iyileştirilir, eğitim müfredatı planlanır, eğitimde verimsiz idari süreçler belirlenerek iyileştirilir, ders içerikleri yeniden yapılandırılır, eğitmen ve idarecilerin öğrenci performansını takip etmesi sağlanır[13]. Türkiye’de eğitim sektöründeki büyük veri uygulamalarına MEBBİS, e-Okul, FATİH, e-YAYGIN örnek olarak verilebilir[9].
- Güvenlik Sektörü: Güvenlik sektöründeki büyük veri uygulamalarına örnek olarak Los Angeles Polis Departmanı ve Kaliforniya Üniversitesi’nin işbirliği ile hazırlanan çalışma gösterilebilir. Bu çalışmada son 80 yılda işlenmiş olan 13 milyon suç dosyasının analiz edilmesi sonucunda oluşabilecek yeni suçlar tahmin edilebilmektedir. Çalışma sonucunda polisin suç daha gerçekleşmeden uyguladığı önlemler ile suç oranlarında düşüşler yaşanmıştır[8].

Büyük veri farklı sektörlerde de önem kazanarak mihenk taşı konumuna gelecektir. Büyük verinin önemini kavrayan şirketler ise yükselişe geçecektir. Bu noktada şirketlerin yapması gereken maliyetlerini azaltıp kâr elde edebilecekleri bir alan olan büyük veri ve büyük veri araçlarına yatırım yapmaktır.

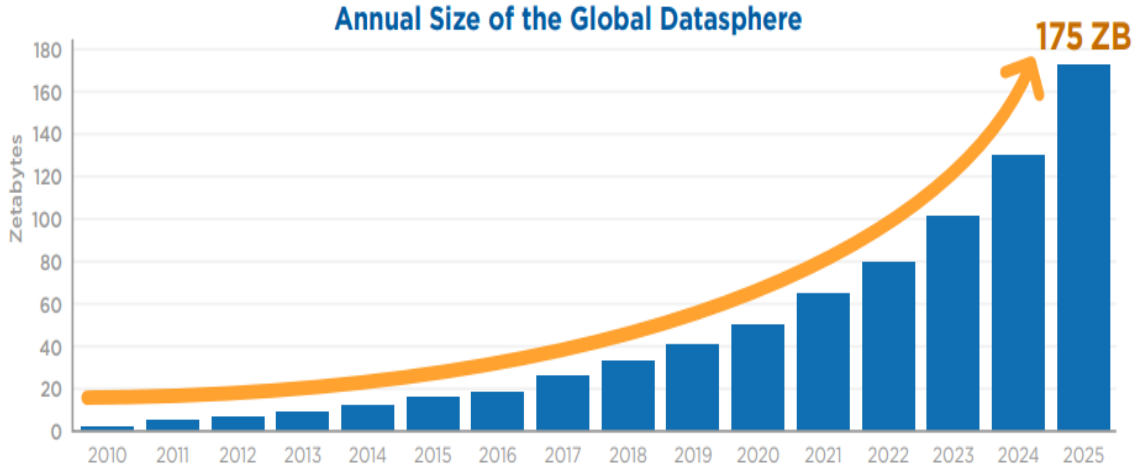
Büyük verinin sağladığı avantajların ve kolaylıkların yanında bazı olumsuz yanları da bulunmaktadır. Önceleri kilobayt ile ölçülen veri artık exabayt ve zetabayt ile ölçülür duruma gelmiştir. Aradaki büyüklük farkının anlaşılması için bazı birimler ve değerleri Tablo 2.2’de gösterilmiştir.

TABLO 2.2. Bayt birimleri ve deęerleri

Sembol	İfade	Bayt Deęeri	İkili Karşılıklar
KB	Kilobayt	10^3	$2^{10}=1024^1$
MB	Megabayt	10^6	$2^{20}=1024^2$
GB	Gigabayt	10^9	$2^{30}=1024^3$
TB	Terabayt	10^{12}	$2^{40}=1024^4$
PB	Petabayt	10^{15}	$2^{50}=1024^5$
EB	Exabayt	10^{18}	$2^{60}=1024^6$
ZB	Zettabayt	10^{21}	$2^{70}=1024^7$
YB	Yottabayt	10^{24}	$2^{80}=1024^8$

Tablodan da görüleceęi gibi exabayt ve zettabayt gibi deęerlerin yanında kilobayt çok küçük kalmaktadır. Bu kadar fazla miktardaki veri yığınının depolanması geleneksel yöntemler ile mümkün deęildir. Verinin hacminin yanında farklı tipten üretilmesi de bu depolamaya engel teşkil etmektedir. Depolanamama durumu büyük verinin olumsuzluklarından bir tanesidir. Bir başka zorluk ise bu büyük veri yığınının anlamlı verilerin çıkartılmasının zor olmasıdır. Önceleri az ve güvenilir kaynaklardan elde edilen veriler vardı. Günümüzde ise farklı farklı kaynaklardan doğruluęu kesin olmayan veriler üretilmektedir. Bu veri havuzundan kurum ve kuruluşların uygun verileri çekmesi analiz problemlerini de beraberinde getirmektedir. Ancak geliştirilen büyük veri araçları sayesinde depolama, analiz gibi işlemlere çözümler sağlanmıştır ve bu olumsuzluklar ortadan kaldırılmıştır.

IDC (International Data Corporation)'nin yaptığı bir araştırmaya göre 2025 yılında üretilcek veri miktarının 175 ZB olacağı öngörülmektedir. Şekil 2.2'de farklı yıllarda üretilmesi beklenen veri miktarları da görülmektedir.



