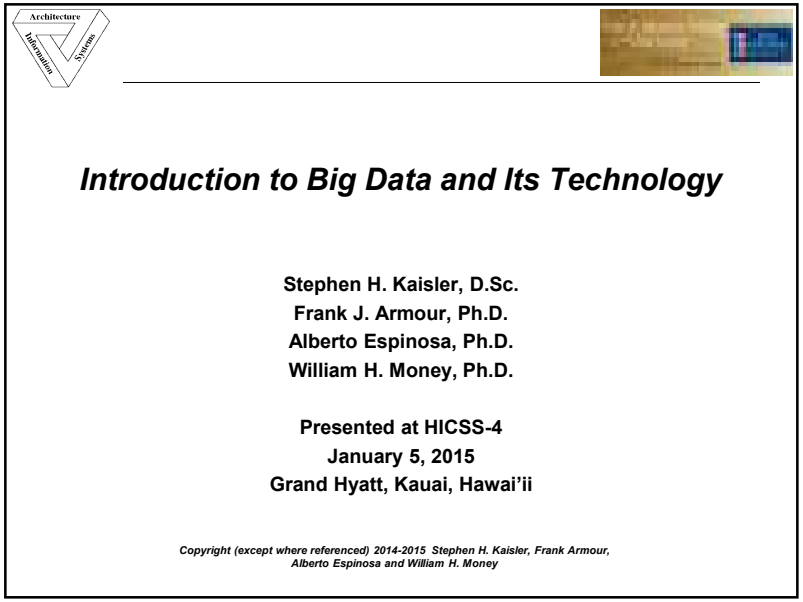
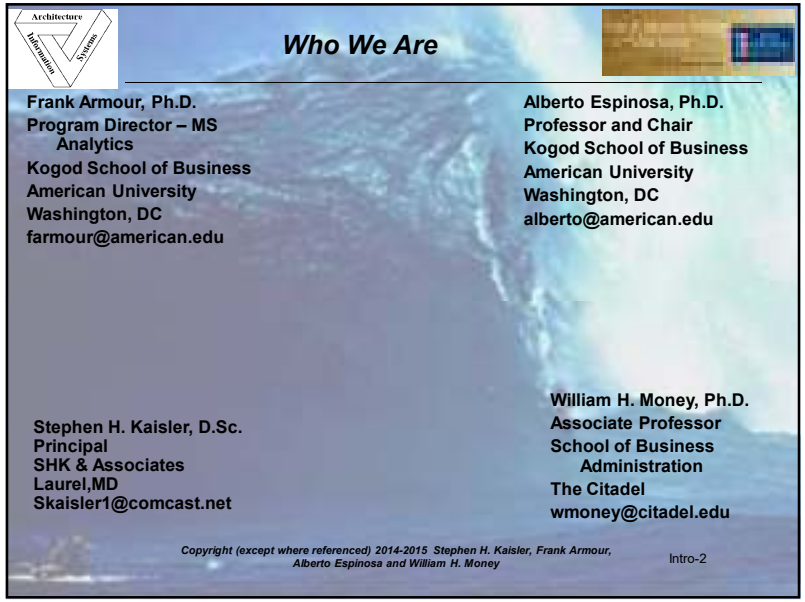


Introduction to Big Data and Its Technology

Stephen H. Kaisler, D.Sc.
Frank J. