CS585/DS503
Big Data Management

Introduction & Logistics

WPI, Mohamed Eltabakh

Theme of this Course

Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

- How to manage very large amounts of data and extract value and knowledge from them

Introduction to Big Data

What is Big Data?

What makes data, "Big" Data?

Data Management Applications

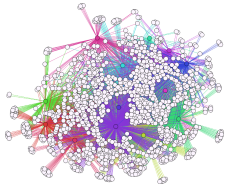
Banking

Retail Sys

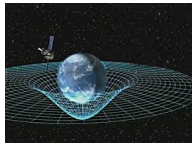
Airlines

Physics

Biology



Graph Data



Spatio-Temporal

Streaming

Social Media

Big Data

Traditional

Scientific and More advanced

Big Data

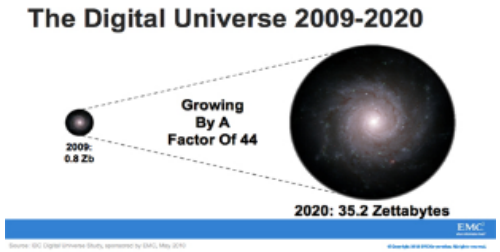
Big Data Definition

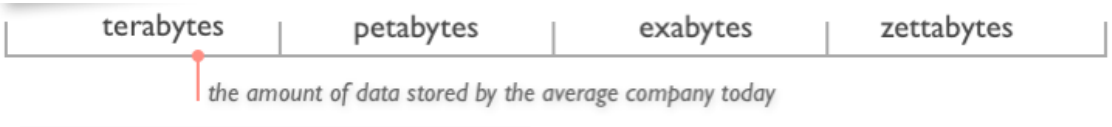
- No single standard definition…

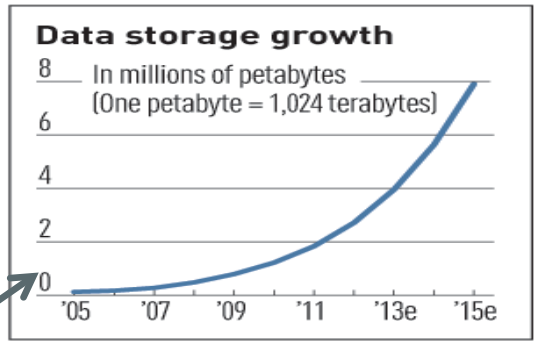
"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

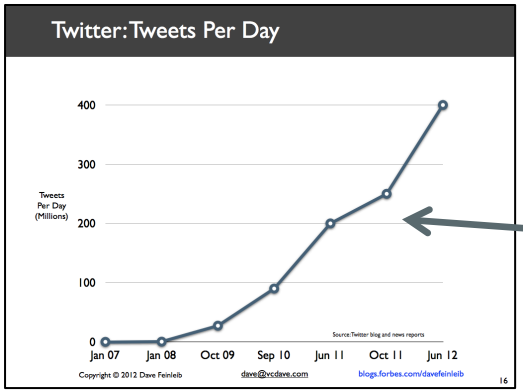
Characteristics of Big Data: 1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





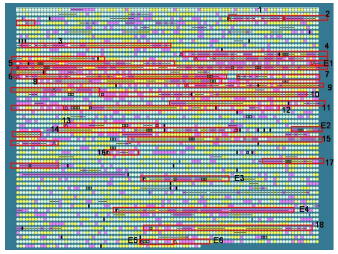


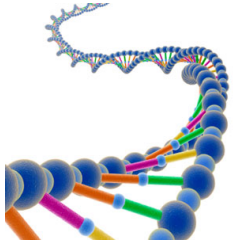


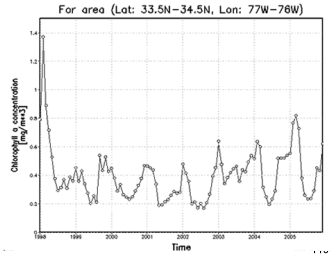
Exponential increase in collected/generated data