Şekil 2.2. 2025 tarihine kadar üretilmesi öngörülen veri miktarları

Şekil incelendiğinde her yıl veri miktarının daha da katlanacağı tahmin edildiği görülmektedir. Bu durum kaçınılmaz bir gerçektir. Teknoloji geliştikçe ve teknolojiyi kullanan insan sayısı çoğaldıkça üretilen veri miktarı elbette bir önceki yıla göre artacaktır. Büyük veri probleminin ortaya çıkmasını önemli kılan olayları Sociomantic Labs “The Big Data Bang” isimli bir çalışmada ortaya koymuştur. Bu önemli olaylar ise şöyle sıralanabilir[14]:

- Amerika Birleşik Devletleri Savunma Bakanlığı’nın ileri düzeydeki araştırma projeleri birimi için 1969 yılında ilk bilgisayar ağı kuruldu.
- Tim Berners-Lee tarafından CERN’de (Avrupa Nükleer Araştırma Merkezi) geliştirilen World Wide Web yani www (Dünya Çapında Ağ), 1991 yılında insanlığın hizmetine sunuldu.
- HotWired 1994 yılında ilk banner reklamını yayınladı. Banner reklam, web sitelerine belirli bir ücret karşılığında yerleştirilen reklam görselidir.
- Yahoo! “golf” kelimesi ile ilk anahtar kelime reklamını 1995 yılında yaptı.
- 1996 yılında Double Click, ilk reklam teknoloji servislerinden birini Amerika’da sundu.
- 1997 yılında Finlandiyalı bir haber sağlayıcısı tarafından SMS yoluyla ücretsiz haber başlığı sunan ilk mobil reklam kullanıldı.
- 2000 yılında Google, Google Adwords’ü sundu. Google Adwords, kelime sorguları baz alınarak çalışan bir reklamcılık türüdür.
- 2004’te Amerika’da toplam internet reklamı harcamaları, 9.6 milyar dolara ulaştı.
- 2005 yılında dünya çapında internet kullanıcı sayısı 1 milyara ulaştı.
- 2007 yılında Facebook, kullanıcılarının sosyal etkileşimleri ve demografik bilgilerini baz alan ve davranışa göre hedeflenen reklam uygulamasını sundu.
- 2008’de Youtube platformu videolara reklam eklemeye başladı.
- 2009 yılında kullanıcıların ilgi alanlarına göre doğru zamanda reklam gösterimi yapan Real Time Bidding, Jason Knapp tarafından sunuldu.
- 2010 yılında Twitter reklam tweetleri kullanmaya başladı.
- 2011’de internet 1 trilyon üzerinde sayfa sayısına ulaştı.

- 2012’de Facebook, yeniden hedefleme (retargeting) tekniğini kullanan Facebook Exchange (FBX) uygulamasını piyasaya sundu. Bu uygulama diğer sitelerdeki aktivitelere göre reklamlar sunulmasına dayanır.
- 2012’de ilk defa internet reklam gelirleri, televizyon reklam gelirlerini aştı.
- 2014’te Pinterest, reklam pinlerini uygulamaya soktu.

Bu olaylar ışığında üretilen veri miktarının 5V özelliklerini sağlaması ile büyük veri ile karşı karşıya gelindi. Büyük veri son zamanlarda popülerleşen ama uzun yıllardır hayatımızda olan bir kavramdır. Büyük veri ve onun avantajlarını kullanan şirket, kurum ve kuruluşlar isimlerini ve markalarını büyütüp birer dünya devi haline gelmişlerdir. Büyük veri kullanan şirket sayısı yıllar geçtikçe daha da fazlalaşmıştır. Bu durumdan şirketlerin şu anda ve ilerleyen zamanlarda büyük veri sayesinde değerlerine değer katacakları sonucunu çıkardıkları görülür. Büyük veriyi etkin olarak kullanan 10 şirket şöyle listelenebilir[15]:

- Amazon
- Apple
- Google
- Spotify
- Facebook
- Instagram
- Starbucks
- Netflix
- McDonald’s
- LinkedIn

Bu şirketlerin büyük veriyi nasıl etkin kullandıkları hakkında çeşitli örnekler verilebilir. Netflix farklı kullanıcılardan elde ettiği verileri analiz ederek aynı film, dizi zevkine sahip insanlara analiz ettiği verilere uygun olarak çeşitli önerilerde bulunur. Bu şekilde kullanıcılarını iyi tanıyarak piyasadaki yerini korur. Spotify da Netflix ile aynı büyük veri avantajını kullanır. Kullanıcı verilerini iyi analiz ederek kullanıcılarına sevebilecekleri şarkı önerilerinde bulunur, onlara özel çalma listeleri hazırlar. Bir diğer örnek olarak Amazon büyük verinin önemini fark eden şirketlerdendir. Öyle ki sitesinde büyük veri kullanarak kullanıcı deneyimini iyileştirmesinin yanında büyük verinin etkin kullanılması için Amazon Web Services (AWS) platformunu kurmuştur. Birçok kurum ve kuruluş bu platformu kullanarak imkanlarından yararlanıyor.

Farklı kaynaklardan elde edilen verilerin aynı türde olmadıklarından ilgili bölümde bahsedildi. Önceleri deney, gözlem, araştırma sonucunda elde edilen güvenilir ve yapısal veriler vardı. Günümüzde ise veriler; yapısal, yapısal olmayan ve yarı yapısal olarak nitelendiriliyor. Bu verileri önceden kullanılan yöntemler ile analiz etmek, depolamak, paylaşmak ve aktarmak mümkün değildir. Yapısal verilerin içeriği bilindiğinden ilişkisel veri tablolarında tutulurdu ve analiz edilebilirdi. Şu an üretilen gerçek zamanlı verilerin içeriği bilinmediğinden uygun bir kullanım durumu sağlayamazlar. Bu olumsuzluklar Büyük Veri Araç ve Teknolojileri için çalışma yapılmasına yol açmıştır. Bu durum sonucunda ise Büyük Veri Araçları olarak isimlendirilen teknolojiler geliştirilmiştir. Bu teknolojiler sayesinde her türden veriyi toplamak, depolamak, analiz etmek, verilere erişimi sağlamak, verilerin güvenliğini sağlamak olanaklı hale gelmiştir. Bir sonraki bölüm olan Büyük Veri Araçları bölümünde bu teknolojilerden bahsedilecektir.

3. BÜYÜK VERİ ARAÇLARI



Şekil 3.1. Büyük veri nerelerden elde edilir temalı görsel

Büyük verinin olumsuzluklarından olan geleneksel yöntemlerle depolama, işleme, aktarma gibi işlemlerin yapılamaması bu işlemlerin yapılabilmesine olanak sağlayan teknolojilerin geliştirilmesi yolunu açtı. Bu durumun sonucunda çeşitli teknolojiler geliştirildi. Bu işlemlerin yapılamaması kurum, kuruluş ve şirketleri olumsuz etkiler. Bu işlemler rekabet gücünü artıran işlemlerdir. Günümüzde veri odaklı pazarlama yapılmaktadır ve kurumların hedeflerini büyük veri belirler. Bu nedenle büyük veri teknolojilerini kullanan şirketler piyasada rakiplerini başarı olarak geride bırakır.

