BÜYÜK VERİ ANALİZİ

DR.ÖĞR.ÜYESI BETÜL AY

BÜYÜK VERİ NEDİR?

- Web tıklama geçmişi, mobil cihazlar üzerinden online etkileşimler gibi internet üzerinden ve tweets, Facebook güncellemeleri, Youtube videoları gibi sosyal medya sitelerinden üretilen veri miktarı IBM'e göre günlük yaklaşık olarak 2.5 kentilyondur.
- Twitter 2010 yılında günlük ortalama 35 milyon tweet sayısına sahipken bugün, 500 milyon üzerinde kayıtlı kullanıcıya ve günlük ortalama 400 milyon tweet sayısına ulaşmıştır
- Geleneksel veri tabanı sistemleri ve yazılım teknikleri kullanarak işlenmesi zor olan böylesi geniş veri kümeleri "büyük veri" olarak adlandırılmaktadır.
- Büyük veri kavramı aynı zamanda bu verilerin depolanması, sorgulanması ve analizi için geliştirilen teknolojileri de ifade etmektedir.

5V İLE BÜYÜK VERİ

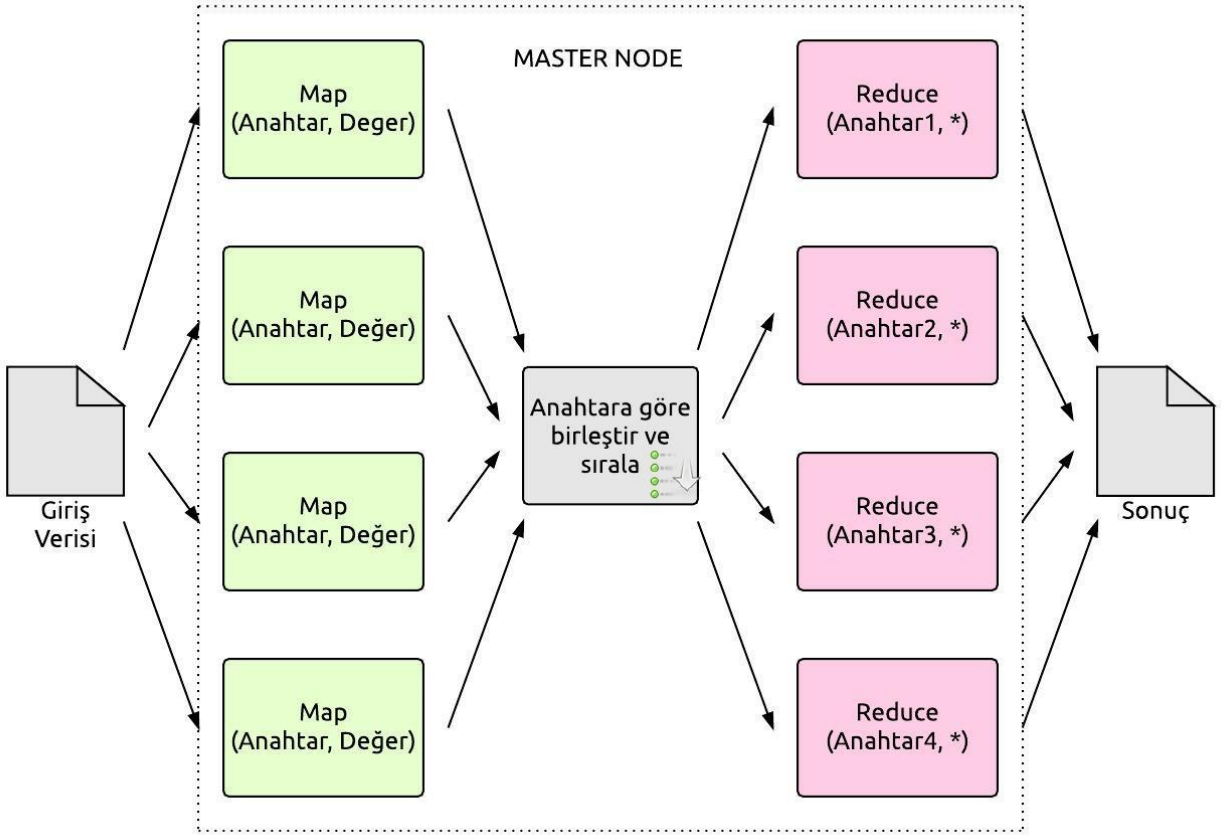
- Büyük veri çeşitlilik, hacim, hız, doğruluk ve değer gibi 5V (variety, volume, velocity, veracity, value) olarak adlandırılan çeşitli sorunlar nedeniyle pek çok zorluğu beraberinde getirmektedir.

