PROCEEDINGS BOOK

7th ASIA PACIFIC International Modern Sciences Congress

November 4-5, 2022
Jakarta, Indonesia
STIE GANESHA College of Economics

Editors

Dr. Adhy FIRDAUS
Dr. Joned Ceylendra SAKSANA
Samira KHADHRAOUI ONTUNC

Institute Of Economic Development And Social Researches Publications®
(The Licence Number of Publicator: 2014/31220)
TURKEY

E posta: info@iksadkongre.org

Issued: November 25, 2022
ISBN - 978-625-8246-59-9

7th ASIA PACIFIC International Modern Sciences Congress

BÜYÜK VERİ ve HADOOP

Emine BAŞ
Lecturer Dr., Department of Software Engineering, Faculty of Engineering and Nature Sciences, Konya Technical University, 42075, Konya, Turkey
ORCID: 0000-0003-4322-6010

ÖZET

Günümüzde teknolojinin yaygın bir şekilde kullanılmasıyla artan bir very (büyük veri) oluşmuştur. Büyük veri, geleneksel veri işleme uygulamalarının üstesinden gelemeyeceği kadar büyük veya karmaşık veri setlerini analiz etme ve bu veri setlerinden sistematik olarak bilgi elde etmeyi sağlayacak yöntemler arayan bilişim bilimleri sahasıdır. Bir diğer deyişle Big Data, çoğunluğu yapılandırılmamış olan ve sonu gelmez bir şekilde birikmeye devam eden, geleneksel ilişki bazlı veri tabanı teknikleri yardımıyla çözülemeyecek kadar yapısallıktan uzak, çok çok büyük, çok ham ve üstel bir şekilde büyümekte olan veri setleridir. Büyük very çeşitlilik, hız ve hacim olmak üzere üç ana bileşeni ile karakterize edilen geleneksel veri analizinden devrim niteliğinde bir adım gerektirir. Bu verinin şekli itibariyle klasik yöntemlerle işlenmesi zordur. Çeşitlilik (*Variety*), büyük verileri gerçekten büyük hale getirir. Verinin hacmi veya boyutu *(Volume)* artık terabayt ve petabayttan daha büyüktür. Hız (*Velocity*) sadece büyük veri için değil, tüm süreçler için gereklidir. Zaman sınırlı süreçler için, değerini en üst düzeye çıkarmak için kuruluşa akarken büyük veri kullanılmalıdır. Verilerin büyük ölçeği ve yükselişi, geleneksel depolama ve analiz tekniklerini geride bırakır. Araştırmacılar bu verinin kolay bir şekilde işlenmesi için bir arayış içine girmiştir. Büyük veri, *MapReduce* gibi mimarileri destekleyen yepyeni bir endüstri yaratmıştır. Hadoop bu büyük verinin sınıflandırılması ve işlenmesi konusunda çıkmış bir yazılımdır. Hadoop JAVA programlama dili ile geliştirilmiş popüler, açık kaynaklı bir Apache projesidir. Üretilme amacı ise büyük verilerin daha hızlı işlenmesidir. Temel olarak yazılımı dağıtık dosya sistemi olarak tanımlayabiliriz. Bu dağıtık dosya sistemi HDFS yani Hadoop Distributed File System olarak adlandırılır. Hadoop bileşenleri şunlardır: HDFS, MapReduce, HBase, Pig, Hive ve ZooKeeper dir. Bu bildiride büyük veri ve hadoop konusunda bir araştırma sunulmuştur.

Anahtar Kelimeler: Hadoop, Büyük veri, Dağıtık sistemler

BIG DATA AND HADOOP

ABSTRACT

Today, with the widespread use of technology, an increasing amount of data (big data) has occurred. Big data is the field of information science that seeks methods to analyze and systematically extract information from data sets that are too large or complex to be handled by traditional data processing applications. In other words, Big Data is a very large, very raw and exponentially growing dataset, most of which is unstructured and continues to accumulate endlessly, too unstructured to be solved with the help of traditional relationship-based database techniques. Large data requires a revolutionary step forward from traditional data analysis, characterized by its three main components: diversity, speed, and volume. This data is difficult to process with classical methods due to its shape. Variety makes big data really big. The volume or size (Volume) of data is now larger than terabytes and petabytes. Velocity is essential for all processes, not just big data. For time-limited processes, big data should be used as it flows into the organization to maximize its