Rekabetin yanında büyük veri günlük hayatı da kolaylaştırır. Büyük veri sayesinde sağlık sektöründe önceki hasta kayıtlarından ve teşhis sonuçlarından yararlanarak ilgili hastaya hızlı ve doğru teşhisler konulur. Aynı zamanda ilaç sektöründe de çeşitli iyileştirmeler yapar. İlgili hastalığa uygun ilaç geliştirilebilir. Ulaşım sektöründe gerçek zamanlı veri analizi sayesinde hızlı bir seyahat planlaması yapılabilir. Böylece insanlar zamandan tasarruf ederek kazandıkları vakitleri daha yararlı işler yapmak için harcayabilir.

Günümüze kadar çok fazla büyük veri aracı geliştirildi. Bazı teknolojiler popülerken bazıları popülerliğini yitirdi ya da kullanışlı olmadı. Burada dikkat edilmesi gereken nokta büyük veri araçları için her gün bir iyileştirme ve geliştirme yapılmasıdır.

Bu bölümde incelenmesi planlanan büyük veri araçlarının listesi şu şekildedir:

- Apache Hadoop
- Apache Spark

- Apache Hive
- MongoDB
- ElasticSearch
- Apache Kafka ve Apache ZooKeeper
- Apache Cassandra

3.1. Apache Hadoop



ŞEKİL 3.1.1. Apache Hadoop logosu

Apache Hadoop; dağıtık depolama ve dağıtık işleme imkanları sağlayan, Java diliyle yazılmış, açık kaynak kodlu bir projedir[16]. Açık kaynak kodlu olması sebebiyle ücretsiz sunulur ve kullanıcıların değiştirmesine, geliştirmesine imkan sağlanır. Böylece hem kullanıcılar projeyi kendi yararları için kullanmış olur hem de projenin patent sahibi ürününe yapılan değişiklikler sayesinde ürününü geliştirmiş olur. Apache Hadoop'un dağıtık depolama ve dağıtık işleme mekanizmaları düşük maliyet ile iş yapılmasına olanak sağlar. Dağıtık depolamada veri kaybı önlenir. Herhangi bir makinede veri kaybı durumu yaşandığında dağıtık depolama sayesinde yedeği alınan veriler kullanılır.

Apache Hadoop sayesinde büyük veri kümeleri gerçek zamanlı olarak, birden fazla makinede paralel olarak işlenir. Makinelerin birbirlerine paralel bağlanmasına cluster denilir. Cluster'larda veriler bloklara ayrılarak saklanır aynı zamanda bloklara ayrılarak işlenir[16]. Veri işleme hızı sayesinde de zamandan tasarruf sağlanır. Aynı zamanda farklı tipteki verilerle işlem yapılmasına olanak sağlanır. Bu sayede büyük veri probleminde ortaya çıkan geleneksel veritabanlarında depolanamama ve işlenememe sorunu aşılmış olur.

Apache Hadoop'un 4 temel bileşeni vardır. Bu bileşenler şu şekildedir[17]:

- Hadoop Common
- HDFS
- MapReduce
- YARN

Hadoop Common; diğer bileşenlere yardımcı programlar ve kitaplıklar sağlayan, aynı zamanda bu bileşenler arasındaki iletişimi sağlayan bir modüldür. Örnek olarak diğer modüllere Java kitaplığı desteği sağlar. Bu kitaplıkta çeşitli dosyalar ve komutlar bulunur. Bu sayede büyük verinin işlenmesine yardımcı olur[18].

HDFS(Hadoop Distributed File System), verilerin dağıtık şekilde saklandığı bir dosya depolama sistemidir. Verileri bloklara ayırarak saklar. Aynı zamanda bu verilerin kopyalarını oluşturup onları da bu bloklarda tutar. Verilerin kaç kere kopyalanacağı bilgisi replikasyon faktöründe (replication factor) tutulur[16]. HDFS'nin çalışma prensibi Tablo 3.1.1'deki gibi detaylı olarak açıklanabilir.

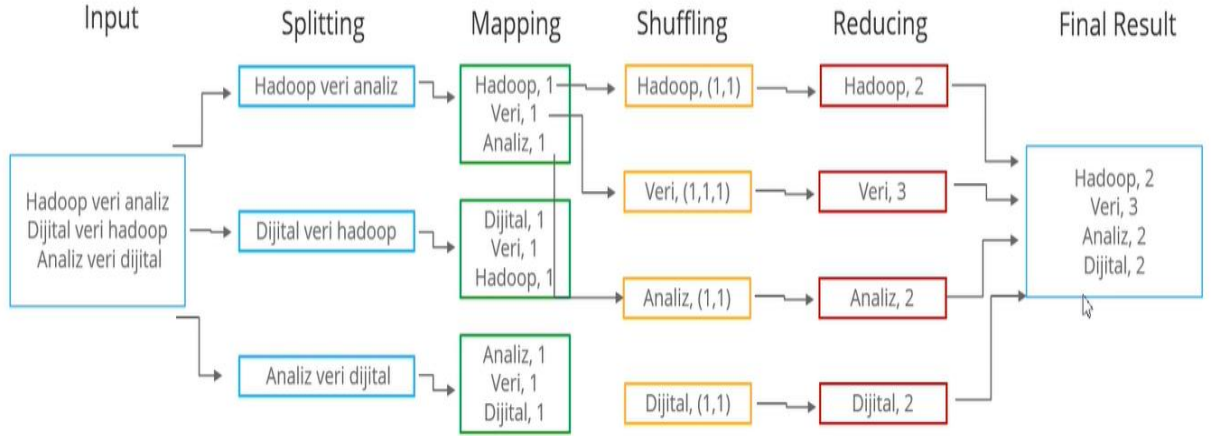
TABLO 3.1.1. HDFS çalışma prensibi

A	Makine 1	Makine 2	Makine 3	Makine 4	Makine 5
B	A	B	E	C	D
C	B	A	B	E	C
D	C	D	A	D	E
E					

İlk tabloda A, B, C, D ve E verilerinin Apache Hadoop tarafından depolanması istenmektedir. Replikasyon faktörünün 3 olması sebebiyle 5 farklı makinede A, B, C, D ve E verileri varsayılan 3 kopya ile bloklara ayrılarak saklanır. Örneğin 4 numaralı makinede bir problemle karşılaşıldığında bu makinede saklanan B, C ve E verileri 1.makine ve 3.makineden elde edilebilir. Farklı kombinasyonlar ile bu verileri farklı makinelerden elde etmek mümkündür. Bu sayede veri kaybı önlenir.

MapReduce büyük verilerin paralel olarak işlenmesini sağlayan modüldür. Verilerin paralel olarak işlenmesi sayesinde veriler aynı anda ve daha hızlı olarak işlenir. MapReduce modülü ismini aldığı Map ve Reduce fonksiyonlarından oluşur. Map fonksiyonu ile büyük veriler filtrelendir, Reduce fonksiyonu ile filtrelenen veriler analiz edilir ve bu analiz edilen verilerden sonuç çıkartılır. MapReduce fonksiyonu 6 aşamalık bir çalışma prensibine sahiptir. Bu aşamalar ile ilgili örnek Şekil 3.1.2'de verilmiştir[16].

MapReduce Kelime Sayma İşlemi



ŞEKİL 3.1.2. MapReduce kelime sayma işlemi

Şekil incelendiğinde input kısmından veri girişi yapılır. Splitting aşamasında verilerin daha hızlı ve kolay analiz edilebilmesi için bloklara ayrılır. Mapping aşamasında bloklara ayrılan veriler için istenen görev yerine getirilir. Burada işlem kelime sayma işlemi olduğu için bu aşamada hangi kelimedenden kaç tane olduğu bilgisi tutulur. Shuffling aşamasında bloklara ayrılan verilerin hangi bloklarda olduğu bilgisi tutulur. Örneğin “veri” kelimesi splitting aşamasında belirlenen 3 blokta da olduğu için (1,1,1) şeklinde gösterilmiştir. Bir başka örnek olarak “analiz” kelimesi 1. ve 3. blokta olduğu için (1,1) şeklinde gösterilmiştir. Reducing aşamasında ise sayım yapılır. Yine aynı örnekten “veri” kelimesi 3 bloğun her birinde geçtiği için 3, “analiz” kelimesi 2 blokta geçtiği için 2 olarak sayılır. Kısaca bu aşamada sonuçların birleştirildiği söylenebilir. Final Result aşamasında ise sonuç dosyası elde edilir ve istenilen işlem sonucu görülür.

YARN(Yet Another Resource Negotiator) kaynak yönetim platformudur[17]. Hangi uygulamanın ne kadar kaynak kullanacağını belirleyerek maksimum verimde bir çalışma düzeni oluşturur. Büyük verilerde kaynak yönetiminin yanında verilerle iletişimi sağlama, iş zamanlama, uygulama kullanımı ve yönetimi görevleri de vardır.

Bu şekilde 4 modülden oluşan Apache Hadoop’un büyük firmalar tarafından tercih edilmesindeki sebepleri şöyle sıralanabilir[16, 17]:

- Dağıtık işleme mekanizması ile veriler hızlı bir şekilde ve aynı anda işlenir bu sayede zamandan tasarruf edilir.
- Farklı tipteki verilerin depolanmasına ve işlenmesine olanak sağlar.
- Cluster’daki herhangi bir makinede sorun ortaya çıktığında diğer makineler tarafından iş akışı devam ettirilir. Bu sayede işler aksamaz.
- Pahalı donanımlara harcama yapmadan güvenilir ve devasa depolama fırsatı sağlar.
- Verilerin birden çok kopyası dağıtık şekilde depolanır böylece veri kaybı önlenir.
- Gerçek zamanlı analizleri sayesinde kullanıcıların ihtiyaçlarına cevap verir.
- Açık kaynaklı olması sayesinde geliştirilme imkanı yüksektir ve kullanımı ücretsizdir. Mevcut cluster’a yeni makine eklemek düşük maliyet gerektirir.

- Verileri işleme zorunluluğu olmadığından farklı türdeki veriler belli bir süre kısıtı olmadan depolanır.

Apache Hadoop ekosisteminde birçok araç bulunmaktadır. Bu araçlardan bazıları değişen şartlara uyum sağlayamadığı için kullanılmıyor bazıları ise sıklıkla tercih ediliyor. Bu araçlar şöyle sıralanabilir[17, 19]:

- Apache Pig
- Apache Hive
- Apache Hbase
- Apache Spark
- Apache ZooKeeper
- Apache Phoenix
- Apache Oozie
- Apache Storm
- Apache Flume
- Apache Sqoop

3.2. Apache Spark



ŞEKİL 3.2.1. Apache Spark logosu

Apache Spark büyük verileri dağıtık işlem ile analiz etmeyi sağlayan Scala dili ile geliştirilmiş bir büyük veri kütüphanesidir[16]. Apache Hadoop gibi açık kaynaklıdır. Dahili depolama birimi yoktur ve in-memory çalışır. Yani verileri önce RAM'e alır ve verilerin analizini RAM üzerinde gerçekleştirir. In-memory prensibi sayesinde Apache Hadoop'un MapReduce modülünden kat kat hızlı çalışır. Ekosistemindeki Spark Streaming sayesinde verileri anlık olarak analiz etmesi tercih edilmesindeki en büyük etkenlerden biridir. Her tipte ve boyuttaki verilerin analizinde oldukça etkilidir.

Apache Spark ve Apache Hadoop karşılaştırmaları çok fazla yapılır. Apache Spark Apache Hadoop'taki MapReduce modülünde hızlı iyileştirmeler yapılarak geliştirilmiştir. Apache Hadoop'ta veriler HDFS ile depolanıp MapReduce ile analiz edilir. Bu nedenle verilerin okunup işlenmesi daha yavaştır. Bu durum Apache Spark'ta dahili depolama biriminin olmaması ve tekrar eden işlem sayısının azaltılması ile iyileştirilmiştir[20].

Apache Spark'ta verilerin RAM üzerinde tutulmasını ve dağıtılmasını sağlayan yapıya RDD(Resilient Distributed Datasets/Esnek Dağıtılmış Veri Kümesi) denir[21]. Herhangi bir tipteki veri RDD'ye dönüştürülür. Analiz yapılır ve analiz sonuçlarından yeni RDD oluşturulur.

