Hafta 10 (NLP - Big Data - PySpark)

@mebaysan

07/11/2021

İlgili Okuma Listesi:

> https://s3-us-west-2.amazonaws.com/secure.notion-static.com/e11c3eca-6c52-4375-9fe9-a69d8544a350/nlp.pdf

Benim Yazdığım Yazılar:

-

Python ile Büyük Veri

Büyük Veri Nedir

İlişkisel veri tabanlarında işlenmesi zor olan durumlarda büyük veri ortaya çıkıyor.

Veri işlemeye yeni bir vizyondur.

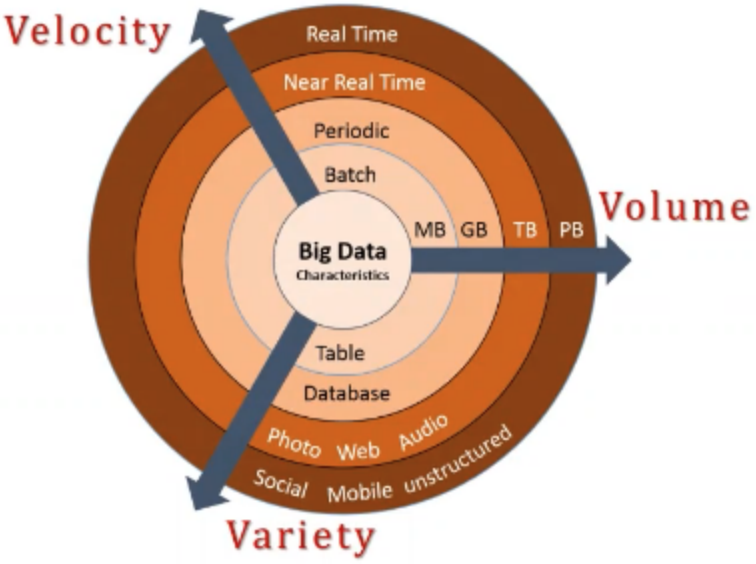
Geleneksel yöntemlerle işlenemeyen verilere büyük veri denir.

Bu problemin çözümü birden fazla bilgisayarın bir araya gelerek tek bir bilgisayar gibi hareket etmesidir. Yani bir işi yapmak için birden fazla bilgisayarın tek bir bilgisayar gibi davranmasıdır.

Veri analitiği alanında yeni ufuklar açtı, hesaplama gücünün artmasıyla makine öğrenmesi algoritmalarının performansları arttı. Daha büyük miktar ve çeşitteki verilerin kullanılması ile veriden faydalı bilgi çıkarma süreci için çok önemli bir kaynak/araç sağlanmış oldu.

Büyük Verinin Bileşenleri

- Ortaya çıkan veri işleme güçlüğü verinin hacmi, çeşitliliği ve hızı ile alakalıdır.
- Büyük veriyi ifade eden özellikler hacim, çeşitlilik ve hızdır.
- Büyük veri araçları veriden faydalı bilgi çıkarma süreçleri için çok güçlü bir araçtır.



Apache Hadoop

Açık kaynak kodlu, güvenilir, ölçeklenebilir paralel hesaplama yazılımı projesidir.

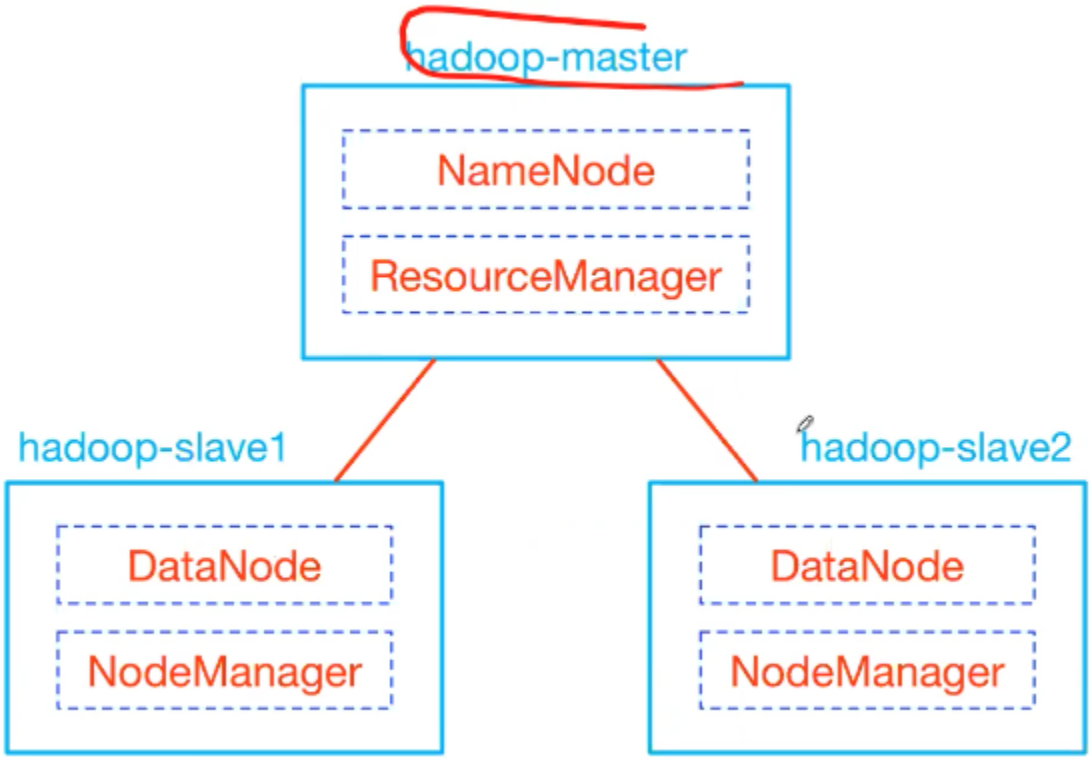
- Büyük veri teknolojilerinin temelini oluşturur
- Geleneksel yöntemler ile etkin olarak işlenmesi mümkün olmayan verilerin işlenebilmesine olanak sağlamaktadır.
- Bir bilgisayar kümesinin belirli bir işi yapmak için tek bir bilgisayar gibi hareket etmesini sağlamaktadır.

