SOCIAL SCIENCES STUDIES JOURNAL

SSSjournal (ISSN:2587-1587)

Economics and Administration, Tourism and Tourism Management, History, Culture, Religion, Psychology, Sociology, Fine Arts, Engineering, Architecture, Language, Literature, Educational Sciences, Pedagogy & Other Disciplines in Social Sciences

Vol:3, Issue:6	pp.873-883	2017
sssjournal.com	ISSN:2587-1587	sssjournal.info@gmail.com

Article Arrival Date (Makale Geliş Tarihi) 05/10/2017 | The Published Rel. Date (Makale Yayın Kabul Tarihi) 13/11/2017

Published Date (Makale Yayın Tarihi) 16.11.2017

BÜYÜK VERİ TEKNOLOJİLERİNİN İŞLETMELER İÇİN ÖNEMİ

IMPORTANCE OF BIG DATA TECHNOLOGIES FOR BUSINESSES

Arş. Gör. Sadullah ÇELİK

Adnan Menderes Üniversitesi, Nazilli İİBF, Ekonometri Bölümü, ssadullah.celik@gmail.com, Aydın/Türkiye

ÖZ

Veri üretimi her geçen gün katlanarak artmakta. Üretilen bu verinin hacmi, çeşitliliği ve hızı mevcut verilerden daha fazla olup analiz edilmesi ve yorumlanması oldukça zordur. Milyarlarca ağa bağlı sensörler akıllı telefonlar, otomobiller, sosyal medya siteleri, dizüstü bilgisayarlar, PC'ler ve endüstriyel makineler gibi verileri çalıştıran, üreten ve ileten cihazlara yerleştirilmiştir. Bu tür kaynaklardan elde edilen veriler yapılandırılmış, yarı yapılandırılmış veya yapılandırılmamış formattaki verilerdir. Geleneksel veritabanı sistemleri bu veri türlerini işlemede yetersiz kalmaktadır. Bu nedenle yeni teknolojilere ihtiyaç duyulmuştur. Bugün geliştirilen teknolojiler büyük veri setlerini; toplama, işleme, analiz etme ve görselleştirmede oldukça başarılıdır. Bu teknolojiler özellikle yapısal olmayan büyük veri setlerini kolayca analiz ederek, şirketlere büyük avantajlar sağlamaktadır. Bu çalışmanın amacı, Büyük Veri Analizinde kullanılan Hadoop ve Spark teknolojilerinin yapılarını tanıtmak ve bunların şirketler açısından sağladığı avantajları ele almaktır.

Anahtar Kelimeler: Büyük Veri, Hadoop, Spark, EşleİndirgeBüyük Veri, Hadoop, Spark, MapReduce

ABSTRACT

Data production is increasing day by day. The volume, diversity and speed of this generated data is more than the available data and is difficult to analyze and interpret. Billions of networked sensors are embedded in devices that run, generate and transmit data such as smartphones, automobiles, social media sites, laptop computers, PCs and industrial machines. The data obtained from such sources is the data in the structured, semistructured, or unstructured form. Traditional database systems are insufficient to process these data types. That's why new technologies are needed. The technologies developed today are large data sets; Collecting, processing, analyzing and visualizing. These technologies provide great advantages for companies, especially by easily analyzing large unstructural data sets. The purpose of this study is to introduce the practices of Hadoop and Spark technologies used in Big Data Analysis and to discuss the advantages they provide for companies.

Keywords: Big Data, Hadoop, Spark, MapReduce

1. GİRİŞ

Büyük Veri, geleneksel veritabanları kullanarak analizi yapılamayan ve yönetilemeyecek kadar büyük miktardaki veri setleri olarak adlandırılır (Ohlhorst, 2013:19). Bugün dünyadaki veri miktarı birçok nedenden dolayı üssel olarak artmakta. Çeşitli kaynaklar ve gündelik faaliyetlerimiz birçok veri üretiyor. Web'in icadıyla birlikte tüm dünya çevrimiçi haline geldi ve yaptığımız her şey dijital bir iz bırakıyor. Akıllı nesneler çevrimiçi hale geldikçe, veri büyüme oranı daha da hızlı arttı. Büyük Veri'nin ana kaynakları, sosyal medya siteleri, sensör ağları, dijital görüntüler/videolar, cep telefonları, satın alma işlemi kayıtları, web günlükleri, tıbbi kayıtlar, arşivler, askeri gözetimler, e-ticaret, karmaşık bilimsel araştırmalar vb. tüm bu bilgiler, yaklaşık beş katrilyon veri baytı civarındadır. Bilimadamları, 2020 yılına gelindiğinde, veri hacminin büyüklüğünün yaklaşık 40 zettabayt olacağını tahmin etmekte (Edureka, 2017).

Peki zaman içinde bu noktaya nasıl gelindi, dünya nerede sürüm 1.0'dan sürüm 2.0'ye yükseldi, yükseltilen daha akıllı olan sürüm hangisi? Hiç kuşku yok ki tüm bu değişimlerin ortaya çıkmasında birçok trend etkin rol oynamıştır. Dünya'nın 2.0 dijital yönü üzerine odaklanırsak, bilgi devriminin yaşanmasına sebep olan üç büyük trend üzerinde durmak gerekir (Datafloq, 2017):

- ✓ Buluttaki sınırsız işlem gücü

- ✓ Her şeyi akıllı hale getirecek sensörler kümesi
- ✓ Akıllı algoritmalar sayesinde Yapay Zeka ve Makine öğrenme

Bu üç büyük trend yaptığımız iş dahil olmak üzere, toplumun herhangi bir bölümünü etkileyecektir. Şimdi bunların her birine bakalım:

1.1. Sınırsız İşlem Gücü

Son yıllarda, veri depolama fiyatı Gigabayt başına yaklaşık olarak $0,03'a kadar düşmüş, Moore yasasına göre bu fiyatın önümüzdeki yıllarda daha da düşmesi bekleniyor (Datafloq, 2017). Öyle ki bugün artık veri silme veri kaydetmeden daha pahalı hale geldi. Bu sayede kuruluşlar için veri toplama ve depolama sorununu büyük ölçüde ortadan kalkmıştır (Bkz. Şekil 1). Bu da herhangi bir verinin toplanabilmesi ve depolanabilmesi, akıllı algoritmalar kullanılarak analiz edilebilmesinin yanında çok daha fazlasının yapılabileceği anlamına geliyor.