value. The massive scale and rise of data outstrips traditional storage and analysis techniques. Researchers have been in a quest for an easy processing of this data. Big data has created a whole new industry supporting architectures like MapReduce. Hadoop is a software for classification and processing of this big data. Hadoop is a popular, open source Apache project developed with the JAVA programming language. Its purpose is to process big data faster. Basically, we can define software as a distributed file system. This distributed file system is called HDFS or Hadoop Distributed File System. Hadoop components are HDFS, MapReduce, HBase, Pig, Hive and ZooKeeper. In this paper, a research on big data and hadoop is presented.

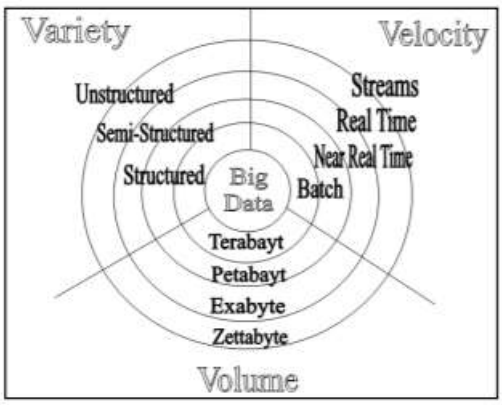
Keywords: Hadoop, Big data, Distributed systems

GİRİŞ

Big Data ("Büyük Veri"), geleneksel veri işleme uygulamalarının üstesinden gelemeyeceği kadar büyük veya karmaşık veri setlerini analiz etme ve bu veri setlerinden sistematik olarak bilgi elde etmeyi sağlayacak yöntemler arayan bilişim bilimleri sahasıdır. Bir diğer deyişle Big Data, çoğunluğu yapılandırılmamış olan ve sonu gelmez bir şekilde birikmeye devam eden, geleneksel ilişki bazlı veri tabanı teknikleri yardımıyla çözülemeyecek kadar yapısallıktan uzak, çok çok büyük, çok ham ve üstel bir şekilde büyümekte olan veri setleridir (https://evrimagaci.org). Başka bir deyişle büyük veri, daha sonraki süreçler veya sonuçlar için depolama, analiz etme ve görselleştirme zorluklarıyla birlikte büyük, daha çeşitli ve karmaşık yapıya sahip büyük veri kümeleri için kullanılan bir terimdir. Büyük veri analitiği olarak adlandırılan gizli kalıpları ve gizli korelasyonları ortaya çıkarmak için büyük miktarda veri üzerinde araştırma sürecidir. Bu yararlı bilgiler, şirketler veya kuruluşlar için daha zengin ve daha derin iç görüler edinme ve rekabette avantaj sağlama yardımı ile kullanılırlar. Bu nedenle, büyük veri uygulamalarının mümkün olduğunca doğru bir şekilde analiz edilmesi ve yürütülmesi gerekmektedir (Sagiroglu ve Sinanc, 2013).

BÜYÜK VERİ (BIG DATA)

Büyük Veri, Şekil 1'de gösterildiği gibi çeşitlilik, hız ve hacim olmak üzere üç ana bileşeni ile karakterize edilen geleneksel veri analizinden devrim niteliğinde bir adım gerektirir (Sagiroglu ve Sinanc, 2013; Gerhardt ve ark., 2012).



Şekil 1. Büyük very bileşenleri

Çeşitlilik (*Variety*), büyük verileri gerçekten büyük hale getirir. Büyük veri çok çeşitli kaynaklardan gelir ve genellikle üç türe sahiptir: yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış. Yapılandırılmış veriler, önceden etiketlenmiş ve kolayca sıralanan bir veri ambarı ekler, ancak