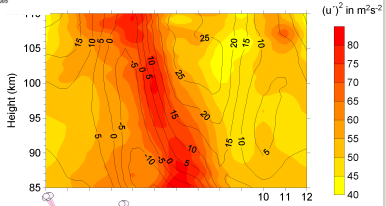
Characteristics of Big Data: 2-Complexity (Varity)

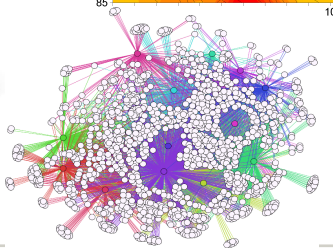
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data











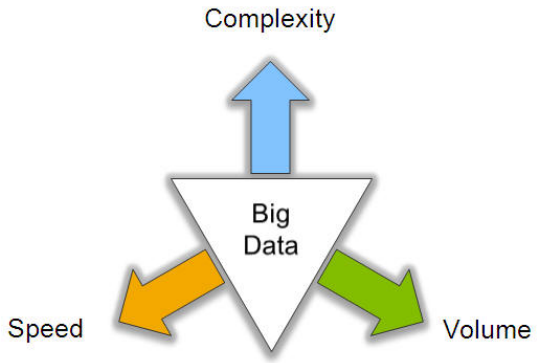
To extract knowledge➔ all these types of data need to linked together

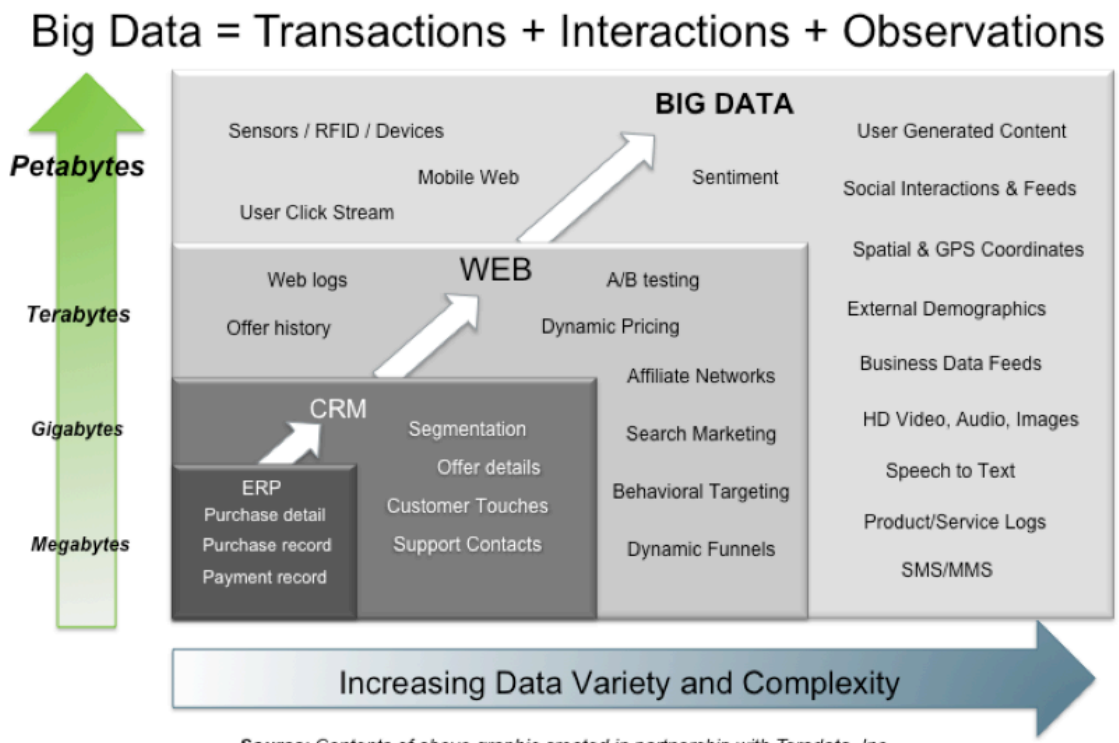
Characteristics of Big Data: 3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