Şekil 1. Veri depolama merkezi (Datafloq, 2017)

Fakat donanım sınırsız hesaplama gücünün tek bir bileşeni değildir. Son yıllarda, gördüğümüz yeni teknolojiler tüm veri analizlerini yapabilecek kapasiteye sahiptir. Burada sadece saniyeler içinde verinin Terabayt ya da Petabayt'ından bahsedilmektedir.

Apache Spark gibi açık kaynak teknolojileri 1000'lerce düğüm sayesinde dağıtık bir ağ üzerinden verinin bellek analizini yapabilme kapasitesine sahiptirler (Datafloq, 2017). Bu açık kaynak teknolojileri sensörlerdeki büyük miktardaki veriyi, internette yaratılan bütün bilgiyi, ayrıca dünyada tüketiciler tarafından yaratılan tüm yapılandırılmamış verinin analizini yapabilecek kapasiteye sahiptirler.

1.2. Sensörler Kümesi

Sınırsız işlem gücü bize geçmiş yıllara bağlı bir dünya yarattı. Bu dünya giderek daha ve daha fazla bağlantılı olmakla birlikte gelecek on yıl içinde bir trilyon kadar bağlanabilir cihazın olacağı tahmin edilmekte. Tüm bu sensörler dünyamızı akıllı hale getirecek ve bizler bu sensörleri Nesnelerin İnterneti (Internet of Things) olarak arayacağız. Dünyaca tanınmış ağ teknoloji şirketi Cisco'ya göre, Nesnelerin İnterneti gelecek on yıl içinde 19 trilyon dolarlık bir pazara sahip olacaktır. Bu üretim için 2.9 trilyon dolar öngörülmektedir (Datafloq, 2017).

Gelecekte, aklınıza gelebilecek herhangi bir yerde sensörleri bulabilirsiniz. Trafiği izlemek için yollara belirli aralıklarla sensörler yerleştirilir. Sensörler makinalarda tahmine dayalı bakım ya da araçlardaki sürüş davranışlarını izlemek ve buna göre sigorta poliçesi oluşturmak için kullanılır. Sensörler gelecekte çok ucuz ve çok küçük olacak öyle ki onları giysilerin ve ilaçların içerisinde de bile bulabileceksiniz. Daha şimdiden insan derisinin altına konulan sensörler kullanılmaya başlanmıştır.

1.3. Akıllı Algoritmalar

Trilyonlarca sensörün ürettiği büyük miktardaki veriyi anlamak için akıllı algoritmalara ihtiyaç duyulur. Neyse ki geçmişteki algoritmaların geliştirilmesi Yapay Zeka, Makine-Öğrenme ve Derin Öğrenme sayesinde bir sonraki seviyeye gelindi. Yapay Zeka'nın günümüzde çok büyük bir kullanım potansiyeli vardır. 2013 yılında Oxford Üniversitesi tarafından yapılan bir araştırmada, Yapay Zeka'nın yakın gelecekte ABD'deki tüm işlerin neredeyse yarısından fazlasını elinden alabileceği tahmin ediliyor (Datafloq, 2017). Yapay Zeka'nın en yaygın uygulaması verinin büyük miktarlarındaki desenlerinin bulunması ve otomatik olarak bağımsız harekete geçme üzerinedir. Bu otomatik ve gelişmiş karmaşık tanımlayıcı, tahmin ve kuralcı analitik görev, firmalara eşi benzeri görülmemiş düzeyde değer yaratmaya izin verir.

2. BÜYÜK VERİ TEKNOLOJİLERİ

Büyük veri setlerini depolamak, işlemek, yönetmek ve analiz etmek için kullanılan teknolojilerin sayısı her geçen gün artmakta. 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 sahiptir. Büyük Veri de özellikle büyük hacimlerde ve yapılandırılmamış olan formattaki veriyi yönetmek için gerekli olan teknolojiye odaklanılır. Büyük Veri Teknolojileri yeni olmalarına rağmen bugün büyük ilgi uyandırmıştır (McKinsey Global Institute, 2011). Büyük Veri Teknolojilerinin en önemli özelliği kuruluşlara ne şekilde değer kazandırabileceği, maliyeti düşürme, veri işleme süresini azaltma, yeni ürün ve hizmet kalitesini geliştirme veya daha iyi kararlar almak için yeni veri modellerinin kullanılmasına imkan sağlamasıdır (Thomas, 2014).

2.1. Hadoop Bileşenleri Ve Mimarisi

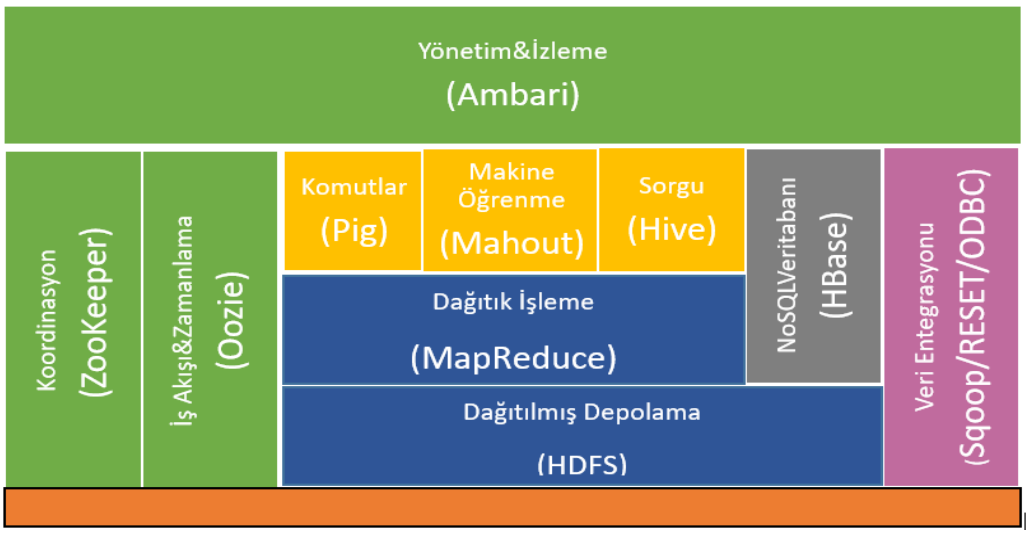
Hadoop, bir makineden başlayarak, yüzlerce makine üzerine dağılabilen Büyük Veri kümelerini işlemek için kullanılan, Java ile geliştirilmiş (ücretsiz) yazılım çatısıdır. Bu uygulamalarda genellikle Web üzerinde kullanılabilen ve çoğunlukla kullanılan açık uygulama programlama arayüzleri aracılığıyla açık veri kaynaklarından erişilen veriler kullanır (Dezyre, 2017).