yapılandırılmamış veriler rastgeledir ve analiz edilmesi zordur. Yarı yapılandırılmış veriler sabit alanlara uymaz ancak veri öğelerini ayırmak için etiketler içerir (Eaton ve ark., 2012; Singh ve Singh, 2012).
Verinin hacmi veya boyutu *(Volume*) artık terabayt ve petabayttan daha büyüktür. Verilerin büyük ölçeği ve yükselişi, geleneksel depolama ve analiz tekniklerini geride bırakır (Eaton ve ark., 2012; Madden, 2012).
Hız (*Velocity*) sadece büyük veri için değil, tüm süreçler için gereklidir. Zaman sınırlı süreçler için, değerini en üst düzeye çıkarmak için kuruluşa akarken büyük veri kullanılmalıdır (Eaton ve ark., 2012; Madden, 2012).
Bu bilgilerin yoğunluğu sırasında bir diğer bileşen veri akışının doğrulanmasıdır. Büyük veriyi kontrol etmek zor olduğundan veri güvenliğinin sağlanması gerekir. Ayrıca büyük verinin üretilip işlenmesinden sonra kurum için artı değer yaratmalıdır.
Veri yönetimi uzmanlarına sorulan TDWI anketinden aşağıda özetlenen bazı sorular ve önemli cevaplar vardır (Russom, 2011).
Kuruluş bir tür büyük veri analitiği uyguladıktan sonra, şu faydalar ortaya çıkar: daha iyi hedeflenmiş pazarlama, daha doğru iş anlayışları, müşteri tabanlı segmentasyon, satışların ve pazar fırsatlarının tanınması.
Büyük veri analitiğini uygularken, bu sorunlar potansiyel engellerdir: uzman olmayan personel, maliyet, işletme sponsorluğundan yoksunluk, analitik sistemlerin tasarlanması zor, analitikte mevcut veritabanı yazılımının eksikliği.
Önemli bir kitle, büyük verileri şimdi ve gelecekte tanımlıyor, kapsamlı analitikler nedeniyle bir fırsat olsa da, bazıları yönetim nedeniyle büyük verileri sorun olarak görüyor.
Günümüzde ileri tekniklerle saklanan ve kullanılan büyük veri türleri: yapılandırılmış, yarı yapılandırılmış, karmaşık, olay ve yapılandırılmamış veridir.
Analitik platformları değiştirilirken şu sorunlar ortaya çıkıyor: büyük miktarda veriye sığamıyor, ihtiyaç duyulan analitik modellere destek verilmiyor, veri yüklemesi çok yavaş, gelişmiş analitik platformu gerekliliği, BT talepleri karşılayamıyor.

Büyük veri örnekleri

Literatürdeki birçok alanda büyük veri örneği mevcuttur. Bunlar astronomi, atmosfer bilimi, genomik, biyojeokimyasal, biyolojik bilim ve araştırma, yaşam bilimleri, tıbbi kayıtlar, bilimsel araştırma, hükümet, doğal afet ve kaynak yönetimi, özel sektör, askeri gözetim, özel sektör, finansal hizmetler, perakende, sosyal ağlar, web günlükleri, metin, belge, fotoğraf, ses, video, tıklama akışları, arama indeksleme, arama detay kayıtları, POS bilgileri, RFID, mobil telefonlar, sensör ağları ve telekomünikasyondur (http://en.wikipedia.org).
McKinsey Global Institute, büyük verinin potansiyelini beş ana başlıkta belirlemiştir (Gerhardt ve ark., 2012; Manyika ve ark., 2011).
Sağlık hizmetleri: Klinik karar destek sistemleri, hasta profili için uygulanan bireysel analitik, kişiselleştirilmiş tıp, personel için performansa dayalı fiyatlandırma, hastalık modellerini analiz etme, halk sağlığını iyileştirme, vb.
Kamu sektörü: Erişilebilir ilgili verilerle şeffaflık yaratmak, ihtiyaçları keşfetmek, performansı iyileştirmek, uygun ürün ve hizmetler için eylemleri özelleştirmek, riskleri azaltmak için otomatik sistemlerle karar vermek, yeni ürün ve hizmetler geliştirmek, vb.
Perakende: Mağaza içi davranış analizi, çeşitlilik ve fiyat optimizasyonu, ürün yerleştirme tasarımı, performansı artırma, işgücü girdileri optimizasyonu, dağıtım ve lojistik optimizasyonu, web tabanlı pazarlar, vb.
İmalat: Gelişmiş talep tahmini, tedarik zinciri planlaması, satış desteği, gelişmiş üretim operasyonları, web araması tabanlı uygulamalar, vb.
Kişisel konum verileri: Akıllı yönlendirme, coğrafi hedefli reklamcılık veya acil müdahale, şehir planlaması, yeni iş modelleri, vb.