Büyük Veri Özellikleri	Tanımı
Variety(Çeşitlilik)	Sosyal medya yorumları, videolar, fotoğraflar ve mesajlar gibi farklı şekillerde yapılandırılmamış veriler
Volume (Hacim)	Özellikle internet teknolojilerinin gelişimi ile katlanarak büyüyen veriler
Velocity (Hız)	Verinin nasıl hızla üretildiği ve nasıl hızla analiz edilmesi gerekliliği
Veracity (Doğruluk)	Önemli kararlar için veri güvenirliğinden emin olma
Value (Değer)	Veri analitiği için kullanılmak üzere farklı kurumlar tarafından depolanan verinin değeri

-

MAPREDUCE PARADİGMASI

- Büyük veriyi işlemek için 5V ile açıklanan temel sorunlar, kurumlar tarafından toplanan tüm verilerden en önemli olanları filtreleme ve böylesi geniş miktarda verileri etkili bir şekilde analiz eden algoritmaları tasarlamaktır.
- Büyük veriyi işlemek için pek çok dağıtık çözümler sunulmuştur.
- Bu çözümlerden en yaygını, Google tarafından geliştirilen MapReduce yaklaşımıdır.
- **MapReduce** dağıtık mimari üzerinde çok büyük verilerin kolay bir şekilde analiz edilebilmesini sağlayan bir sistemdir.



MAPREDUCE PARADİGMASI-WORD COUNT ÖRNEĞİ

- Bu uygulama, kendisine verilen bir metin dosyası içerisinde hangi kelimenin ne sıklıkla geçtiğini bulmaktadır.
- Map aşamasında her bir dosyadaki kelimelerin sayılma işlemi gerçekleşmektedir.
- Reduce aşaması ise, map aşaması sonucunda çıkan değerleri birleştirerek bir kelimenin verilen tüm metin dosyalarında ya da dosya kümelerinde ne sıklıkla geçtiğini sayarak sonuç üretmektedir.

4 Functions are used :

- Map (the mapper function)
- EmitIntermediate(the intermediate key,value pairs emitted by the mapper functions)
- Reduce (the reducer function)
- Emit (the final output, after summarization from the Reduce functions)

```
Map(String anahtar, String değer)
//anahtar: metin dosyasının ismi
//değer: dosya içerikleri
for each kelime w in değer:
    EmitIntermediate (w, "1");

Reduce(String anahtar, String değer)
//anahtar: bir kelime
//değerler: sayıların listesi
int sonuc -0;
for each v in degerler;
sonuc += ParseInt(v);
    Emit(AsString (sonuc));
```

MAPREDUCE ÖZELLİKLERİ

- Farklı makinelere dağıtılan map görevleri, diğer map görevleri ile iletişim halinde değildir; aynı durum reduce görevleri için de geçerlidir.
- Verinin kopyaları ayrı makinelerde tutulduğu için, bir makinenin donanımsal ya da yazılımsal herhangi bir nedenle çökmesi durumunda görevler devam etmektedir.
- Ayrıca map işlemleri diğer map işlemlerini beklemek zorunda değildir, bir map görevinin arkasından reduce görevi başlayabilir yani sıralı bir şekilde çalışmak zorunda değildir.
- Böylelikle MapReduce veri depolama sisteminden ya da programlama dilinden bağımsız, yüksek derecede ölçeklenebilir, basit ve hata toleransı (fault tolerance) yüksek olan bir uygulama çatısı sunar.