Apache Spark'ın farklı özelliklere sahip modülleri vardır. Bu modüller şöyle sıralanabilir[16]:

- Spark SQL
- Spark Streaming
- Spark MLlib
- GraphX

Spark SQL, SQL komutları ile kolay bir şekilde analiz yapmayı sağlar. Yapısal verilerin ya da yapısal veri haline dönüştürülmüş verilerin analizini SQL sorguları kullanarak yapar[21].

Spark Streaming, verileri anlık olarak analiz etme olanağı sağlar. En popüler modüldür.

Spark MLlib(Machine Learning Library), makine öğrenmesi modülüdür[22]. Makine öğrenmesi algoritmaları ve yardımcı programları içeren bir kitaplardır.

GraphX, graflar yardımıyla analiz yapılmasını sağlayan modüldür[21].

Apache Spark'ın avantajları şöyle sıralanabilir[16]:

- Veri işleme RAM üzerinde gerçekleştiğinden veri analiz hızı oldukça fazladır.
- Farklı diller kullanılarak veriler işlenir.
- Açık kaynaklı olması sayesinde farklı kullanıcılar tarafından geliştirilir.
- Verileri anlık analiz etme imkanı sağlar.
- Büyük bir ekosistemi vardır ve bu ekosistem sayesinde sorunlara uygun çözümler bulur.
- Verinin türünden bağımsız olarak analiz yapar.

3.3. Apache Hive



ŞEKİL 3.3.1. Apache Hive logosu

Apache Hive, Facebook tarafından Hadoop ekosisteminin bir alt modülü olarak geliştirilmiş açık kaynak bir kütüphanedir[16]. Açık kaynak olması sayesinde başta Facebook tarafından geliştirilse de daha sonradan farklı büyük şirketler tarafından kullanılmış ve çeşitli iyileştirmeler yapılmıştır. Hadoop alt modülü olması sayesinde HDFS’de depolanan veriler MapReduce ile analiz edilir. Aynı zamanda Apache Hive şeması ile veriler HDFS’de saklanır.

Apache Hive platformunun amacı büyük verileri SQL (Structured Query Language) benzeri sorgulama dili olan HiveQL (Hive Query Language) sorguları ile analiz etmektir[23]. SQL’de yapılandırılmış verilerin tutulmasına karşın Apache Hive’da yapılandırılmış ve yapılandırılmamış veriler desteklenir. Yine aynı şekilde Hive kodlama adımlarında SQL’de olduğu gibi veritabanı ve tablo oluşturma adımları bulunur. Tablo kolonlarına isim verilir ve veri tipleri belirlenir. Daha sonrasında HiveQL sorgusu yazılır. Apache Hive’da şema bilgilerinin, tabloların, kolonların ve kolon tiplerinin tutulduğu bölüme Hive MetaStore denir[23].

Apache Hive platformunun avantajları şöyle sıralanabilir[23]:

- Yüksek bayttaki veriler hızlı bir şekilde işlenir.
- Açık kaynaklı olması sayesinde farklı şirketler tarafından amaçlarına daha iyi hizmet edecek şekilde geliştirilir ve kullanıma sunulur.
- Desteklenen programlama dillerini (Scala, Java, Python) bilmeyen kişilerin bile basit SQL’e benzer HiveQL ile işlem yapması sağlanır.
- Farklı tipteki verileri destekler.
- Veriler HDFS ile depolandığından normal veritabanlarından kat kat fazla veri tutulur.

3.4. MongoDB



ŞEKİL 3.4.1. MongoDB logosu

MongoDB açık kaynaklı bir NoSQL (Not Only SQL) teknolojisidir[16]. NoSQL teknolojileri içinde en popüler olanıdır. MongoDB konusunda devam etmeden önce NoSQL kavramından bahsetmek daha faydalıdır.

NoSQL ilişkisel veritabanı sistemlerine alternatif olarak ortaya çıkan sistemlerdir[24]. İlişkisel olmayan veritabanlarını sorgulama dilidir. NoSQL büyük veri tarafında yaygın olarak kullanılır çünkü büyük verileri yönetmede kolaylık sağlar.

Geleneksel veritabanlarındaki kolonlarda değer ataması olmasa bile NULL değeri atanması durumu, herhangi bir sorgulama yapıldığında bu değerlerin de üzerinden geçildiği için işlem hızını yavaşlatan faktörlerden biridir[16]. Aynı zamanda bu atama gereksiz yer kaybına neden olmaktadır. NoSQL’de ise bu durumun önüne geçilerek işlem hızı artırılmış, zamandan tasarruf sağlanmıştır. Bir başka NoSQL avantajı ise geleneksel veritabanlarında farklı tablolar arası ilişkilerin kurulmasındaki sıkıntıların giderilmesidir. Geleneksel veritabanlarında proje büyüdükçe tablo sayısı da artar ve bu durum tablolar arası ilişkilerin kurulmasında uzmanlık ve zaman gerektiren bir problem haline gelir[16]. NoSQL veritabanı ilişkisel veritabanı olmadığından bu gibi dezavantajların önüne geçen bir çözüm olmuştur ve sayesinde çok kısa sürelerde istenilen sorgu sonucu getirilir.

MongoDB verileri JSON (JavaScript Object Notation) benzeri BSON (Binary JSON) formatında saklar[16]. JSON formatı verileri anahtar-değer ilişkisi içinde saklar. Farklı platformlar arasındaki veri akışında oldukça kullanışlı olduğundan sıklıkla tercih edilir. JSON formatına Şekil 3.4.2 örnek olarak verilebilir.

```
( ) deneme.json > ...  
1  {  
2    "name": "Aysenur",  
3    "surname": "Capkan",  
4    "age": "21"  
5  }  
6
```

ŞEKİL 3.4.2. JSON örneği

Şekil 3.4.2 incelendiğinde anahtar-değer ilişkisi daha net bir şekilde görülür. Farklı kullanıcılar eklendiğinde ve örnek olarak bir kullanıcı için yaş bilgisi eklenmediğinde JSON formatında bu anahtar için değer bilgisi beklenmeyecektir ya da NULL değer ataması yapılmayacaktır. Böylece hem depolamadan tasarruf sağlanacak hem de herhangi bir sorgu yapıldığında olabilecek gereksiz aramalar önlenecektir.