2.1.1. Apache Foundation Tarafından Tanımlanan Hadoop

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 imkanı sunmaktadır. Yüksek erişilebilirlik sağlamak için donanıma güvenmek yerine, kütüphane kendisi, başarısızlıkları uygulama katmanında algılamak ve ele almak üzere tasarlanmıştır. Bu nedenle Hadoop, her biri başarısızlıklara eğilimli olabilen bir bilgisayar kümesinin üstünde yüksek oranda mevcut bir hizmet sunmaktadır (Mohammad, 2011:7). Apache Hadoop, anlamlı bilgiler elde etmek için analitikten yararlanmak için büyük miktarda veri kullanıldığında, Büyük Verileri işlemek için iyi bir çözümdür. Apache Hadoop mimarisi, çeşitli hadoop bileşenleri ve karmaşık iş problemlerini çözmek için muazzam özellikleri olan farklı teknolojilerin birleşmesinden oluşur.

Hadoop mimarisinin bütünsel yapısını Hadoop Ekosistemi'ndeki; Hadoop Common, Hadoop YARN (Yet Another Resource Negotiator), HDFS (Hadoop Distributed File System) ve MapReduce elemanları oluşturmaktadır. Bu ana bileşenlerin altında ise başka araçlar bulunmaktadır. Hadoop Common, 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 MapReduce, Büyük Veri kümelerinin YARN tabanlı paralel işlenmesini sağlar.

Verilen iş sorunlarına doğru çözümler üretmek için Hadoop mimarisine ve bileşenlerine derinlemesine girmek gerekir (Bkz. Şekil 2).



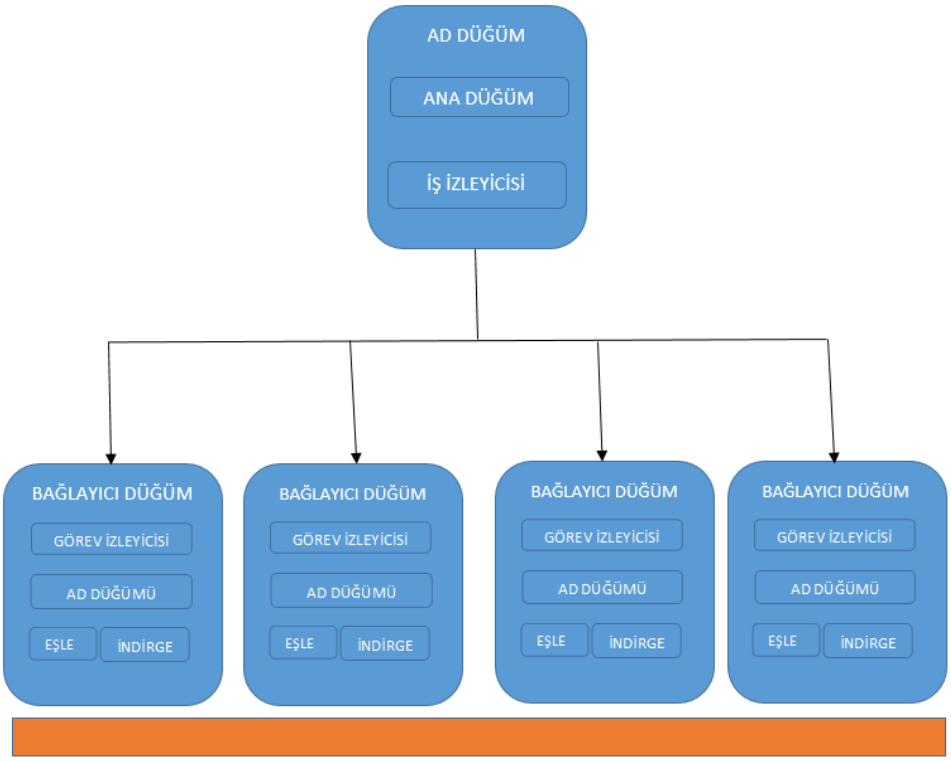
Şekil 2. Apache Hadoop Ekosistemi (Slidershare, 2016)

2.1.2. Hadoop Büyük Veri Ekosisteminin Mimari Bileşenleri

Hadoop Ekosistemi; Hadoop Common, Hadoop Dağıtılmış Dosya Sistemi (Hadoop Distributed File System-HDFS), MapReduce (Apache Hadoop'un Dağıtık Veri İşleme Çerçevesi) ve YARN olmak üzere dört temel bileşenden oluşmaktadır.

Hadoop Common, Apache Foundation, Hadoop ekosistemindeki diğer modüller tarafından kullanılabilen önceden tanımlanmış bir dizi yardımcı program ve kütüphaneye sahiptir. Örneğin, HBase ve Hive, HDFS'ye erişmek istiyorsa, Hadoop Common'da saklanan Java arşivlerini (JAR dosyaları) oluşturmaları gerekir.

HDFS, Apache Hadoop için varsayılan Büyük Veri depolama katmanıdır. Kullanıcılar, Büyük Veri kümelerini HDFS'ye dökebilecekleri için HDFS, Apache Hadoop bileşenlerinin "Gizli Sosu" olarak adlandırılır ve verilerin analiz için hazır hale gelmesi burada olur. 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 oluşur. 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 (Master/Slave Node) (Bkz. Şekil 3) Düğümü mimarisi modelinde çalışır (Dezyre, 2017).