Web, büyük veriler için de bir tür fırsatlar sunmaktadır. Örneğin; daha hedefli reklamcılık için kullanıcı zekasını anlamak, pazarlama kampanyaları ve kapasite planlaması, müşteri davranışı ve satın alma kalıpları gibi sosyal ağ analizi de duygu analitiği gibi. Bu çıkarımlara göre firmalar içeriklerini ve öneri motorlarını optimize etmektedir (Vailaya, 2012). Google ve Amazon gibi bazı şirketler çalışmalarıyla ilgili makaleler yayınlamaktadır. Yayınlanan yazılardan ilham alan geliştiriciler, Lucene, Solr, Hadoop ve HBase gibi açık kaynaklı yazılımlara benzer teknolojiler geliştiriyorlar. Facebook, Twitter ve LinkedIn bir adım daha ileri giderek Cassandra, Hive, Pig, Voldemort, Storm, IndexTank gibi büyük veriler için açık kaynak projeleri yayınlamışlardır.

Büyük Veri Metotları

Çoğu kuruluş, birçok farklı biçimde gelen çok sayıda yeni veriyle karşı karşıyadır. Büyük veri, her işletmeyi dönüştürebilecek iç görüler sağlama potansiyeline sahiptir. Büyük veri, *MapReduce* gibi mimarileri destekleyen yepyeni bir endüstri yarattı. MapReduce, Google tarafından karmaşık büyük veri sorunlarını küçük iş birimlerine bölmek ve bunları paralel olarak işlemek için böl ve yönet yöntemini kullanarak oluşturulan dağıtılmış bilgi işlem için bir programlama çerçevesidir (Schneider, 2012). MapReduce iki aşamaya ayrılabilir (Bakshi, 2012):
Harita Adımı: Ana düğüm verileri birçok küçük alt probleme bölünür. Bir çalışan düğüm, JobTracker düğümünün kontrolü altındaki daha küçük sorunların bazı alt kümelerini işler ve sonucu, bir indirgeyicinin erişebildiği yerel dosya sisteminde saklar.
Azaltma Adımı: Bu adım, harita adımlarından gelen girdi verilerini analiz eder ve birleştirir. Toplama işlemini paralelleştirmek için birden fazla azaltma görevi olabilir ve bu görevler, JobTracker'ın kontrolü altındaki çalışan düğümlerinde yürütülür.
Hadoop, Google'ın veri depolama sistemi olan BigTable, Google Dosya Sistemi ve MapReduce'dan ilham almak için oluşturuldu (Begoli ve Horey, 2012). Hadoop, Java tabanlı bir çerçeve ve heterojen açık kaynak platformudur. Veritabanı, ambar veya ETL (Çıkart, Dönüştür, Yükle) stratejisinin yerine geçmez. Hadoop, dağıtılmış bir dosya sistemi, analitik ve veri depolama platformları ve paralel hesaplama, iş akışı ve konfigürasyon yönetimini yöneten bir katman içerir (Sagiroglu ve Sinanc, 2013). Akışlar gibi gerçek zamanlı karmaşık olay işleme için tasarlanmamıştır. HDFS (Hadoop Dağıtılmış Dosya Sistemi), bir Hadoop kümesindeki düğümler arasında çalışır ve birçok girdi ve çıktı veri düğümündeki dosya sistemlerini birbirine bağlayarak onları büyük bir dosya sistemine dönüştürür (Sagiroglu ve Sinanc, 2013).