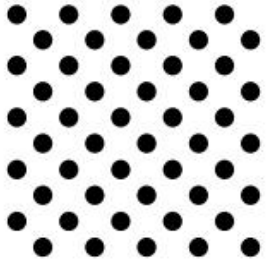
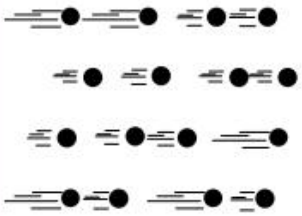
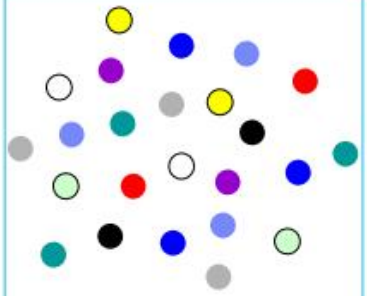
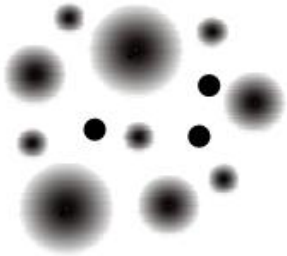
Big Data: 3V's



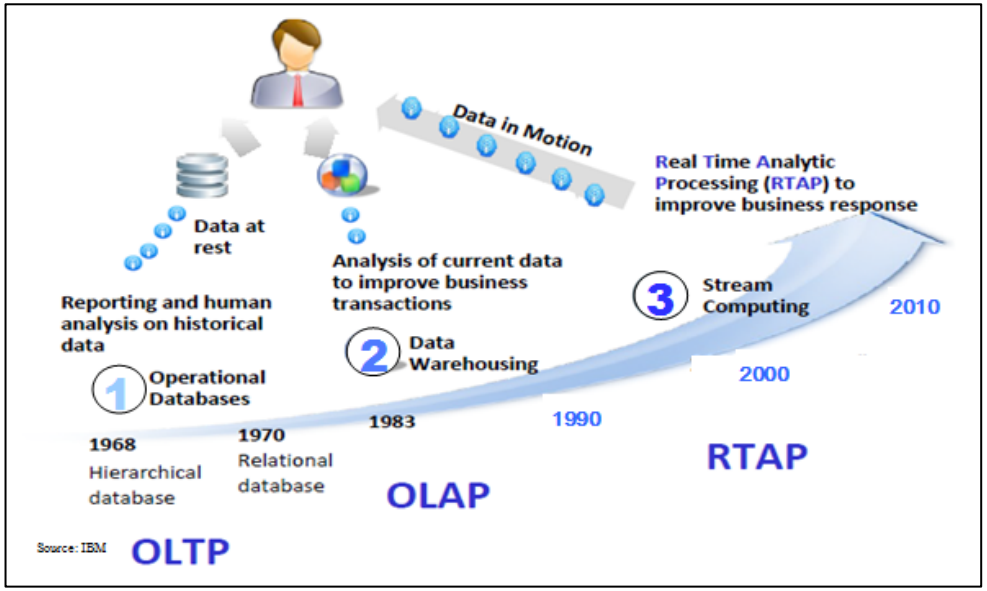


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Data Explosion

2 Billion **Internet users by 2011**

1.3 Billion RFID tags in 2005

30 Billion **RFID tags by 2010**

4.6 Billon
Mobile Phones World Wide

Capital market data volumes grew ***1,750%,*** **2003-06**

twitter

Twitter process ***7 terabytes*** **of data every day**

World Data Centre for Climate
- ***220 Terabytes*** **of Web data**
- ***9 Petabytes*** **of additional data**