Şekil 3. Hadoop Ana/Bağlayıcı Düğüm Mimarisi (Slidershare, 2016)

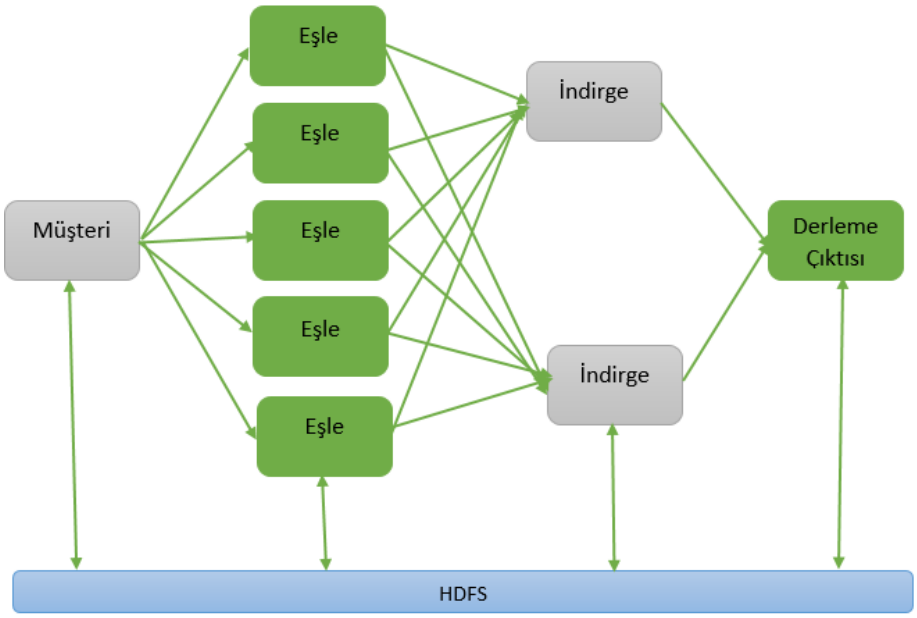
Tipik Bir Hadoop dağıtık dosya sisteminde makine rolünü üstlenen İstemci makineleri, Ana düğümleri ve Bağlayıcı düğümleri bulunur. Ana düğümler, Hadoop'u oluşturan iki önemli fonksiyonel parçayı denetlemektedir: çok miktarda veri depolamak (HDFS) ve bu verilerin hepsine paralel hesaplamalar yürütmek (Map Reduce). Ad Düğümü, veri saklama işlevini (HDFS) denetler ve koordine ederken, İş İzleyici (Job Tracker), Map Reduce'ı 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 (Gürsakal, 2014).

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 MapReduce 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 sunucunuz olabilir. Orta ve büyük kümelerde her rol için tek bir sunucu makinesi gereklidir (Brad, 2017).

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

MapReduce'un arkasındaki temel çalışma prensibi şöyledir: Map (Eşle) ve Reduce (İndirge) birer fonksiyon olup bu fonksiyonlar sayesinde işlenecek veriler birbirinden bağımsız parçalara ayrılır. Ayrılan bu parçaların her biri Map'e anahtar-değer çiftleri şeklinde eşlenerek iletilir (Bkz. Şekil 4). Daha sonra Map'ten çıkan veriler gruplanıp sıralanarak tekrar Reduce'e iletilir (Vaccari, 2014). Reduce kısmında ise bütün çiftler aynı anahtar değeri ile indirgenir. Bu arada, görev giriş ve çıkışları bir dosya sisteminde saklanır. MapReduce, bu işlerin zamanlamasını yapar, işleri izler ve başarısız olan görevi yeniden gerçekleştirir (Murthy, Markha, Vavilapalli, and Eadline, 2014).



Şekil 4. MapReduce çalışma prensibi (Dezyre, 2017)

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 MapReduce çerçevesinden yerleşik kod kullanır (Dezyre, 2017).

YARN, Hadoop 2.0 olarak bilinip günümüzde dağıtılan Büyük Verilerin işlenmesi ve yönetilmesi için yaygın olarak kullanılmaktadır. Hadoop YARN, Ekim 2013'te (Vavilapalli, 2013) piyasaya sürülen en son teknolojidir. Hadoop YARN, Hadoop veritabanı ve Hadoop HBase veritabanı 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 MapReduce'yi Hadoop sistemlerinde kullanmak zorunda kalmayan iş planlaması ve kaynak yönetimi görevlerini yerine getirir. Hadoop YARN, Hadoop 1.0'ın özgün özelliklerinden farklı olarak geliştirilmiş bir mimariye sahiptir, böylece sistemlerin yeni seviyelere kadar ölçeklenebilir ve Hadoop HDFS'deki çeşitli bileşenlere sorumluluklar kolayca atanabilmektedir (Dezyre, 2017).

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 hale getirmek veya yeni özellikler ve işlevler geliştirmek için kullanılmaktadır. Şirketler tarafından büyük

ölçüde kullanılan Hadoop bileşenleri şunlardır: Pig, Hive, Sqoop, Flume, HBase, Oozie ve Zookeeper'dır (Dezyre, 2017).

2.1.3. Hadoop Ekosisteminin Veri Erişim Bileşenleri-Pig ve Hive

Apache Pig, Büyük Veri kümelerinin etkili ve kolay bir şekilde analiz edilmesi için Yahoo tarafından geliştirilen kullanışlı bir araçtır. Pig Latin (Apache Pig Tutorial, 2017) dilini kullanarak optimize edilmiş, genişletilebilir ve kullanımı kolay yüksek düzeyde bir veri akışı sağlar. Pig programlarının göze çarpan önemli özelliği, yapılarının büyük veri setlerinin işlerliğini kolaylaştıran paralelleştirmeye açık olmasıdır (Dezyre, 2017).

Apache Hive, Facebook tarafından geliştirilmiş, Hadoop'un üzerine kurulmuş bir veri ambarıdır (Facebook, 2017). Hive, yapılandırılmamış Büyük Veri setlerinin yönetilmesi, sorgulanması, özetlenmesi ve analizi için SQL'e benzer HiveQL olarak bilinen basit bir dil kullanır (Hortonworks, 2017). Hive, sorgulamayı dizine ekleme yoluyla daha hızlı hale getirir. Hadoop YARN üzerine kurulmuş olan Hive'ın ölçeklenebilirlik, kullanılabilirlik ve hata toleransı gibi avantajları vardır (Apache Hive, 2017).

2.1.4. Hadoop Ekosisteminin Veri Bütünleştirme Bileşenleri - Sqoop ve Flume

Apache Sqoop, Hive RDBMS'lerden veri almak ve göndermek için Sqoop'ı kullanır. Sqoop, Hadoop ve harici ilişkisel veritabanları ile Hadoop ve NoSQL sistemleri arasındaki toplu veri aktarımını destekler. Sqoop, işletim sistemlerinden bağımsızdır ve konektör tabanlı mimarisi, ek harici sistemlere olan bağlantıyı desteklemektedir. Sqoop, verileri içe ve dışa aktarmak için MapReduce kullanır (Apache Sqoop, 2017). 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(Dezyre, 2017).