Büyük Veriden Bilgi Keşfi

Veriden Bilgi Keşfi (KDD), karmaşık veri kümelerinden bilgi almak için tasarlanmış bazı işlemler olarak adlandırılmaktadır (Begoli ve Horey, 2012). Referans (Fayyad ve ark., 1996), KDD'yi dokuz adımda özetlemektedir:
1. Bilgi öncesi uygulama alanı ve müşterinin bakış açısıyla sürecin amacını tanımlama.
2. Bilgi keşfi için alt küme veri noktası oluşturun.
3. Gürültünün giderilmesi, eksik veri alanlarının ele alınması, modelleme için gerekli bilgilerin toplanması ve zaman bilgisi ile bilinen değişikliklerin hesaplanması.
4. İşin amacına bağlı olarak verileri sunmak için faydalı özellikler bulma.
5. Amaçları belirli bir veri madenciliği yöntemleriyle eşleştirme.
6. Veri modellerini aramak için veri madenciliği algoritmasını ve yöntemini seçin.
7. Kalıpları anlatım biçiminde araştırmak.
8. Yinelemeler için 1'den 7'ye kadar olan adımların döndürülmesi, bu adım, kalıpların görselleştirilmesini de içerebilir.
9. Bilgiyi doğrudan kullanma, bilgiyi birleştirme

Referans (Begoli ve Horey, 2012), Hadoop kullanarak büyük verilerden bilgi keşfini üç ilkede analiz eder. Bunlar:

KDD, dağıtılmış programlama, örüntü tanıma, veri madenciliği, doğal dil işleme, duygu analizi, istatistiksel ve görsel analiz ve insan bilgisayar etkileşimi gibi çeşitli analiz yöntemlerini içerir. Bu nedenle mimari, çeşitli yöntem ve analiz tekniklerini desteklemelidir (Sagiroglu ve Sinanc, 2013).
Büyük veri kümelerini özetlemek, verileri anlamak ve tahmin için modeller tanımlamakla ilgilenen istatistiksel analiz.
Veri madenciliği, büyük veri kümelerinde kendi başına faydalı modelleri keşfetmekle ilişkilidir, makine öğrenimi, veri madenciliği ve makinelerin veri kümelerini anlamasını sağlayan istatistiksel yöntemlerle birleşir.
Görsel analiz, büyük veri kümelerinin kullanıcılara zorlu yollarla sunulduğu, ilişkileri anlayabilecekleri gelişmekte olan bir alandır.
Proses hattını korumak ve işletmek için kapsamlı bir KDD mimarisi temin edilmelidir (Sagiroglu ve Sinanc, 2013).
Hatalar, eksik değerler ve kullanılamayan formatta uygun sorun giderme için veri ve toplu analizlerin hazırlanması yapılır.
Yapılandırılmış ve yarı yapılandırılmış verilerin işlenmesi
Sonuçları erişilebilir ve kusursuz kılmak esastır. Bu nedenle bu sorunu aşmak için aşağıdaki yaklaşımlar kullanılmaktadır (Sagiroglu ve Sinanc, 2013).
Açık kaynak ve popüler standartları kullanma
WEB tabanlı mimarileri kullanın
Herkese açık sonuçlar

Hadoop

Apache Hadoop projesi, güvenilir, ölçeklenebilir, dağıtılmış bilgi işlem için açık kaynaklı yazılım geliştirir. Apache Hadoop yazılım kitaplığı, basit bir programlama modeli kullanarak büyük veri kümelerinin bilgisayar kümeleri arasında dağıtılmış olarak işlenmesine izin veren bir çerçevedir. Uygulamaların binlerce hesaplamadan bağımsız bilgisayar ve petabaytlarca veri ile çalışmasını sağlar. Hadoop, Google'ın MapReduce ve Google Dosya Sisteminden (GFS) türetilmiştir (Patel ve ark., 2012; http://hadoop.apache.org).
Hadoop, kullanıcı etkileşimine ihtiyaç duymadan verilerin ilgili düğüme taşınmasını ve olası bir donanımsal sorunda kurtarma işlemlerinin otomatik olarak gerçekleştirilmesini sağlar. Kullanılan Mapreduce tekniğinde, yazılan program bilgisayar kümesinde bulunan bir düğümde (node) çalışabilecek küçük parçalara ayrılır. Ayrıca Hadoop bilgisayar kümesi için yüksek bant genişliğine olanak sağlayan dağıtık dosya sistemini (Hadoop Distributed File System - BAŞLA HADOOP MAP REDUCE MANNKENDAL SPEARMAN'S RHO BİTİR 21 HDFS) kullanıcıya sunar (Kaya, 2018). Hem Mapreduce hem de HDFS dosya sisteminin birlikte kullanımıyla düğüm hataları otomatik olarak düzeltilir. Böylece kullanıcıya çok daha yüksek güvenilirlikli bir sistem sunulmuş olunur (White, 2009).

HDFS (Hadoop Distributed File System)

Hadoop Dağıtılmış Dosya Sistemi (HDFS), hata toleransı sağlayan ve ticari donanım üzerinde çalışmak üzere tasarlanmış dağıtılmış bir dosya sistemidir. HDFS, uygulama verilerine yüksek verimli erişim sağlar ve büyük veri kümelerine sahip uygulamalar için uygundur. Hadoop, binlerce sunucuda veri depolayabilen dağıtılmış bir dosya sistemi (HDFS) ve bu makineler arasında işi (Haritalama/Küçültme işleri) çalıştırarak verilerin yakınında çalıştıran bir araç sağlar. HDFS, master/slave mimarisine sahiptir. Büyük veriler, hadoop kümesindeki farklı düğümler tarafından yönetilen parçalara otomatik olarak bölünür (Patel ve ark., 2012; http://hadoop.apache.org).

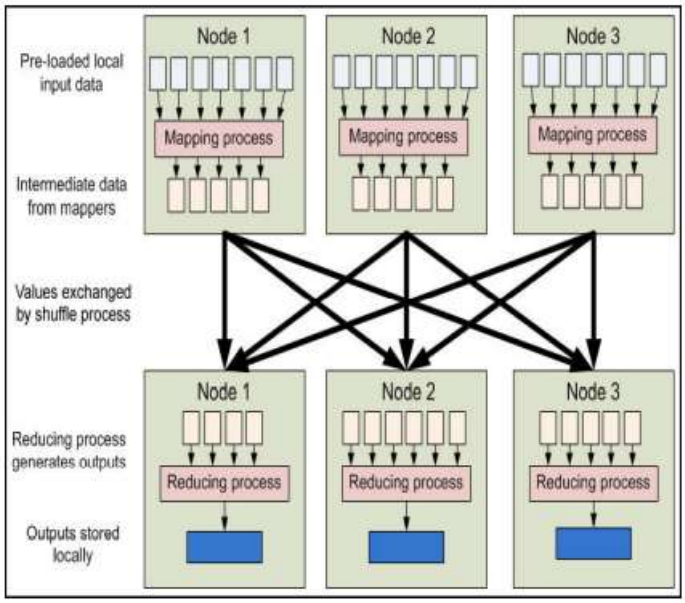
MapReduce Programming Framework

MapReduce, bilgisayar kümelerindeki büyük veri kümeleri üzerinde dağıtılmış hesaplamayı desteklemek için Google tarafından 2004 yılında tanıtılan bir yazılım çerçevesidir. MapReduce, büyük veri kümelerini işlemek ve oluşturmak için bir programlama modelidir. Kullanıcılar, bir ara anahtar/değer çiftleri kümesi oluşturmak için bir anahtar/değer çiftini işleyen bir eşleme işlevi ve aynı ara anahtarla ilişkili tüm ara değerleri birleştiren bir azaltma işlevi belirtir (Yang ve ark., 2007; http://hadoop.apache.org).
"*Map*" *adımı*: Ana düğüm girişi alır, onu daha küçük alt problemlere böler ve bunları çalışan düğümlere dağıtır. Bir işçi düğümü bunu sırayla tekrar yapabilir ve çok seviyeli bir ağaç yapısına yol açabilir. Çalışan düğüm, daha küçük sorunu işler ve yanıtı ana düğümüne geri iletir. Harita, bir veri etki alanındaki bir türe sahip bir çift veri alır ve farklı bir etki alanındaki çiftlerin bir listesini döndürür:

$Map(k1, v1) \rightarrow list(K2, v2)$ (Patel ve ark., 2012)

"Reduce" adımı: Ana düğüm daha sonra tüm alt problemlerin cevaplarını toplar ve bunları bir şekilde çıktıyı oluşturmak için birleştirir - başlangıçta problemin cevabını çözmeye çalışır.
Azaltma işlevi daha sonra her gruba paralel olarak uygulanır ve bu da aynı etki alanında bir değerler koleksiyonu üretir:

$Reduce\big(K2, list(v2)\big) \rightarrow list(v3)$ (Patel ve ark., 2012)

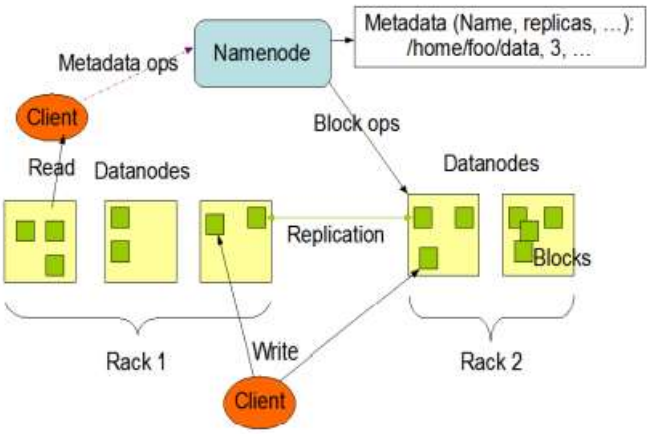


Şekil 2. Dağıtık map ve reduce süreçleri (Patel ve ark., 2012)

Sistem mimarisi

Sistem mimarisi, hadoop mimarisi, hadoop çok düğümlü küme kurulumu, HDFS kurulumu ve yoğun veri sorununu çözmek için Harita Azaltma programlama çalışmasının uygulanmasından oluşur.
HDFS Mimarisi: Şekil 3'te gösterildiği gibi, bir HDFS kümesi, dosya sistemi ad alanını yöneten ve istemciler tarafından dosyalara erişimi düzenleyen bir ana sunucu olan tek bir NameNode'dan oluşur. Ek olarak, üzerinde çalıştıkları düğümlere bağlı depolamayı yöneten, genellikle kümedeki düğüm başına bir tane olmak üzere bir dizi DataNode vardır. HDFS, bir dosya sistemi ad alanını ortaya çıkarır ve kullanıcı verilerinin dosyalarda saklanmasına izin verir. Dahili olarak, bir dosya bir veya daha fazla bloğa bölünür ve bu bloklar bir DataNode kümesinde depolanır. NameNode, blokların Datanode'lara eşlenmesini belirler. HDFS, çok büyük dosyaları büyük bir kümedeki

makineler arasında güvenilir bir şekilde depolamak için tasarlanmıştır. Her dosyayı bir dizi blok olarak saklar.



Şekil 3. HDFS mimarisi (Patel ve ark., 2012)

Hadoop Kümesi Yüksek Düzey Mimarisi: Hadoop kümesi, tek bir ana ve birden çok bağımlı veya "işçi düğümünden" oluşur. JobTracker, MapReduce görevlerini kümedeki belirli düğümlere, ideal olarak verilere sahip düğümlere veya en azından aynı rafta bulunan düğümlere dağıtan Hadoop içindeki hizmettir (White, 2010).
TaskTracker, kümede bir JobTracker'dan görevleri (Haritalama, Küçültme ve Karıştırma işlemleri) kabul eden bir düğümdür. Ana düğüm bir JobTracker, TaskTracker, NameNode ve DataNode'dan oluşur. Bir bağımlı veya çalışan düğüm, hem DataNode hem de TaskTracker görevi görür. Daha büyük bir kümede, HDFS, dosya sistemi dizinini barındırmak için ayrılmış bir NameNode sunucusu ve bunu sağlayan ikincil bir NameNode aracılığıyla yönetilir. İsim düğümünün (Namenode) bellek yapılarının anlık görüntülerini oluşturabilir, böylece dosya sisteminin bozulmasını önleyebilir ve veri kaybını azaltabilir.