facebook

Facebook process ***10 terabytes*** **of data every day**

Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

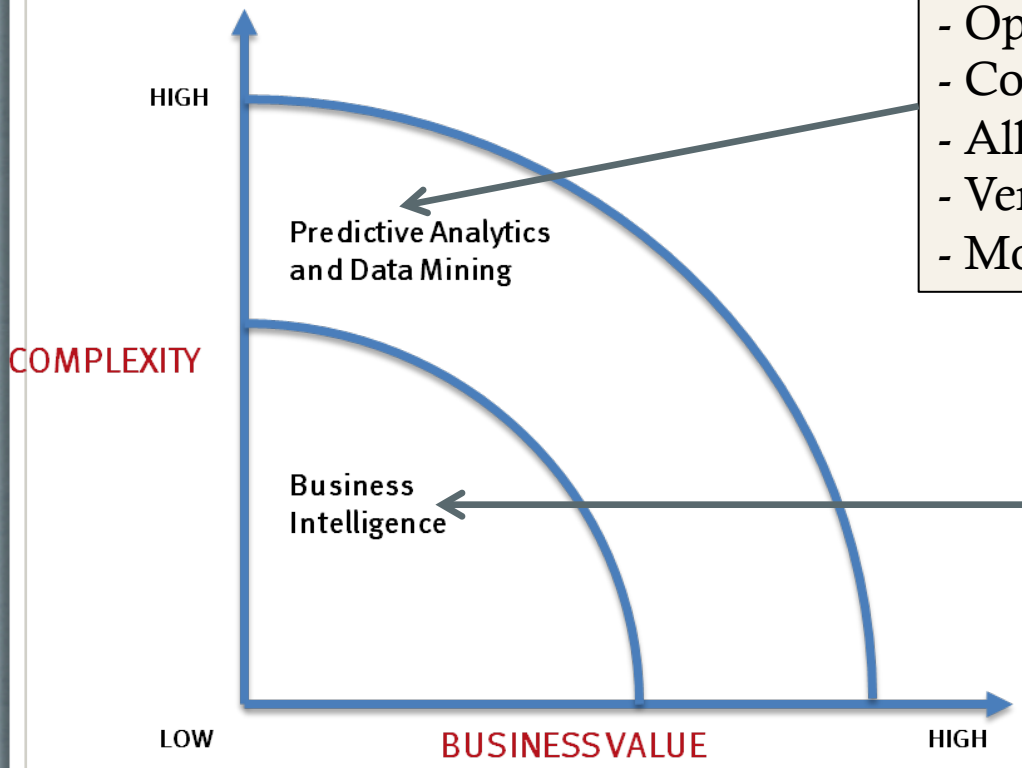
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data

