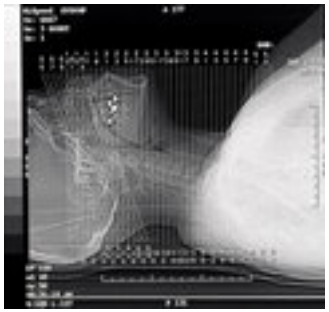
Büyük Veri

Yrd. Doç. Dr. Özgür Yılmazel
Gökhan Çapan
Anadolu Üniversitesi

Büyük Veri Nedir?

- Büyük veriler her yerde
  - Bilimsel hesaplamalar
  - Medikal görseller
  - Web Sunucu log dosyaları



- Ama büyük veri ne kadar büyük?
  - Boyut her zaman önemli mi?

Büyük Veri Ne Kadar Büyük?

- CERN – 40 TB/ Saniye

- 12,233 Twitter posts / saniye – Superbowl maçında

Büyük Veri Ne Kadar Büyük?

Yeri geldiğinde:

- 40MB Powerpoint – Distributed to hundreds of people via email
- 1TB Medikal Görüntü eş zamanlı konsultasyon alımında kullanılmalı
- 1PB 3 Boyutlu filmin derlenmesi

Boyutu kadar Ne Yapılacağı da Önemli

Büyük Veriyi tanımlayan özellik, sistemin kapasitesini veya iş isterlerini zora sokan tüm özelliklerden birisi olabilir.

- Verinin oluşma veya geliş hızı
- Kaynakların çeşitliliği ve sayısı
- Tüm büyük veriler eşit yaratılmamış
  - Yapısal Veriler
  - Yapısal Olmayan Veriler

Büyük Veri

- Yapısal Veriler – ilişkisel veri tabanları, banka transactionları, alış veriş geçmişleri
- Yapısal Olmayan Veriler – Bloglar, email, sosyal medya, sensor bilgileri, fotograflar

- Otomatik Oluşturulan Veriler
- Şahısların Oluşturduğu Veriler

Ne Yapmalı?

Ne Yapmalı?

Veri Toplama:

- Bir çok farklı tipde veri
  - Yapısal
  - Yapısal Olmayan
  - XML
  - Metinsel
  - Görseller
- Yüksek hız – bilinen geçikme – Saniye de 12Bin Tweet
- Yüksek sayıda veri
- Esnek Veri Yapıları

Ne Yapmalı?

Düzenle:

- Yüksek verim – çıkış hızı
- Anlık ve Yerinde düzenlemeler
- Tüm kaynaklara ve tüm yapılara uyum

Ne Yapmalı?

Yorumlama:

- Derinlemesine Analiz
- Çevik Geliştirmelere uyumlu
- Çok büyük ölçeklenebilirlik
- Gerçek Zamanlı Sonuçlar

Ne Yapmalı?

- Neredeyse her zaman paralel
- Onlarca yüzlerce hatta bazen binlerce bilgisayara dağınık durumda