SONUÇ

Bu bölümde, büyük verinin içeriğine, kapsamına, örneklerine, yöntemlerine, avantajlarına ve zorluklarına genel bir bakış ve gizlilik endişesi ele alınmıştır. Sonuçlar göstermiştir ki, literatürde mevcut veriler, araçlar ve teknikler mevcut olsa bile dikkate alınması, tartışılması, iyileştirilmesi, geliştirilmesi, analiz edilmesi vb. pek çok nokta bulunmaktadır. Büyük veri sorunu gelecekte daha fazla tartışılacaktır. Ayrıca hadoop yazılımı da incelenmiştir. Hadoop, Mapreduce yöntemini kullanarak büyük veri dosyalarının paralel şekilde işlenmesini sağlayan açık kaynak kodlu bir yapıdır (Lam, 2010). Bilgisayar kümeleri üzerinde çalışan dağıtık programlar için alt yapı desteği sunmaktadır (Er, 2013).

Bu makale, bu önemli konu hakkındaki tüm konuyu açıkça çözmemiş olsa da, umarım araştırmacılar için bazı yararlı tartışmalar ve bir çerçeve sağlamıştır.

KAYNAKLAR

https://evrimagaci.org/big-data-nedir-buyuk-veri-yapay-zekanin-zincirlerini-kirmasini-saglayacak-anahtar-olabilir-mi-11347 (Erişim Tarihi: 19.01.2022)
Sagiroglu S, Sinanc D, 2013. "Big data: A review," *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 42-47, doi: 10.1109/CTS.2013.6567202.
Gerhardt B, Griffin K, Klemann R, 2012. Unlocking Value in the Fragmented World of Big Data Analytics, Cisco Internet Business Solutions Group, June 2012,

http://www.cisco.com/web/about/ac79/docs/sp/InformationInfomediaries.pdf. (Erişim Tarihi:11.03.2022)
Eaton C, Deroos D, Deutsch T, Lapis G, Zikopoulos PC, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012.
Singh S, Singh N, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011.
Madden S, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6.
Russom P, "Big Data Analytics ", TDWI Best Practices Report, TDWI Research, Fourth Quarter 2011, http://tdwi.org/research/2011/09/best-practices-report-q4-big-dataanalytics/asset.aspx. (Erişim Tarihi:11.03.2022)
http://en.wikipedia.org/wiki/Big_data, (Erişim Tarihi: 20.01.2022).
Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburg C, Byers AH, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx. (Erişim Tarihi:11.03.2022)
Vailaya A, What's All the Buzz Around "Big Data?, IEEE Women in Engineering Magazine, December 2012, pp. 24-31.
Schneider RD, Hadoop for Dummies Special Edition, John Wiley&Sons Canada, 978-1-118-25051-8, 2012.
Bakshi K,"Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012.
Begoli E, Horey J,2012. Design Principles for Effective Knowledge Discovery from Big Data, Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, August 2012.
Fayyad U, Piatetsky-Shapiro G, Smyth P, 1996. From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, AI Magazine, Fall 1996, pp. 37- 54.
Patel AB, Birla M, Nair U, 2012, Addressing Big Data Problem Using Hadoop and Map Reduce, 2012 NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-2012, 06-08DECEMBER.
Apache Software Foundation. Official apache hadoop website, http://hadoop.apache.org/, Aug, 2012. (Erişim Tarihi:11.03.2022)
Yang HC, Dasdan A, Hsiao RL, 2007. Map-ReduceMerge: Simplified Data Processing on Large Clusters, paper published in Proc. of ACM SIGMOD, pp. 1029– 1040.
The Hadoop Architecture and Design, Available: http://hadoop.apache.org/common/docs/r0.16.4/hdfs_des ign.html, Aug, 2012. (Erişim Tarihi:11.03.2022)
White T, Hadoop The Definitive Guide 2nd Edition. United States : O'Reilly Media, Inc., 2010.
Kaya M, 2018. HADOOP KULLANARAK METEOROLOJİ VERİLERİNDEN BİR İKLİM DEĞİŞİMİ EĞİLİM ANALİZİ, SÜLEYMAN DEMİREL ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ, Yüksek Lisans Tezi, 83s, Isparta.
White T, 2009. Hadoop: The Definitive Guide. O'Reilly Media, Inc., 625p, Sebastopol, USA.
Lam C, 2010. Hadoop in Action. Manning Publications Co., 301p, Stamford,USA.
Er HR, 2013. Gezgin Satıcı Probleminin Hadoop Üzerinde Çalışan Paralel Genetik Algoritma İle Çözümü, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 83s, Kocaeli.