Apache Flume, büyük miktarda veriyi toplamak ve entegre etmek için kullanılır. Apache Flume, orijinalinden veri toplamak ve onu depolama yerine (HDFS) geri göndermek için kullanılır. Flume bunu kaynaklardan ve lavabo yapılarından oluşan veri akışlarını özetleyerek gerçekleştirir. Veri akışını flume ile çalıştıran süreçlere aracılar denir ve flume yoluyla akan veri parçacıkları olaylar olarak bilinir. 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 (Dezyre, 2017).

2.1.5. Hadoop Ekosistemi Veri Depolama Bileşeni –Hbase

Apache HBase, verilerin depolanması için HDFS'yi kullanan sütuna yönelik bir veritabanıdır. HBase, MapReduce kullanarak rasgele okumaları ve toplu hesaplamaları destekler. HBase ile NoSQL veritabanı kuruluşu, donanım makinesinde milyonlarca satır ve sütun içeren geniş matrisler oluşturulabilir (Gates, 2011).

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

Apache 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 depolayan bir veritabanı kullanır (Dezyre, 2017).

Apache 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 (Dezyre, 2017). Zookeeper araçları, geliştiricilerin kısmi arızaları güvenle idare eden dağıtılmış uygulamalar oluşturmasına olanak tanır (Mukhammadov, 2013).

Diğer taraftan yaygın olarak kullanılan Hadoop ekosistemi bileşenleri arasında Avro, Cassandra, Chukwa, Mahout, HCatalog, Ambari ve Hama bulunur. Kullanıcılar, bir veya daha fazla Hadoop ekosistem bileşenini kullanarak Hadoop'u uygulayarak, değişen iş gereksinimlerini karşılamak için Büyük Veri deneyimlerini kişiselleştirebilirler.

Apache Cassandra, bir açık kaynak olup dağıtılmış bir sistem üzerinde çok büyük miktarlarda veriyi işlemek için tasarlanmış (ücretsiz) veritabanı yönetim sistemidir. Bu sistemi ilk olarak Facebook geliştirildi ve şuanda Apache Software temelli bir proje olarak yönetilmektedir(Dezyre, 2017).

Apache Ambari, bir Hadoop bileşeni olan Ambari, Hadoop yönetimi için web kullanıcı arayüzü kullanımı kolay bir RESTful API'dir. Ambari, Hadoop hizmetlerini başlatmak, durdurmak ve yeniden yapılandırmak için merkezi yönetimle donatılmıştır. Ambari, Hadoop kümesinin sağlıklı işleyiş durumunu izleyebilen metrik toplama, uyarı çerçevesini kolaylaştırır (Dezyre, 2017).

Apache Mahout, makine öğrenimi için önemli bir Hadoop bileşenidir. Mahout, MapReduce için yazılan makine öğrenme algoritmalarının bir kütüphanesi, ancak makine öğrenme algoritmaları yinelemeli olduğundan birçok MapReduce işine ihtiyaç duyar. 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 bulutu 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 MapReduce etkin kümeleme uygulaması ile matris ve vektör kütüphanelerini içerir. Bugün; Adobe, Facebook, LinkedIn, Foursquare, Twitter ve Yahoo gibi şirketler Mahout'u yoğun olarak kullanmaktadır. Foursquare belirli bir alanda mevcut yerleri, yiyecek ve eğlence bulma konusunda Mahout'u kullanır (Dezyre, 2017).

Apache Kafka, Linkedln tarafından geliştirilmiş ve daha sonra 2011'de açık kaynaklı bir Apache projesi haline geldi. Apache Kafa, 2012'de birinci sınıf bir Apache projesi oldu. Apache Kafka Scala ve Java dillerinde yazılmış, yayın abone temelli hataya dayanıklı mesajlaşma sistemidir. Mesajlar, hızlı, ölçeklenebilir ve tasarım yoluyla dağıtılır (Apache Kafka, 2017).

Apache Hama, MapReduce'un ötesinde gelişmiş analizler yapmaya izin veren Apache üst düzey açık kaynak projesidir. Makine öğrenimi ve grafik algoritmaları gibi pek çok veri analiz tekniği yinelemeli hesaplamalar gerektirir. Hama, Toplu Eşzamanlı Paralel modelin (Bulk Synchronous Parallel model) ''düz'' MapReduce'dan daha etkili olabileceği yerdir (Dezyre, 2017).

Apache HCatalog, Apache Pig, Hive ve MapReduce kullanıcılarının HDFS'de saklanan verilerin ilişkisel görünümünü şemaları paylaşmalarını sağlayan Hive meta veri deposunun üzerine kurulmuş bir depolama yönetimi katmanıdır (Robert, 2006). HCatalog, harici sistemlere Hive meta verilerine erişime izin veren Representational State Transfer (REST) arabirimi (Robert, 2006) (WebHCat (WebHCat, 2016) vasıtasıyla) içerir. HCatalog, Hive SerDe'yi (Confluence, 2016) kullanarak çeşitli dosya formatlarını (örneğin RCFile, CSV, JSON ve SequenceFile biçimleri) okuma ve yazmayı destekler.

2.2. Apache Spark

Apache 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 büyük çaplı endüstrilerdeki işletmeler tarafından hızla benimsenmiştir. Netflix, Yahoo ve eBay gibi İnternet santralleri, toplu olarak 8000'den fazla düğümün kümeleri üzerinde birden fazla petabayt veri işleyen Spark'ı büyük çapta kullanıma açtı. 250'den fazla şirketin 1000'in üzerinde katkıda bulunanların, Büyük Veri alanındaki en büyük açık kaynak topluluğu haline geldi (Databricks, 2016). Apache Spark, hızlı hesaplama için tasarlanmış yıldırım hızlı küme bilgi işlem teknolojisidir. Apache Spark, Hadoop MapReduce'u temel alır ve MapReduce modelini etkileşimli sorgular ve akış işleme içeren daha fazla hesaplama türleri için verimli bir şekilde kullanmak için genişletir. 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. Bütün bu işyükünü ilgili bir sistemde desteklemenin yanı sıra, ayrı araçları korumanın yönetim yükünü de azaltmaktadır. Apache Spark aşağıdaki özelliklere sahiptir (Apache Spark Tutorial, (2017).