MAPREDUCE AVANTAJLARI

Büyük ölçekli veri analizi için MapReduce tabanlı sistemlerin gittikçe yaygınlaşmasının pek çok sebebi vardır. Bu sebeplerden bir kaçı aşağıda sıralanmıştır:

- ***MapReduce arayüzü basit fakat etkileyicidir.*** MapReduce sadece map ve reduce fonksiyonlarını kullansa da, SQL sorgulama, veri madenciliği, makine öğrenmesi, graf işlemeyi içeren sayısız veri analitik görevleri MapReduce ile yerine getirilebilir.
- ***MapReduce esnektir (flexible).*** MapReduce depolama sistemlerinden bağımsız tasarlanmıştır. Analiz edilecek tüm veriyi bir yerden başka bir yere taşımak yerine verilerin olduğu sistem üzerinde küçük analizler gerçekleştirilmektedir ve yapılandırılmış ya da yapılandırılmamış farklı çeşitlerdeki tüm veriler analiz edilebilmektedir.
- ***MapReduce ölçeklenebilirdir (scalable).*** Paylaşılan bir kümede binlerce düğüm üzerinde MapReduce çalışabilirken, sistemde bir hata oluştuğu zaman sadece hatanın olduğu düğümdeki görevleri tekrar çalıştırarak hata toleransını sağlayabilmektedir.

HADOOP AÇIK KAYNAK PROJESİ

- MapReduce paradigması, farklı lokasyonlarda bulunan büyük ölçekli veri kümeleri üzerinde veri işleme problemlerini çözmek için ortaya çıkmıştır. Büyük veri işlemek için popüler bir paradigma olan MapReduce, Hadoop gibi pek çok açık kaynak projelerinde uygulanmıştır.
- Hadoop, iki temel bileşenden oluşmaktadır: Yarn ve HDFS.
- **Yarn (Yet Another Resource Negotiator)**, bir Hadoop kümesi üzerinde çalışan uygulamaları saklamakta ve CPU, hafıza yönetimi sağlamaktadır. Hadoop'un ilk jenerasyonu yalnızca MapReduce uygulamalarını çalıştırabilirken Yarn, Hadoop'un yanısıra Spark gibi diğer uygulama platformlarının çalışmasına da olanak sağlamaktadır.
- **HDFS (Hadoop Distributed File System)**, veri depolama için bir Hadoop kümesinde bütün düğümlere verileri dağıtan bir dosya sistemidir. Yüksek hız ile büyük miktardaki veriye erişim sağlayabilen bu dağıtık dosya sistemi olarak, bir çok makinedeki dosya sistemlerini birbirine bağlayarak tek bir dosya sistemi gibi kullanılabilmelerine olanak sağlamaktadır.

APACHE HADOOP İLE TWİTTER VERİLERİNİN ANALİZ EDİLMESİ

- Ayşe adlı bir kullanıcı pek çok kişiyi takip etmektedir ve aynı şekilde Ayşe de pek çok kişi tarafından takip edilmektedir.
- Ayşe bir tweet attığında, yani bir düşüncesini yazdığında tüm takipçileri tarafından bu tweet görülebilmektedir.
- Aynı zamanda Ayşe başka kullanıcıların tweet'lerini de retweet yapabilmektedir. Retweet, bir e-postayı iletmek gibidir.
- Ayşe, Fatmadan bir tweet görürse, onu retweet eder ve Ayşe'nin takipçileri Fatma'nın tweet'ini görür.