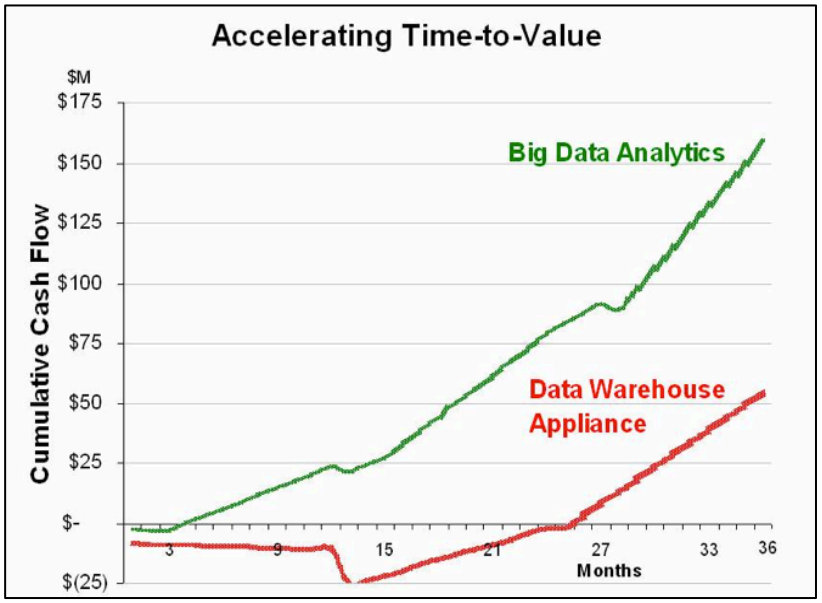


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

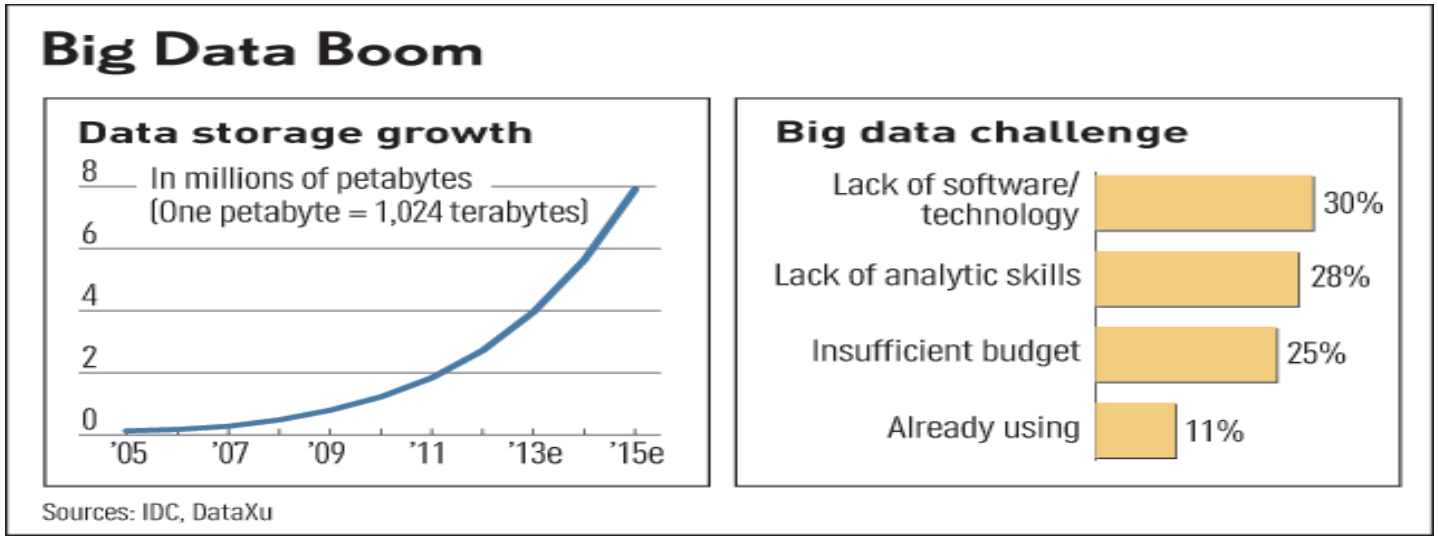
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

What Technology Do We Have For Big Data ??

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO
KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service