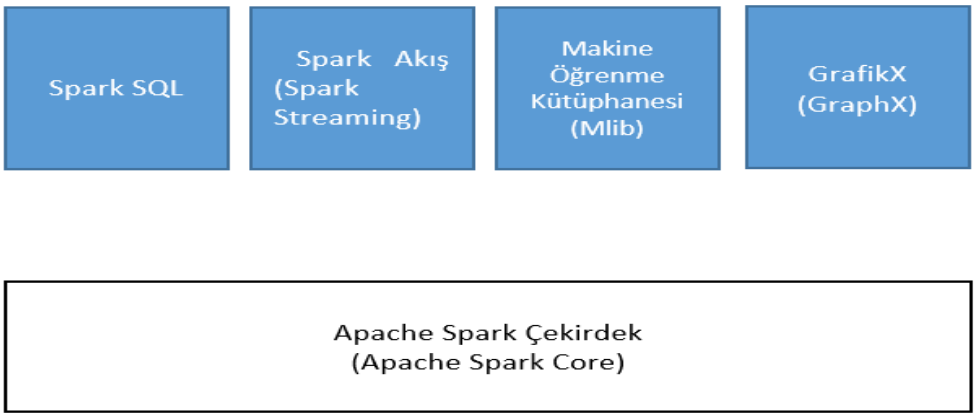
1) Speed: Spark Hadoop kümesinde bir uygulamayı çalıştırmaya yardımcı olur, bellekte 100 kat daha hızlıdır ve disk üzerinde çalışırken ise 10 kat daha hızlıdır. Spark'ta, diske okuma / yazma işlemlerinin sayısını azaltmak mümkündür. Ayrıca, ara işlem verisini belleğe kaydedilebilir.

2) Birden fazla dili destekler: Spark, Java, Scala veya Python'da yerleşik API'ler sağlar. Bu nedenle, uygulamaları farklı dillerde yazılabilir. Spark, interaktif sorgulama için 80 üst düzey operatörle birlikte gelir.

3) Gelişmiş Analiz: Spark yalnızca 'Map' ve 'Reduce' ile değil. Ayrıca, SQL sorguları, Akış verileri, Makine öğrenme (ML) ve Grafik algoritmalarını destekler.

2.2.1. Spark Bileşenleri

Aşağıdaki şekil Spark'ın farklı bileşenlerini göstermektedir (Apache Spark Tutorial, 2017).



Şekil 5. Spark bileşenleri (Tutorialspoint, 2016)

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

Spark SQL, yapılandırılmış ve yarı-yapılandırılmış verilere destek sağlayan SchemaRDD adlı yeni bir veri soyutlaması getiren Spark Çekirdeğinin üst kısmında yer alan bir bileşendir.

Spark Akış, akış analitiklerini gerçekleştirmek için Spark Çekirdeğinin hızlı çizelgeleme özelliğini kullanır. Spark Akış, verileri küçük parçacıklar halinde alır ve bu küçük veri yığınları üzerinde RDD (Resilient Distributed Datasets-Esnek Dağıtılmış Veri Kümeleri) dönüşümleri gerçekleştirir.

Makine Öğrenme Kütüphanesi (Mlib), dağıtık bellek tabanlı Spark mimarisi nedeniyle Spark'ın üstünde dağıtılmış bir makine öğrenme çerçevesidir. Karşılaştırmalara göre, MLlib geliştiricileri tarafından Alternatif En Küçük Kareler (ALS) uygulamalarına karşı yapılır. Spark MLlib, Apache Mahout'un Hadoop disk tabanlı sürümünden dokuz kat daha hızlı (Mahout Spark arayüzünü kazanmadan önce) çalışmaktadır.

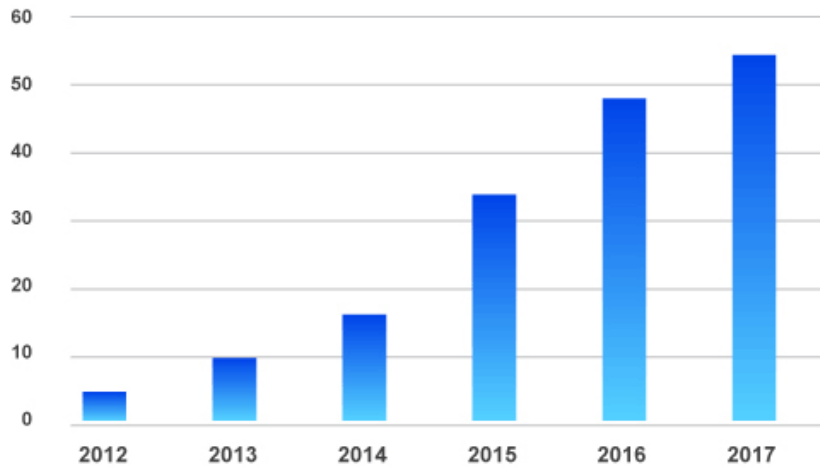
GrafikX, Spark'ın üstünde dağıtılmış bir grafik işleme çerçevesidir. GrafikX, Pregel soyutlama API'sını kullanarak kullanıcı tanımlı grafikleri modelleyebilen grafik hesaplamayı ifade etmek için bir API sağlar. Ayrıca, bu soyutlama için optimize edilmiş bir çalışma zamanı sağlar.

3. BÜYÜK VERİ ŞİRKETLERİ

Büyük verilerin ortaya çıkması, kuruluşların daha iyi rekabet etmesini sağladı. Büyük Veri, kuruluşlara süreçlerini ve sonuçlarını iyileştirmelerini sağlamanın yanı sıra, ürünlerini ve hizmetlerini rakiplerinden farklılaştırmalarına da yardımcı oluyor. Büyük Veri, şirketlerin iş büyümesini destekleyerek masraflardan tasarruf yapmalarını ve şirketlerin ürün geliştirme, pazarlama, satış, finans ve diğer alanlarda gerçek zamanlı bilgiye erişmesine imkan sağladı. Bu da şirketlerin daha iyi ve daha akıllı kararlar almalarını sağlayarak, onlara müşteri ilişkilerini geliştiren ve böylece şirkete rekabet avantajı sağlayan satışları arttırma ve müşteri davranış eğilimlerini belirleme konusunda daha fazla fırsat sundu. Dahası, veri odaklı stratejiler, yenilikçi, rekabet eden ve değeri yakalamak için kullanılabilir. Toplanan ve analiz edilen veriler gelecek ürünleri tasarlamak veya yeni hizmet teklifleri oluşturmak için kullanılabilir.