Apache Hadoop Bileşenleri

Temel bileşenler:

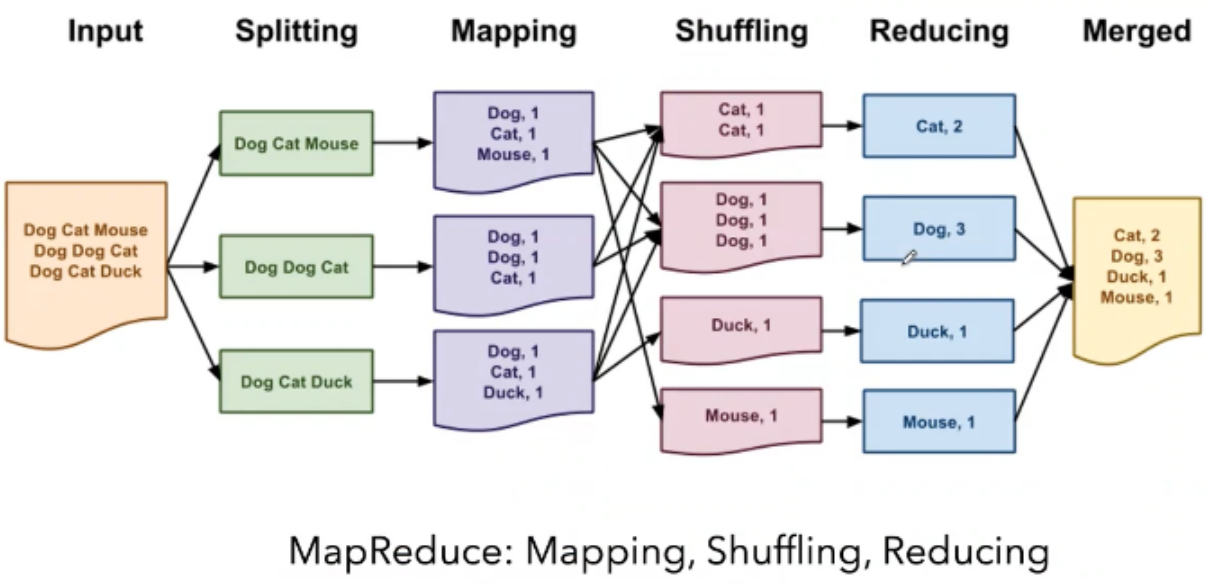
- Hadoop Common
- Hadoop Dağıtık Dosya Sistemi (HDFS)
- Hadoop YARN
- Hadoop MapReduce

Apache Hadoop Küme Yapısı



MapReduce Nedir?

Paralel işleme yapısını yerine getiren programatik bileşendir.



MapReduce: Mapping, Shuffling, Reducing

- Bazı kaynaklarda Split ve Shuffle basamakları gösterilmez
- Split basamağında veri parçalara ayrılır
- Map basamağında görev cihazlara dağıtılır
- Shuffle basamağında birbirinze benzer öğeler bir araya getirilir
- Reduce basamağında benzer öğeler sayılır
- Merge basamağında hangi öğeden kaç tane var sayılır || dağıtık makinalardaki veri birleştirilir