GELENEKSEL ÇÖZÜM:

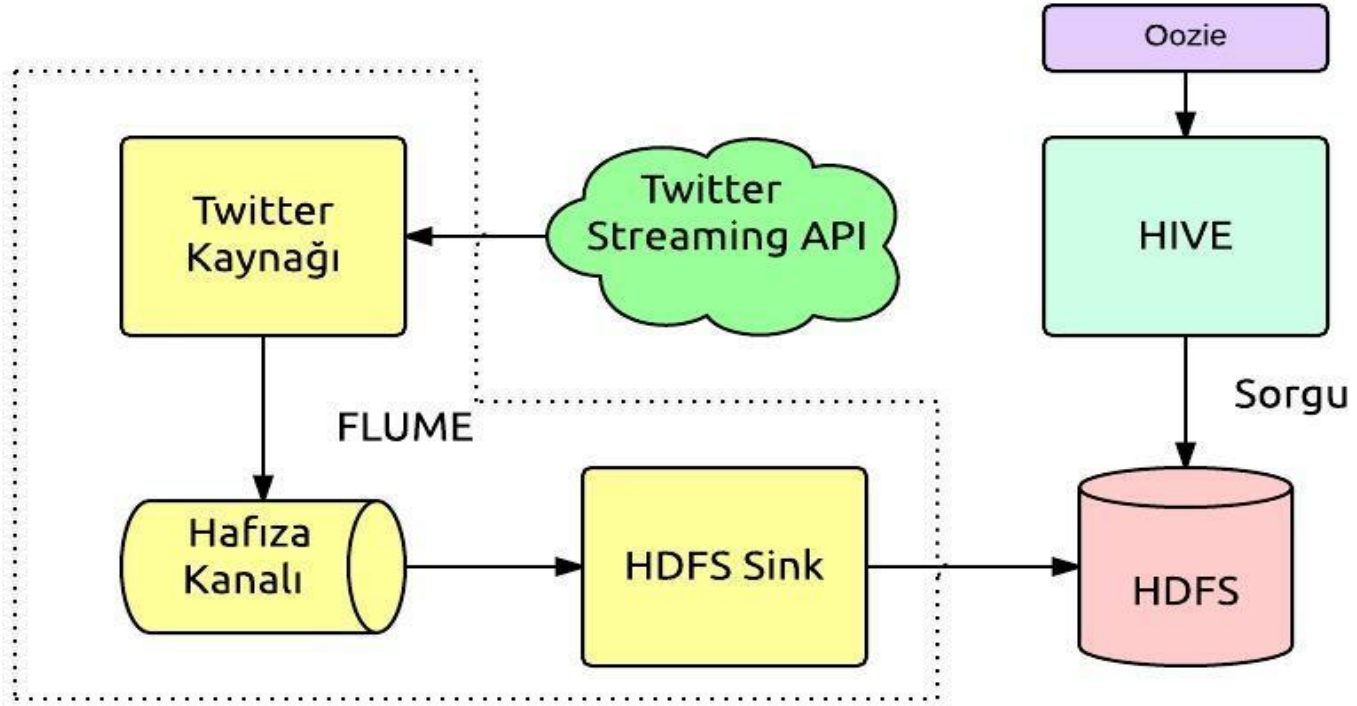
1. Twitter, tüm tweet'ler için retweet sayısını tuttuğu için, kullancı tweet'lerini analiz ederek istediğimiz kullanıcıya ulaşabiliriz.
2. Ayrıca en çok retweet alan kullanıcı kim olduğunu bularak istenilen alanda en popüler kişiyi tespit edebiliriz. SQL sorguları bu soruların cevabı için kullanılabilir.
3. Retweet'ler azalan şekilde sıralandığında, en çok retweet'i alan kişiye bakabiliriz.