factual
kaggle
knoema
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

COUCHBASE
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

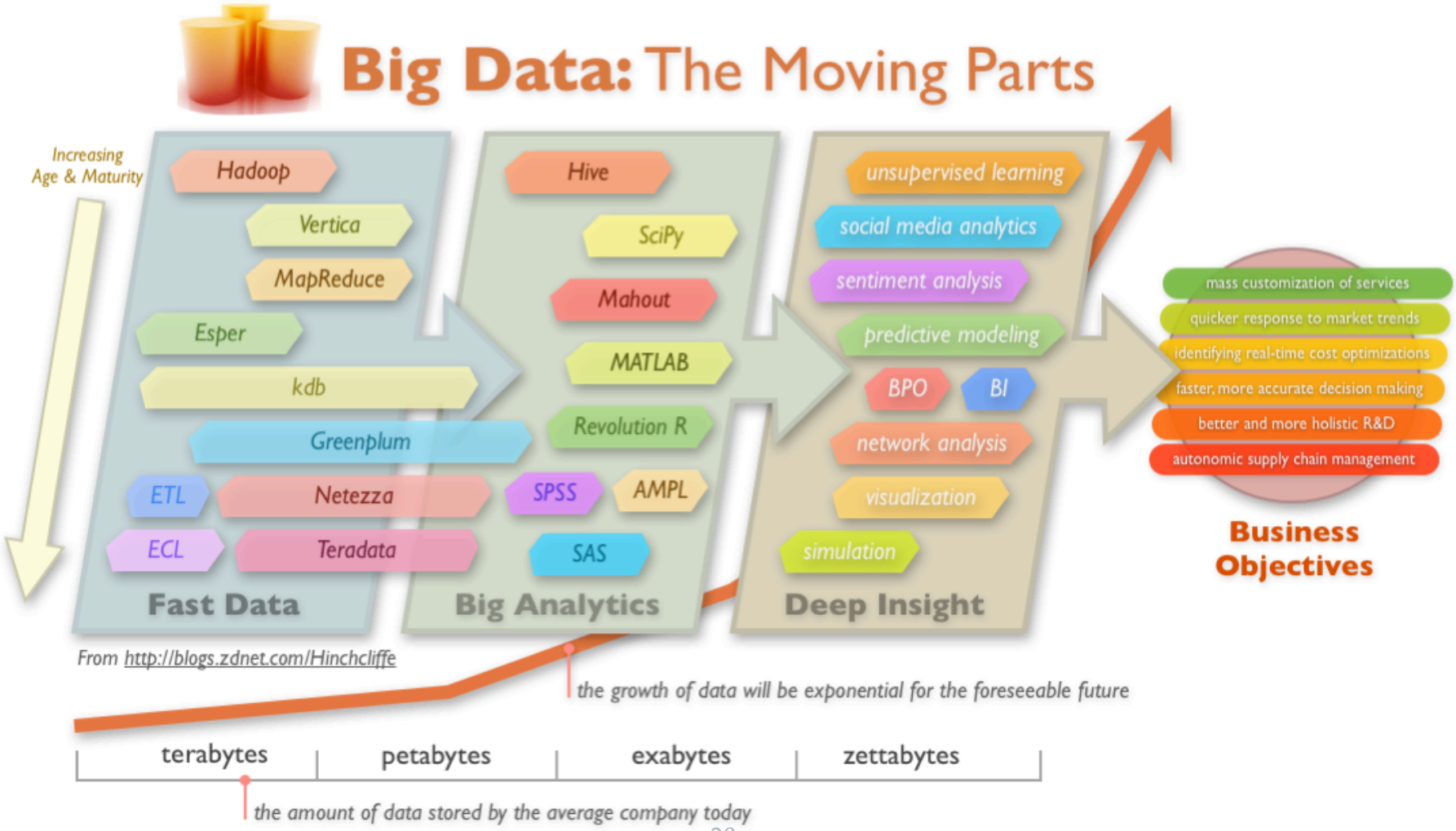
Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Technology



What You Will Learn

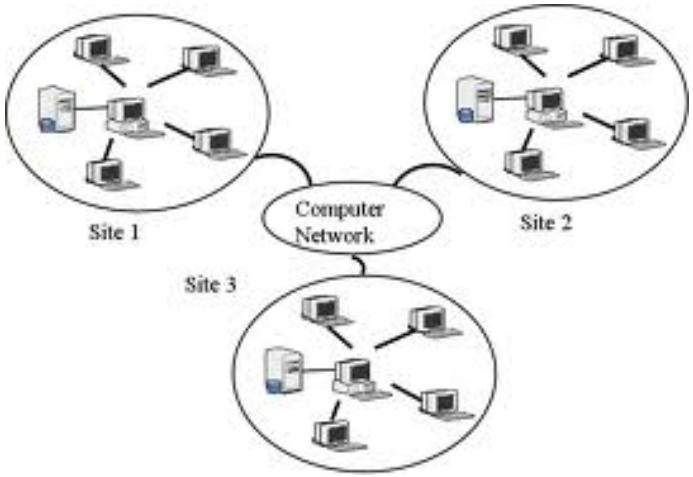
MapReduce / Hadoop

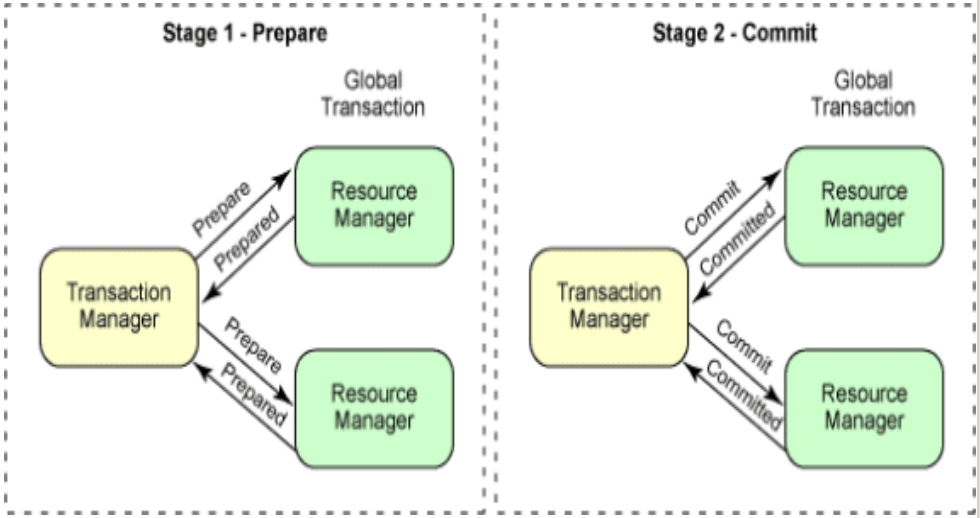
- **Learn the platform (how it is designed and works)**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs in different languages**
  - Programming Languages: Java, C, Python
  - High-Level Languages: Apache Pig, Hive
- **Learn advanced analytics tools on top of Hadoop**
  - RHadoop: Statistical tools for managing big data
  - Mahout: Data mining and machine learning tools over big data
- **Learn state-of-art technology from recent research papers**
  - Optimizations, indexing techniques, and other extensions to Hadoop

What You Will Learn

Distributed DBMSs

- **Differences from traditional DBMSs**
- **Data and Query Models**
- **Several operators and transaction management**

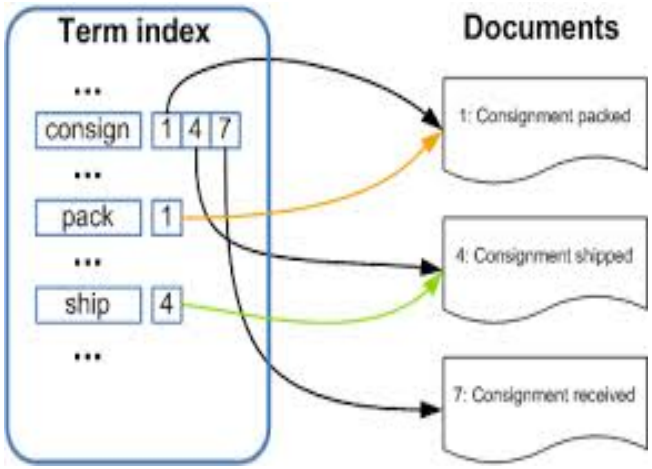




What You Will Learn

NoSQL DBs

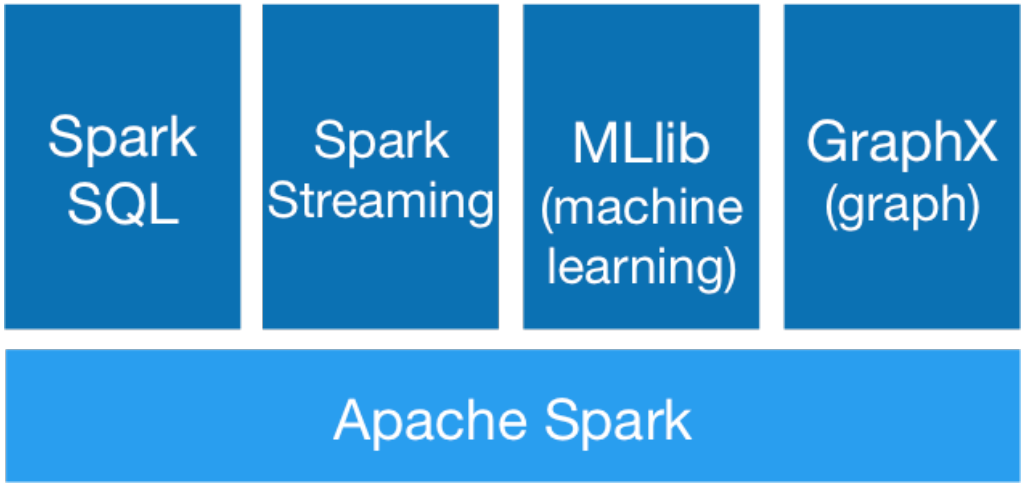
- **MongoDB and Document DBs**
- **Lucene Index**