BSON formatı ise JSON formatında tutulan verilerin binary formatta tutulmasıdır. BSON, JSON'a göre daha hızlıdır ve daha fazla veri tipini destekler[25].

MongoDB'de veriler belge şeklinde saklanır yani MongoDB document-oriented (belge yönelimli) özelliğine sahiptir[26]. Belge şeklinde saklama çeşitli avantajlar sunduğundan tercih edilir. Bu avantajlar içinde en öne çıkanı esnek bir şema kullanımıdır.

MongoDB'nin avantajları şöyle sıralanabilir[16]:

- Kullanılan veri boyutuna göre makine ekleme çıkarma işlemi kolaydır ve düşük maliyetlidir.
- Verilerin BSON formatında tutulması günümüzdeki web teknolojilerinde veri aktarımı için çok uygundur.
- Verilerin birden fazla kopyasının saklanması veri kaybını önler.
- İlişkisel olmayan veritabanları kullandığından işlem hızı yönünden öndedir.

3.5. Elasticsearch




ŞEKİL 3.5.1. Elasticsearch logosu

ElasticSearch büyük veri teknolojileri arasında arama işlemlerinde kullanılan, Java ile yazılmış açık kaynaklı bir NoSQL teknolojisidir[16]. Arama işlemi açısından büyük veri araçları arasında en popüler olanıdır. Genellikle metin arama işleminde kullanılır. Çeşitli web sitelerine ve uygulamalara entegre edilir. Web sitelerinde kullanımı durumuna örnek olarak “Iphone” kelimesi yazıldığında milisaniyeler içinde Iphone 12, Iphone 13 gibi önerilerin çıkması verilebilir.

ElasticSearch teknolojisinin bu kadar hızlı çalışmasının sebebi verileri daha veritabanlarına kaydederken indekslemesidir. ElasticSearch çalışma prensibi 3.5.1’deki tablo örneği ile daha detaylı anlatılmıştır.

TABLO 3.5.1. ElasticSearch çalışma prensibi

1	Bugün hava çok güzel.
2	Hava yarın bulutlu olacak.
3	Bugün yağmur yağabilir.



bugün	<1>,<3>
hava	<1>,<2>
çok	<1>
güzel	<1>
yarın	<2>
bulutlu	<2>
olacak	<2>
yağmur	<3>
yağabilir	<3>

Tablo 3.5.1’de 4 farklı cümlelerin (ElasticSearch’te doküman olarak isimlendirilir) ElasticSearch veritabanlarına kaydedilme işlemi gösterilmiştir. Önce her dokümandaki kelime verileri ayrılır. Ayırma işleminde boşluklar varsayılan olarak belirlenmiştir ve kelimeler boşluklardan ayrılır. Ayrılan kelimeler indeks listesine alınır ve hangi dokümanda geçtikleri bu indeks listesine eklenir. Örnek olarak “bugün” kelimesi hem 1. hem de 3. dokümanda yer aldığından bu 2 doküman numarası indeks listesinde kelimenin yanına yazılır. Başka bir örnek “yarın” kelimesi için verilebilir. Yarın kelimesi sadece 2. dokümanda geçtiği için indeks listesinde yarın kelimesinin yanına doküman numarası olan 2 yazılır.

Arama yapıldığı zaman sorguya bugün yazıldığında indeks listesi taranır ve bugün kelimesinin 1. ve 3. dokümanda geçtiği hızlı bir şekilde görülür. Sorgu sonucunda da bugün hava ve bugün yağmur çıktıları sunulur. Varsayılan olarak boşluklarına göre ayırma durumu farklı bir arama istendiğinde değiştirilebilir. Boşluklarına göre ayırma yerine “_” görünce ayır gibi farklı bir kullanım tercih edilebilir.

Şekilde de incelenilen çalışma prensibi sayesinde bir arama yapıldığı zaman bütün dokümanları dolaşmak yerine arama işlemi indeksler arasında hızlı bir şekilde yapılıyor ve yine aynı hızda sonuçlar yansıtılıyor.

ElasticSearch teknolojisinin avantajları şöyle sıralanabilir[16]:

- Arama işleminin önemli olduğu durumlarda işlem hızı sayesinde tercih edilir.
- Diğer büyük veri teknolojilerinde olduğu gibi cluster yapısı vardır. Bu sayede veriler kopyalanır ve veri kaybı önlenir.
- Ölçeklenebilir olması sayesinde veri miktarına göre kolaylıkla makine eklenebilir veya çıkarılabilir.
- Birçok dil desteği sayesinde proje geliştirme kolaylığı sağlar.

3.6. Apache Kafka ve Apache ZooKeeper



ŞEKİL 3.6.1. Apache Kafka logosu

Apache Kafka hakkında tanım yapılmadan önce Apache ZooKeeper hakkında ufak bir bilgilendirme yapılacaktır. Bu bilgilendirme sayesinde dağıtık işlem mantığı daha iyi bir temele oturtulacaktır.



ŞEKİL 3.6.2. Apache ZooKeeper logosu

Apache ZooKeeper dağıtık işlem yapan teknolojilerde koordinasyon sağlayan açık kaynaklı bir teknolojidir[16]. Kaynak yönetimi, iş koordinasyonu gibi görevleri vardır.

Sağladığı merkezi yönetim ile büyük veri araçlarının işlerini kolaylaştırır[27]. Dağıtık depolama ve analiz gibi işlemler yapmaları sebebiyle Apache projeleri tarafından sıklıkla kullanılır.

Apache Kafka sosyal medya, sensörler, veritabanları ve çeşitli uygulamalar gibi farklı kaynaklardan gelen verileri toplar[16]. Daha sonra işlem yapacak teknolojilere verileri aktarır. Verilerin direkt olarak kaynaktan analiz, arama gibi işlemler yapacak araçlara aktarılmasında veri kaybı yaşanabilir. İlgili araç o anda yaşanan bir problem sonucu çalışmıyorsa veriler kaybedilmiş olur. Bu gibi durumların önlenmesi için Apache Kafka veri kaynağı ve işlem yapacak araç arasında köprü görevi görür[16]. Veri kaybını önlediği için gerçek zamanlı veri kullanımı gerektiren projelerde kullanılır.

Apache Kafka verileri hatasız ve hızlı bir şekilde kuyruk (query) yapısı ile toplar[16]. Kuyruk yapısı FIFO (First In First Out) mantığı ile çalışır. Bu mantık neticesinde Apache Kafka'ya gelen veriler kuyrukta en sona yazılır. Analiz araçları veri istediği zaman ise ilk gelen veriler aktarılır.

Apache Kafka'nın çalışma prensibinde de Cluster yaklaşımı vardır. Veri toplayan makinelere Broker adı verilir. Broker'ların birbirleriyle paralel bağlanma ilişkisine Cluster denir. Bu şekilde veriler dağıtık bir şekilde depolanır. Apache Zookeeper bu noktada koordinasyonu sağlamak için devreye girer. Apache Kafka ile Broker'lar arasındaki ilişkiyi Apache Zookeeper sağlar. Zookeeper'lar hem kendi aralarında hem de verilerin saklandığı Broker'lar arasındaki haberleşmeyi sağlar[16].

Veriler Broker'larda Topic'lere ayrılır[16]. Örneğin gelen verilerden "Araba" bilgisini içerenler araba ile ilgili oluşturulan Topic'te saklanır. İleride analiz aracından ilgili Topic istendiğinde bütün verileri aramak yerine daha önceden kümelenmiş Topic aktarılır. Broker'larda birden fazla Topic bulunabilir.

Verileri gönderen yapılara Producer, verilerin aktarılmasını isteyen yapılara Consumer denir. Producer'lardan alınan veriler Topic'ler içinde bulunan Partition'lara kuyruk yapısı ile depolanır. Partition sayısı uygulamayı kullanan kişilerin isteğine göre belirlenir[16].

Çalışma prensibi verilen Apache Kafka'nın sağladığı avantajlar şöyle sıralanabilir[16]:

- Veri kaybını önleyerek hızlı ve hatasız veri toplar.
- Gerçek zamanlı projelerde etkin kullanılır.
- Analiz araçlarda bir problem olsa bile verileri depolar ve ilgili araçlardaki problem çözüldüğünde veri aktarımına kaldığı yerden devam eder.
- Gönderilen veri bilgisini tuttuğu için aynı verilerin tekrar gönderilmesini önler.
- Replikasyon özelliği sayesinde veriler kopyalanır.
- Kaynaklardan alınacak verilere belli bir sınır getirebilir. Böylece istenilen ölçüde veri depolanır.

3.7. Apache Cassandra



ŞEKİL 3.7.1. Apache Cassandra logosu

Apache Cassandra Facebook tarafından Java dili ile geliştirilmiş, açık kaynaklı NoSQL (Not Only SQL) veritabanıdır[28]. Veriler dağınık bir şekilde düğümlerde tutulur. Bu sayede herhangi bir düğümden problem yaşandığında işlem zamanından kayıp yaşanmamaktadır. Bütün düğümler aynı görevlere sahiptir biri diğerinden üstün değildir[29]. Yani ana düğüm yoktur. Bu sayede bir problem meydana geldiğinde hissettirilmeden diğer düğümlerden işlemler aksamadan ilerler.

Apache Cassandra büyük hacimli verileri işlemek için tasarlanmıştır. Bu sayede okuma ve yazma hızı oldukça fazladır. Sorgular CQL (Cassandra Query Language) adı verilen SQL'e benzer sorgulama dili ile yapılır[30]. Dilin SQL'e benzer olması sayesinde geliştiriciler tarafından kolayca kullanılıp adapte olunabilir.

MongoDB'nin de NoSQL veritabanı olması ile Apache Cassandra ile karşılaştırması yapılabilir. Aradaki temel farklardan bazıları şöyle listelenebilir[30]:

- En temel fark sorgulama dillerinin farklı olmasıdır. MongoDB'de MQL (MongoDB Query Language) kullanılırken Apache Cassandra'da CQL (Cassandra Query Language) kullanılır.
- MongoDB verileri JSON (JavaScript Object Notation) benzeri BSON (Binary JSON) formatında saklarken Apache Cassandra verileri tablo deposunda depolar.
- Apache Cassandra ücretsiz olarak sunulurken MongoDB'de farklı kullanımlara özel ücretli paketleri vardır.
- Apache Cassandra'da mobil geliştirme özelliği yokken MongoDB'de vardır.
- Apache Cassandra'da veri görselleştirme özelliği yokken MongoDB'de vardır.

- Apache Cassandra’da bütün düğümlerin görevleri aynı olduğundan bir problem ile karşılaşıldığında sorun fark ettirilmeden işlemlere devam edilir. Ancak MongoDB’de lider düğümden çözüm beklendiği için işlem akışında ufak bir zaman kaybı yaşanır.
- Apache Cassandra’da MongoDB’ye istinaden daha az dil desteği vardır.

Apache Cassandra’nın kullanıcılarına sunduğu avantajlar şöyle sıralanabilir[31]:

- Açık kaynaklı olması sayesinde farklı kişiler tarafından geliştirilir ve paylaşılır.
- Ücretsiz sunulur.
- Farklı veri türlerini destekler.
- Dağıtık bir sisteme sahip olması sayesinde işlem akışında bir problem olduğunda aksamadan devam eder.
- Okuma ve yazma hızı normal veritabanlarına göre kat kat fazladır.
- Ölçeklenebilir olması sayesinde düğüm sayısı artırılabilir ya da azaltılabilir.
- Hataya dayanıklıdır.
- Çok yüksek veri hacimlerini işler.
- Sorgulama dilinin SQL’e benzer olması sayesinde kolaylıkla adapte olunur.
- Hiçbir başarısızlık noktası yoktur.

4. SONUÇLAR

Büyük Veri günümüz dünyasında değeri anlaşılan ve yatırım yapılan bir alandır. Bu alanda doğru yatırımlar yapan şirketler rakiplerini geride bırakacak güce sahip olur. Şirket bazında bakılmadan kariyer planları yapan bir bireyin de bu alanda kendine yapacağı yatırım ona güzel bir gelecek ve iş imkanıyla neticelenebilir. Büyük veri mühendisleri, veri mühendisleri sektörde aranan ancak bulunamayan mesleklerdendir. Bu nedenle eğer büyük veriyle ilgili araştırma yapma, veri toplama, analiz etme gibi işlemlere ilgi duyuluyorsa bu alanda bir kariyer hedefi kişinin hayatını değiştirecek türdendir. Büyük Veri ve Büyük Veri Araçları ile ilgili internette birçok eğitim videosu, makale, blog yazısı mevcut. Aynı zamanda bazı şirketler bu alanlarda çeşitli eğitimler düzenleyerek farkındalık yaratmaya çalışıyor.