APACHE HADOOP İLE TWİTTER VERİLERİNİN ANALİZ EDİLMESİ

PROBLEM:

- Twitter Streaming API tweet'leri karmaşık gelebilen bir JSON (JavaScript Object Notation) formatında verdiği için geleneksel ilişkisel veri tabanı yönetim sistemleri bu işlem için yetersiz kalmaktadır.
- Hadoop ekosisteminde, Hive [32] projesi Hadoop dağıtık dosya sisteminde bulunan verileri sorgulamak için kullanılabilen bir sorgu arayüzü sağlamaktadır ve HiveQL adı verilen SQL'e benzer bir dil kullanarak veriyi sorgulamaktır.

APACHE HADOOP İLE TWİTTER VERİLERİNİN ANALİZ EDİLMESİ



APACHE HADOOP İLE TWİTTER VERİLERİNİN ANALİZ EDİLMESİ

1. **Apache Flume:** Hadoop'un Cloudera açık kaynak dağıtımını (CDH) kullanarak HDFS (Hadoop dağıtık dosya sistemi) içerisine veriyi getirmenin bir yoludur. Flume büyük veriyi etkili bir şekilde toplayan ve götüren dağıtık, güvenilir ve erişilebilir bir servistir. 3 temel yapıdan oluşur: kaynak, kanal ve sink.
2. **HDFS Sink:** Bir kanal üzerinden gelen tweet'leri dağıtık dosya sistemine yazmak için kullanılır. Hadoop dağıtık dosya sistemine veriler yüklendikten sonra, ilk adım Hive üzerinde harici bir tablo yaratarak verileri sorgulamaktır.
3. **Apache Oozie:** İş akışı (workflow) koordinasyon sistemidir. Oozie iki temel bölümden oluşmaktadır: MapReduce, Pig [35], Hive ve benzeri farklı Hadoop işlerinden oluşan iş akışını çalıştıran ve saklayan bir iş akışı motoru (workflow engine) ve önceden tanımlanmış programlar üzerinde iş akışı işlerini çalıştıran bir koordinatör motoru (coordinator engine).

HADOOP AVANTAJLARI

- Verinin hangi formatta olduğundan ve hangi uygulama altında çalıştığından bağımsız bir şekilde yeni veri düğümleri eklenebilmektedir (ölçeklenebilirlik).
- Hadoop, sunucular arasında etkili bir paralel hesaplama gücü sağlayarak daha fazla verinin depolanmasına ve maliyet açısından daha büyük kazanç elde edilmesine olanak tanımaktadır (maliyet etkinliği).
- Farklı dağıtık kaynaklardan gelen yapısal veya yapısal olmayan tüm veri tipleri Hadoop ile işlenebilmekte ve analiz edilebilmektedir (esneklik).
- Donanımsal ya da yazılımsal bir hata sebebiyle bir veri düğümü düştüğü zaman sistem veri işlemeye devam etmektedir (hata toleransı).

HADOOP DEZAVANTAJLARI

- Bir mapper fonksiyonundan sonra reducer fonksiyonu çalıştırmak zorunda kalınan bir model ile sınırlıdır. Dolayısıyla ihtiyaç duyulduğunda çok fazla mapper ve reducer ile büyük veriyi işlemek zorunda kalmaktadır.
- Sunucu kümeleri üzerinde hata ayıklama (debug), yazılım dağıtımı ve loglama gibi işlemleri gerçekleştirmek ve yönetmek zordur.
- Sık kullanılan karmaşık iteratif (yinelemeli) algoritmaları desteklemez. Iteratif algoritmalarda, bir değerin hesaplanması önceki hesaplanan değerlere bağlı olarak değişir ve map-reduce görevleri birbirinden bağımsız bir şekilde çalıştığı için MapReduce kullanılamaz.
- Bir MapReduce platformunda makine öğrenmesi algoritmaları uygulandığında, her iterasyon sonucu diske yazılır. Diskten okuma ve yazma gibi iteratif işlemler ise sistemi yavaşlatmaktadır.