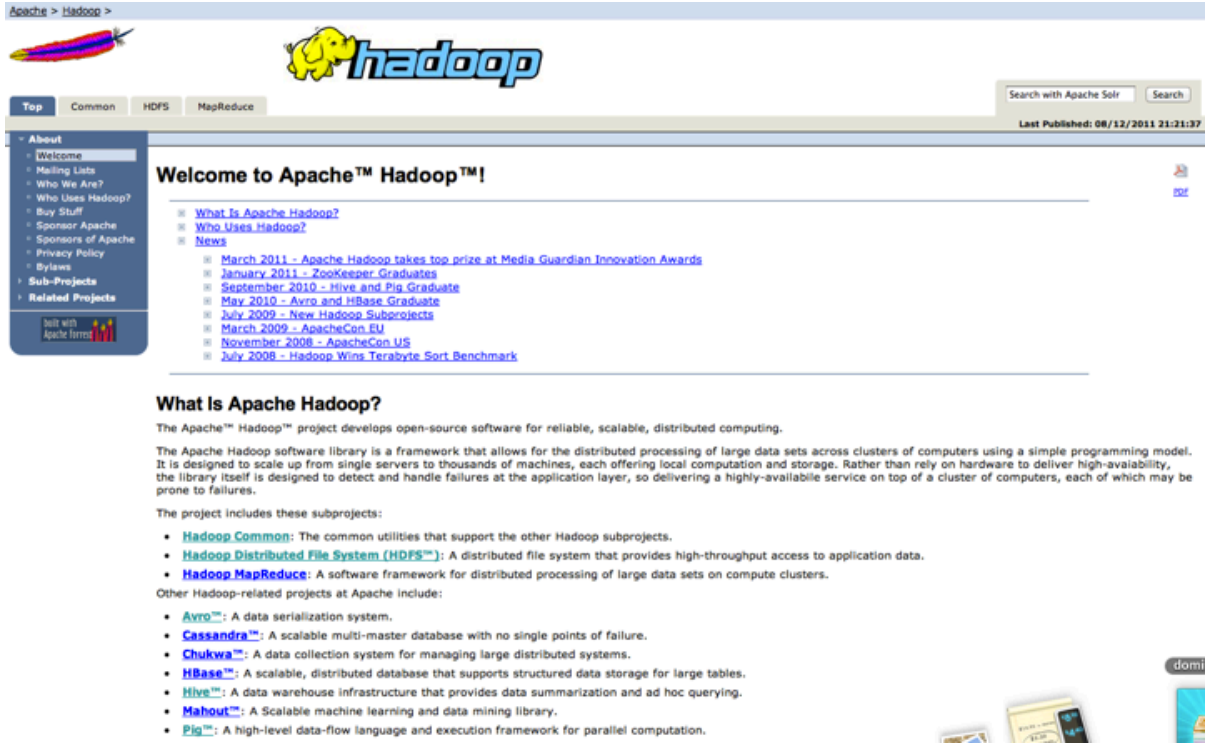
Kimler Yapıyor?

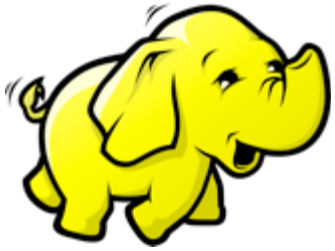
Nasıl Yapalım?

Tamamı açık kaynak kodlu projeler:

- Hadoop
- HDFS
- MapReduce
- Pig
- Hive
- HBase
- Mahout

Hadoop





- Açık kaynak kodlu
- Dağıtık Hesaplama
- Çok basit bir model MapReduce
- Bir birine bağlı bir grup bilgisayar

Örnek Hadoop Kurulumları

- Ebay – 532 node 532x8 Core – 4256 Core
- Facebook – 1100 node 8800 core 12 PB Storage
- LinkedIn – 1200 + 580 + 120 node
- Quantcast – 3000 node 3.5 PB 1PB+ günlük veri

Kaynak:Hadoop wiki – powered by

Hadoop

- Bileşenleri:
  - Veri Saklama - HDFS – dağıtık disk dosya sistemi
  - Veri İşleme - MapReduce

HDFS

- Google File System
- Dagitik bir dosya sistemi
- Dosyaların bir küme üzerinde dağıtılmalarından sorumlu
- Dosya boyutları genelde gigabytelar seviyesinde
- Hata toleransli
- Replikasyon
- Yatay olceklenebilir – HDFS fiziksel lokasyon konusunda bilgi sahibidir.

HDFS

- Hadoop Shell aracılığıyla, Java API veya Web UI ile ulaşılabilir
- Dört tip node:
  - NameNode – dosya sisteminin meta verisini yönetir
  - Job Tracker – MapReduce
  - Task Tracker – MapReduce
  - Data Node – Verilerin saklandığı NameNode tarafından adreslenen nodelardır

MapReduce

- Büyük Veri İşleme - Google
- Dağıtık Hesaplama Modeli
- Anahtar – Değer ikilisi ile veri işleme
- Kolay bir programlama çerçevesi (framework)
  - sadece map() ve reduce() fonksiyonlarını geliştirmeniz yeterli

Map: İlk Adım

- Elemanları bir anahtar ile eşleştirip bir sonraki işleme hazır hale getirir.
  - Veri temizleme
  - Basit bir hesap yapma
  - Virgül veya TAB ile ayrılmış diziyi parçalama

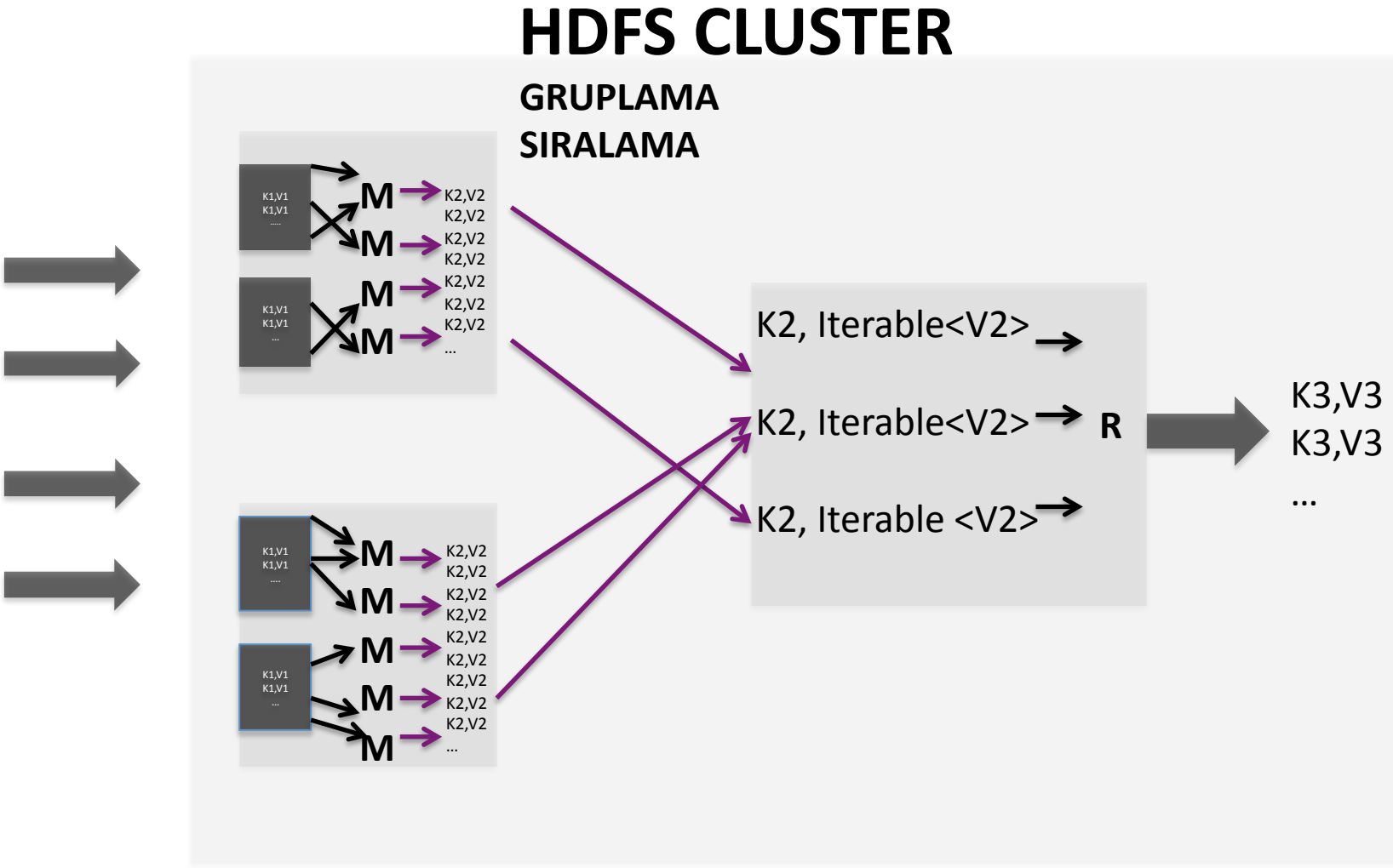
Reduce: Son Adım

- Iterator ile aynı anahtar için bir değerler listesi alır ve bunları
  - Biriktirerek
  - Filtereleyerek
  - Örnekleyerek

  azaltır.

Sonuçlarını HDFS veya HBASEe yazar

MapReduce - Mimari



Apache Pig

- Yahoo!
- Büyük veri setlerinin kolay analiz edilmesi için tasarlanmış
- Javada Map Reduce yazmaktan daha basit
  - Pig Latin dilinde kodlanabilir
  - Geliştirilebilir
- Diğer scripting dillerine çok benzer
- 10 Satırlık bir Pig kodu yüzlerce satır Java koduna denk gelebilir
- Sıklıkla hızlı prototip çıkarmak için ve anlık sorgulamalarda kullanılır

Pig Koşturmak!

- Grunt – Shell
- Java arayüzü
- Eclipse ve IntelliJ IDEA pluginleri

```
%tweets = load ‘/today/tweets’ as (user,
mention, tweet)
%twitters = group tweets by mention
```

Apache Hive

- Hadoop için Veri Ambarı Projesi
- Veri özetleme
- Anlık sorgulamalar
- SQL benzeri bir dil –HiveQL
  - Özel mapper ve reducer tanımlamalarına olanak
- Hive Compileri SQL sorgularını MapReduce operasyonlarına çevirir
- Hadoopun karmaşıklığından son kullanıcıyı kurtarır.

HBase

Her zaman ilişkisel veri tabanlarına ihtiyacımız yok

Ölçeklenebilirlik ihtiyacımız var

Tablo boyutlarımız çok büyük olabiliyor

Çok hızlı erişim istiyoruz

Distributed Key-Sorted Persistent Map

HBase

Google – Big Table klonu

HDFS üzerinde çalışır

- hata toleransı
- ölçeklenebilirlik
- MapReduce girdi çıktısı

Hbase = HDFS + rastgele okuma (Random read/ write)

HBase

Nerelerde Kullanılır:

- Sosyal Medya
- Tavsiye Sistemleri
- Arama Motorları
- İstihbarat ve İzleme Servisleri
- Finansal Sistemler – dolandırıcılık - sahtekarlık

Apache Mahout

- Basit analizlerden öteye geçelim

- Sınıflandırma – Email Spam – Çağrı Merkezi
- Kümeleme – Yeni haber bulma
- Tavsiye Sistemleri

Apache Mahout

- Ölçeklenebilir Hadoop üzerinde çalışan veri madenciliği ve yapay öğrenme kütüphanesi
  - Yapay öğrenme çalışanları petabytelar seviyesinde veri ile uğraşmıyorlar
  - Ticari uygulamalar hazır değil
  - Gerçek problemlerin verileri büyük
    - Milyonlarca kredi kartı başvurusu
    - Milyonlarca telefon konuşması
    - Milyarlarca kullanıcı davranışı logu

Apache Mahout

- Veriden Bilgiye ulaşmak
  - Veri hazırlama
  - Model oluşturma
  - Bilgiye ulaşma
- Mahout tüm adımlar için araçlar sağlar
- Hadoop üzerinde çalışır
- Hbase veya HDFS veri giriş çıkışı için kullanılabilir.

Apache Mahout

Hazır Algoritmalar:

- Sınıflandırma:
  - Logistic Regression
  - Bayesian
  - Random Forests
- Uygulama
  - Metin Sınıflandırma
  - Çağrı yönlendirme
  - Yüz tanıma

Apache Mahout

Hazır Algoritmalar:

- Kümeleme:
  - kMeans
  - Canopy
  - Mean Shift
  - MinHash
  - Latent Dirichlet Allocation (LDA)
  - Uygulama
    - Yeni haber bulma
    - Dolandırıcılık sahtekarlık bulma

Apache Mahout

Hazır Algoritmalar:

- Tavsiye:
  - Distributed Item based
  - Matrix Factorization
  - Non distributed item/user based
- Uygulama
  - eTicaret - Amazon
  - Film tavsiyesi – NetFlix
  - Müzik tavsiyesi – LastFM
  - Mekan Tavsiyesi - FourSquare

Nerede Ne Kullanılmalı?

- Veri Toplamak:
  - Flume - cloudera
  - Scribe – facebook
- Veri Saklamak:
  - HDFS
  - HBASE
- Veri Analizi
  - MapReduce
  - Pig
  - Hive
- Akıllı Uygulamalar
  - Mahout

Teşekkürler