Bu tezde de Büyük Veri hakkında hiçbir şey bilmeyen bir kişiye konuyu sıkmadan ve fazla teknik bir tavır izlemeden Büyük Veri anlatılmaya çalışılmıştır. İlgili tezde Büyük Veri'nin günlük hayattaki kullanım örneklerine sıklıkla yer verilerek aslında Büyük Veri'yle hep karşılaşıldığı gösterilmeye çalışılmıştır. Büyük Veri Araçları bölümünde Apache Hadoop, Apache Kafka, MongoDB, Apache Spark, Cassandra, Apache ZooKeeper, Elasticsearch, Apache Hive teknolojileri incelenmiş olup ilgili araçların çalışma prensipleri basit bir üslupla ve şekillerle anlatılmıştır. Bu teknolojilerin açık kaynaklı olması sebebiyle bir sorun ile karşılaşıldığında internette çözüm yolu bulma ihtimali çok yüksektir. Bu durum da bu teknolojilerle uğraşmak isteyen bir kişiye sağlanacak en büyük avantajlardan biridir.

KAYNAKLAR

- [1] <https://tr.wikipedia.org/wiki/Veri> (26.03.2022)
- [2] Aktan, E. (2018). Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu. *Dergipark*, 1/1, 3-7. <https://dergipark.org.tr/tr/download/article-file/482194>
- [3] <https://medium.com/dusunenbeyinler/big-data-b%C3%BCy%C3%BCk-veri-analizi-d53d8f8ab52b> (28.03.2022)
- [4] (2016). Büyük Veri Kavramı ve Karakteristik Özellikleri. XVIII. Akademik Bilişim Konferansı. <https://ab.org.tr/ab16/bildiri/66.pdf>
- [5] https://tinkdata.com/bigdata_buyuk-veri.html (03.04.2022)
- [6] <https://www.analiyz.com/adan-zye-buyuk-veri-nedir-big-data-klavuzu/> (28.03.2022)
- [7] <https://www.ticimax.com/blog/big-data-buyuk-veri-nedir-e-ticarete-kullanim-ornekleri> (28.03.2022)
- [8] <https://www.dailymotion.com/video/x3dtnii> (29.03.2022)
- [9] Terzi, R., Sağiroğlu, Ş., Demirezen, M. U. (2017). Büyük Veri Ve Açık Veri: Temel Kavramlar. <http://www.ttbilgin.com/2017-2018-bahar/MCH641/unite1.pdf>
- [10] <https://evrimagaci.org/big-data-nedir-buyuk-veri-yapay-zekanin-zincirlerini-kirmasini-saglayacak-anahtar-olabilir-mi-11347> (04.04.2022)
- [11] <https://tusiad.org/tr/fikir-ureten-fabrika/item/8348-buyuk-verinin-iki-yuzu> (29.03.2022)
- [12] <https://iskulubu.com/teknoloji/big-data/> (28.03.2022)
- [13] Odabaşı, H. F., Akkoyunlu, B. ve İşman, A. (Ed.). (2017). Eğitim Teknolojileri Okumaları 2017. TOJET. Sakarya https://www.researchgate.net/publication/318509130_Egitimde_Buyuk_Veri
- [14] Karaca, İ. (2015). Büyük Veri Analizlerinin Kurumsal Faaliyetlerde Kullanım Alanları. Lisans Tezi. Ankara Üniversitesi Bilgi ve Belge Yönetimi Bölümü <https://www.ismailkaraca.com.tr/wp-content/uploads/2015/06/B%C3%BCy%C3%BCk-Veri-Analizlerinin-Kurumsal-Faaliyetlerde-Kullan%C4%B1m-Alanlar%C4%B1-Lisans-Tezi-%C4%B0smail-Karaca.pdf>
- [15] <https://www.analyticssteps.com/blogs/companies-uses-big-data> (30.04.2022)
- [16] <https://www.btkakademi.gov.tr/portal/course/bueyuek-veriye-giris-5577> (28.05.2022)
- [17] <https://www.gtech.com.tr/hadoop-nedir/> (19.05.2022)
- [18] <https://tr.theastrologypage.com/hadoop-common> (18.05.2022)
- [19] <https://tr.ilusionity.com/2624-a-quick-overview-of-the-apache-hadoop-framework> (20.05.2022)
- [20] <https://www.elektrikport.com/teknik-kutuphane/buyuk-veri-islemek-icin-tasarlanmis-araclar-hadoop-ve-spark/22465#ad-image-0> (20.05.2022)

- [21] <https://www.gtech.com.tr/apache-spark/> (19.05.2022)
- [22] <https://spark.apache.org/mllib/> (28.05.2022)
- [23] <https://aws.amazon.com/tr/big-data/what-is-hive/> (20.05.2022)
- [24] <https://aws.amazon.com/tr/nosql/> (20.05.2022)
- [25] <https://medium.com/keove/json-vs-bson-73294d9c012d> (20.05.2022)
- [26] <https://tr.theastrologypage.com/document-oriented-database> (20.05.2022)
- [27] <https://www.datascienceearth.com/zookeeper-bolum-1-nedir-ne-degildir/> (21.05.2022)
- [28] <https://www.gtech.com.tr/apache-cassandraya-giris/> (21.05.2022)
- [29] <https://medium.com/codable/apache-cassandra-nedir-nas%C4%B1-kullan%C4%B1%C4%B1r-c058c2e3a687> (21.05.2022)
- [30] <https://www.mongodb.com/compare/cassandra-vs-mongodb> (21.05.2022)
- [31] https://cassandra.apache.org/_/index.html (21.05.2022)

ÖZGEÇMİŞ

Ad Soyad: Ayşe Nur ÇAPKAN
Doğum Tarihi: 10.07.2000
Doğum Yeri: Malatya
Lise: 2014 – 2018 Suat Terimer Anadolu Lisesi
Staj Yaptığı Yerler: CPM Yazılım -İstanbul- (3 ay)
Çalıştığı Yerler: KoçSistem -İstanbul- (4 ay)