2015 yılına ait Forbes raporuna (Edureka, 2017) göre, küresel kuruluşların yaklaşık %90'ı, Büyük Veri Teknolojilerine orta veya yüksek düzeyde yatırım yaptığını ve üçte birinin yatırımlarının "çok önemli" olduğunu söylüyor. En önemlisi, katılımcıların yaklaşık üçte ikisi, veri ve analitik girişimlerin gelirler üzerinde belirgin ve ölçülebilir bir etkisi olduğunu söylemiştir. Mevcut durumda genişleyen veriler, bizi "Bilgi Ekonomisine" götürüyor. Bu ekonomi çok yönlü olarak toplanan bilgilere dayanıyor ve iş anlayışına dönüştürülerek gelir elde ediyor. Bu, ekonomi de yeni bir "Bilgi Katmanı" nın ortaya çıkmasına yol açtı. Ekonomi de gelir getiren yeni bulunan 'Bilgi katmanı', küresel ekonominin büyümesini de artırıyor. Küresel ekonomideki bu büyüme, milyonlarca yeni işe neden olacaktır. Gartner, 2015 yılına gelindiğinde, Büyük Veri talebinin tüm dünyadaki Bilişim Teknolojisi (BT) Endüstrisinde 4.4 milyon iş yaratacağını öngörüyor (Edureka, 2017). Gartner'a göre, sadece ABD'de 1,9 milyon BT işi yaratılacaktır. Büyük Veri'nin BT Endüstrisini nasıl doğrudan etkilediği büyük öneme sahiptir. Bu 1,9 milyon işe ek olarak, her BT işi de BT

dışı 3 iş olanağı yaratacaktır. Böylece önümüzdeki 4 yıl içinde bu yeni "Bilgi Ekonomisi" tarafından yaratılan 6 milyon ABD işgücünün büyük bir rakamına ulaşacaktır (Edureka, 2017). Gartner, ABD için Büyük Veri'nin ekonomik büyümeyi nasıl sürdüreceğini şekil 6'deki gibi öngörmektedir.



Şekil 6. ABD için Büyük Veri Pazarı Tahmini (Yüzde Milyon Dolar) (Edureka, 2017)

Bugün şirketlerin ayakta kalabilmesini sağlayan unsurların başında organizasyon esneklikleri ve değişime ayak uydurma gelmektedir. Bir firmanın diğer bir firmanın önüne geçmesini sağlayacak olan unsurların başında bir etkinliğin önceden tahmin ederek ona göre stratejiler belirlenmesidir (Oğan, 2014). Örneğin, Amazon kitap satışlarını öneri sistemleri (recomender systems) sayesinde %30 oranında arttırmıştır (Gürsakal, 2014). Lisa Arthur Forbes'teki bir makalesinde şunları söylüyor, eBay'ın küresel çapta 100 milyondan fazla aktif kullanıcısı bulunmaktadır. 2011 yılında bu şirket 68,6 milyar dolarlık bir satış yaptı. eBay da bunu diğer büyük şirketler gibi Büyük Veri sayesinde yapmıştır. Başka bir üst düzey eBay yöneticisi ise, '' Bir teknisyenin göremeyeceği desenleri veri setlerinde gördük. Bu sayede potansiyelimizin altında kullanılan sunucuları ve diğer etkinsizlikleri görerek milyonlarca dolar tasarruf etmemizi sağladı'' diyor (Lisa, 2012). Tractica analisti Bruce Daly'in belirttiği (Datafloq, 2017) gibi, "Verileri farklı düşünmek konusunda ilerleme gösteren şirketler Google ve Uber gibi ekonomiyi değiştiren şirketler".

Özellikle hızlı gelişen pazarlarda ve rekabet sebebiyle kar marjlarının daraldığı alanlarda inovatif iş modelleri şirketler açısından büyük öneme sahiptir. Çünkü inovasyon işletmelere fırsat yaratır ve rakiplerini tehdit eder. Facebook, bugün için dünyanın en popüler sitesi ama hiçbir içerik kendisinin değil. Alibaba.com, dünyanın en değerli perakendecisi ama stok tutmuyor hiçbir mağazası yok. Airbnb, dünyanın en büyük konaklama hizmetleri sağlayıcısı ama tek bir gayrimenkulü yok. Uber, dünyanın en büyük taksi şirketi ama tekbir aracı yok. uber'in doğduğu şehir San Francisco da şehrin elli yıllık en eski köklü taksi firmasının batmasını batırdı ve şehir şiddetli protestolara sahne oldu. Airbnb ve uber'in çok başarılı olmalarındaki ortak özellik her ikisinin de bir uygulama üzerinden bir paylaşım ekonomisi yaratmış olmalarıdır. Airbnb ve uber inovatif bir iş modeline başladı ama veri odaklı iş modeliyle herikisi de çok büyüdü. Uber 62,5 milyar dolar (Hook, 2016) değerindeyken airbnb 30 milyar dolar (Kara, 2017) değerindedir. NETFLIX ve iTunes içinde bulundukları sektörlerde gerçek bir yıkım etkisi yaratmış girişimlerdir. 2000'li yıllarda blokbuster fiziksel mağazalarından filim ve video kiraladığımız bir bir imparatorluktu ve NETFLIX'in internetten video girişimi blokbuster'ın iflas etmesine sebep oldu. iTunes, CD piyasasında taşları yerinden oynatarak aynı yıkımı müzik sektöründe de yarattı. Buradan şu sonuçlara ulaşmak mümkün. Bugün aslında birçok endüstri yenilikçi, teknolojik odaklı, müşteriyi iyi tanıyan veri odaklı şirketlerin tehdidi altına girmeye başladı. Bu nedenle bugünden organizasyonların veri odaklı düşünmeye başlamaları ve veri kültürünü bütün organizasyona yaymayı hedeflemeleri gerekiyor.