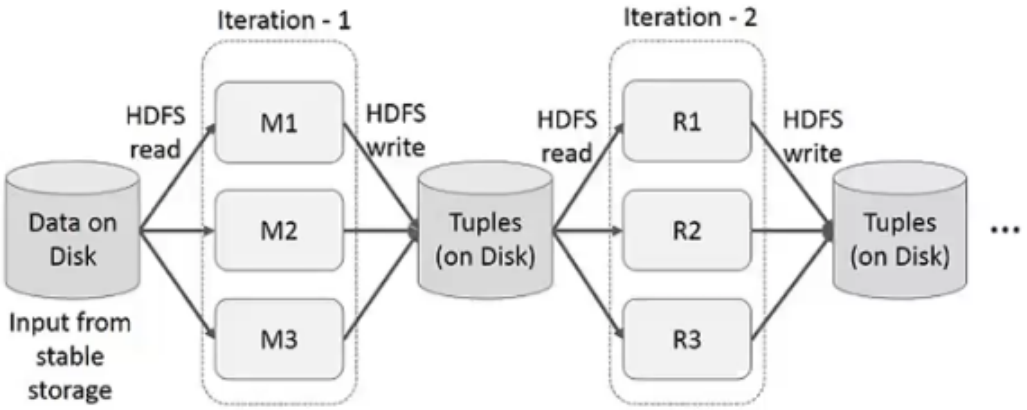
Apache Hadoop Faydaları

- Veri Saklama ve İş Gücü
- Açık Kaynak
- Hız
- Esneklik
- Ölçeklenebilirlik

- Hata Toleransı

Apache Hadoop Çözüm Olarak Yeterli mi?

- Disk tabanlı çalışan bir modeldir
- Her MapReduce görevinde diskten okuma ve diske yazma işlemi yapılır
- İteratif işlemler zaman alır ve kaynakları meşgul eder



MapReduce Disk Bazlı Çalışma Modeli

Apache Spark

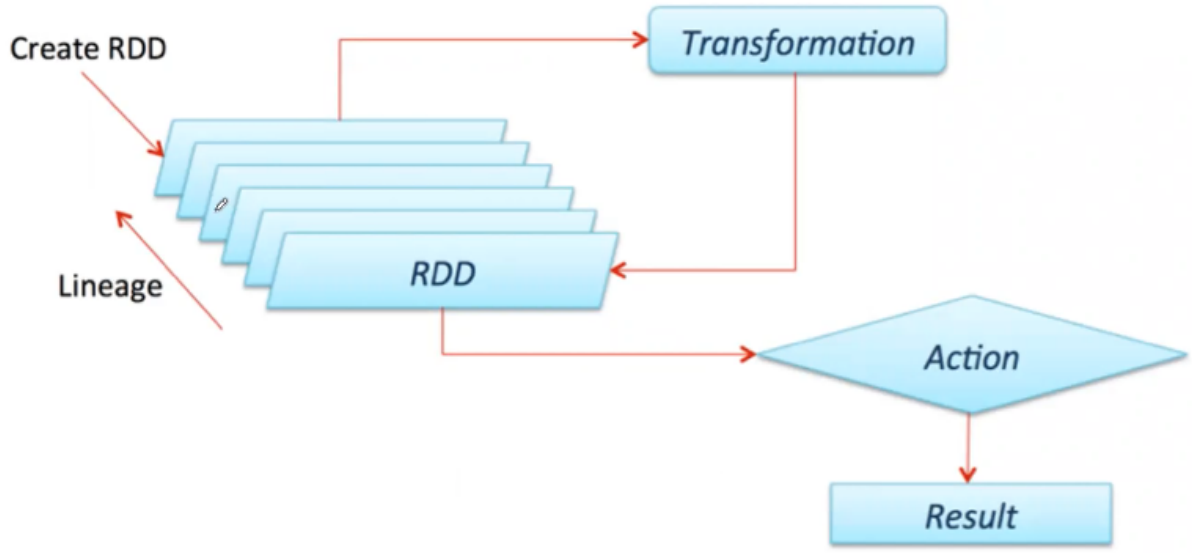
- Küme üzerinde hızlı ve genel amaçlı bilgi işleme sistemidir.
- Hadoop'un alternatifi değildir.
- MapReduce modelinde yer alan disk bazlı çalışma sisteminin yarattığı maliyetlerden dolayı ortaya çıkmıştır.

- Apache Hadoop'a göre 100 kat daha hızlı çalışmaktadır.
- Java, Scala, Python ve R ile uygulama geliştirilebilir.
- Genelleştiricidir: Spark SQL, Spark MLlib, Spark Streaming, GraphX aynı uygulamada kullanılabilir.

Apache Spark'ın Bileşenleri

- Spark Core ve RDD's
- Spark SQL
- Spark MLlib
- Spark Streaming
- GraphX

Dayanıklı Dağıtık Data Setler (RDDs)



Apache Hadoop'un programatik veri işleme yöntemi MapReduce'dur. Apache Spark tarafında bunu yapan RDDs'dir.

- Veri bellekte tutularak (ram) diskten okuma problemi çözülüyor

MapReduce ile RDDs Karşılaştırılması