What You Will Learn

Spark

- **In-Memory Data Processing**



What You Will Learn

State-of-art in Research

- **Covering many research papers**
- **Presentations/discussion/quizzes**

Course Logistics

Course Logistics

- Web Page: http://web.cs.wpi.edu/~cs585/s17/
- Electronic WPI system: blackboard.wpi.edu
- Lectures
  - Thursdays: (6:00 - 8:50pm)
    - Divided into 2 sub-lectures 80, 80, & 10 mins break

Textbook & Reading List

- **No specific textbook**
  - Big Data is a relatively new topic (so no fixed syllabus)
- **Reading List**
  - We will cover the state-of-art technology from research papers in big conferences
  - Many Hadoop-related papers are available on the course website
- **Related books:**
  - Check the course website

Requirements & Grading

- **Seminar-Type Course**
  - Students will read research papers and present them (Reading List)
- **Hands-on Course**
  - Several coding projects covering the entire semester
- **Assignments**
  - Coding Projects
  - Presentations

  } ***Done in teams of two***
- **4 Quizzes**
- **Final Exam**

Late Submission Policy

- **For Projects**
  - One-day late → 10% off the max grade
  - Two-day late → 20% off the max grade
  - Beyond that, no late submission is accepted
  - **Submissions:**
    - Submitted via blackboard system by the due date

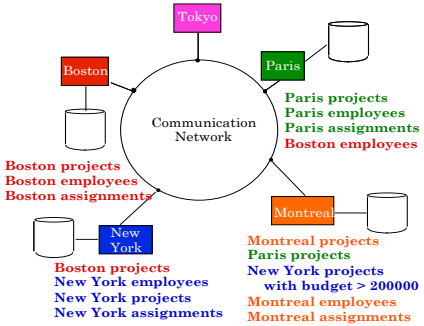
More about Projects

- **A virtual machine is created including the needed platform for the projects**
  - Ubuntu OS
  - Hadoop platform
  - Apache Pig/ Hive
  - Mahout library (Version 0.7)
  - Rhadoop
  - MongoDB
- **Need Virtual Box (Vbox)** [free]

Next Step from You…

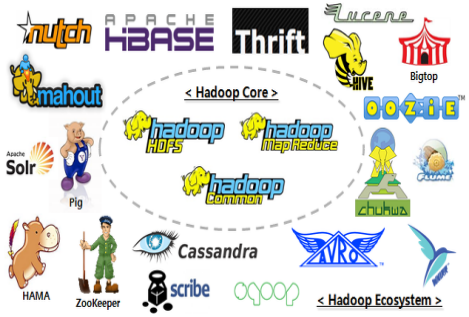
1. Form teams of two
2. Visit the course website
3. Plan ahead:
    1. Decide what/when you want to present??
4. Use Blackboard "Discussion" forum for posts or for searching for teammates

Summary of What You Will Learn



Distributed DBs

+



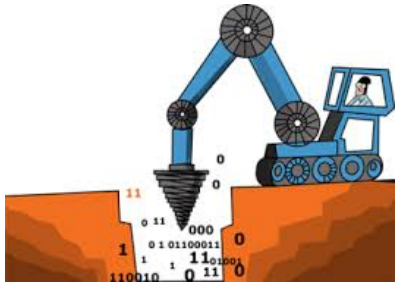
MapReduce & Hadoop

+

NoSQL: NotOnly SQL

APACHE Spark™

+



Big Data Mining & Analytics

+

Presenting & Reading Research Papers

Prerequisites

- **Good knowledge about database systems**
  - CS4432 (DB II) or CS542
- **Programming skills**
  - Java, Python, Scala