4. SONUÇ VE DEĞERLENDİRME

Bugün Büyük Veri Yönetimi her zamankinden daha fazla önem kazandı. Bunun temel nedeni işletmelerin geçmişe göre daha çok veriye önem vermesidir. Böyle bir etkiye neden 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 kaynaklanmakta. 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 kaldı. Söz konusu bu baskılar, Büyük Verileri daha doğru, açık ve güvenilirlikte yönetmek için giderek artan bir zorunluluk yarattı. Bugün tüm bu sorunları çözmek için Hadoop gibi Büyük Veri analizi yapan teknolojiler geliştirilmiştir. Hadoop ilk ortaya çıktığında hızla büyük verilerle özdeşleşmeye başladı ancak yine de istenen seviyeye gelemedi ve Büyük Veri projelerinin birçoğu başarısız oldu. Çünkü Hadoop'un yapısı idrak edilen ve düşünülenden çok daha karmaşıktı. Hadoop, MapReduce ile bu sorunların üstesinden gelmeye çalıştı ama başarısız oldu. Fakat şuan Spark sayesinde Hadoop büyük ölçüde bu sorunların üstesinden geldi. Hadoop sayesinde şirketler bugün hem toplu işleme hem de veri akışı için oldukça başarılı analizler yaparak sorunlara ekonomik ve genel amaçlı çözümler sunmakta. 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 sermaye oranlarını büyük ölçüde arttırmakta. Bu çalışmada Büyük Veri Teknolojilerinin yapıları anlatılarak, bu teknolojilerin şirketler açısından önemi üzerinde durulmuştur. Sonuç olarak, şirketlerin rakipleriyle gelecekte mücadele edebilmesi ve ayakta kalabilmesi için Büyük Veri Teknolojilerine daha fazla yatırım yapmaları gerekir.

KAYNAKLAR

Apache Hive, (2017). https://hive.apache.org/, (Erişim: 25.05.2017)

Apache Kafka, (2017). https://www.tutorialspoint.com/apache_kafka/index, (Erişim: 15.03.2017)

Apache Pig Tutorial, (2017). https://www.tutorialspoint.com/apache_pig/apache,(Erişim: 10.06.2017)

Apache Spark Tutorial, (2017). www.tutorialspoint.com/apache_spark/, (Erişim: 10.05.2017)

Apache Sqoop, (2017). http://sqoop.apache.org/, (Erişim: 10.04.2017)

Brad, H., (2017). "Understanding Hadoop Cluster and the Network", /01.01.2017/ http://bradhedlund.com/2011/09/10/understanding-hadoop-clustersand-the-network/, (Erişim: 15.10.2017)

Confluence, (2016). cwiki.apache.org/confluence/display/Hive/HCatalog/, (Erişim: 20.02.2017)

Databricks, (2016). https://databricks.com/spark/about/20160102/, (Erişim: 10.05.2017)

Datafloq, (2017). https://datafloq.com/read/in-brief-what-is-datamonetization/, (Erişim: 12.04.2017)

Datafloq, (2017). https://datafloq.com/read/unlimited-computing-swarmsensors-algorithms-world/2138, (Erişim: 20.06.2017)

Dezyre, (2017) https://s3.amazonaws.com/files.dezyre.com/images/Tut orials/HDFS, (Erişim: 18.01.2017)

Dezyre, (2017). https://www.dezyre.com/article/hadoop-componentsand-architecture-big-data-and-hadoop-training/114/, (Erişim: 10.07.2017)

Edureka, (2017). www.edureka.co/blog/bigdatatutorial?utm_source, (Erişim: 20.06.2017)

Edureka, (2017). https://cdn.edureka.co/blog/wpcontent/uploads/2014/03 /, (Erişim: 12.07.2017)

Edureka, (2017). www.edureka.co/blog/5-reasons-to-learn-hadoop, (Erişim: 19.07.2017)

Edureka, (2017). https://www.edureka.co/blog/big-data-leads-economic growth, (Erişim: 19.07.2017)

Facebook, (2017). "Facebook for Developers" https://developers.facebook.com/ /, (Erişim: 15.02.2017)

Gates, A., (2011). "Data flow scripting with Hadoop: Programming Pig", First Edition. 1005 Gravenstein Highway North, Sebastopol, CA 95472.: O'Reilly Media.

Gürsakal N., (2014). ''Büyük Veri'' Genişletilmiş 2. Baskı, Dora, Bursa, ISBN:978-605-4798-803.

Hook, L., (2016). "saudi wealth fund takes $3.5bn Uber stake", JUNE 1, 2016, https://www.ft.com/content/a7e31c58-282c-11e6-8b18-91555f2f4fde, (Erişim: 20.10.2017)

Hortonworks, (2017). https://hortonworks.com/apache/hive/#section_5, (Erişim: 10.06.2017)

Kara, M., (2017). "Airbnb, 30 milyar dolar değerlemeden sonra karlığa ulaştı", 27 Ocak 2017, https://webrazzi.com/2017/01/27/airbnb-30-milyar-dolar-degerlemeden-sonra-karliliga-ulasti/ (Erişim: 20.10.2017)

Lisa, A., (2012). "Big Data Pushback? 86 Percent of Americans Vote No On Tailored Political Ads".

McKinsey Global Institute, (2011). "Big Data Report", https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf, (Erişim: 10.02.2016)

Mohammad, H. F., (2011). ''Data Intensive Query Processing For Semantic Web Data Using Hadoop And MapReduce'', Thesis, The University Of Texas At Dallas, s.7.

Mukhammadov, R., (2013). ''A scalable database for a remote patient monitoring system'', Master of Science Thesis, School of Information and Communication Technology (ICT) KTH Royal Institute of Technology Stockholm, Sweden.

Murthy, A., Markha, J., Vavilapalli, V. K., and Eadline, D., (2014). "Apache Hadoop YARN". https://www.puurdata.nl/wpcontent/uploads/Apache.Hadoop.YARN _.Sample.pdf, (Erişim: 10.09.2017)

Oğan, Ö., (2014). ''Büyük Veri Denizi'', Elma yayınevi, Ankara, syf. 21.

Ohlhorst F., (2013). "Big data analytics: turning big data into big Money", New Jersey.

Robert, R., (2006). "Representational State Transfer (REST)", In Pro PHP XML and Web Services, 633–72. Springer.

Slidershare, (2016). slidehshare.net, (Erişim: 12.01.2016)

Thomas, D., (2014). "big data @ work, Türk Hava Yolları Yayınları", Çeviren: Müge Çavdar, syf.127.

Tutorialspoint, (2016). https://www.tutorialspoint.com/apache_spark/images/components_of_spark. (Erişim: 10.10.2016)

Vaccari, C., (2014). "Big Data in Official Statistics", University of Camerino, Doktora of Philosophy in İnformation Science and Complex System-XXVI Cycle, School of Science and Technology.

Vavilapalli, V., (2013). "Apache hadoop yarn: Yet another resource negotiator", in Proceedings of the 4th annual Symposium on Cloud Computing. ACM.

WebHCat, (2016). https:/cwiki.apache.org/confluence/display/Hive/WebHCatUsingWebH, (Erişim: 17.01.2016)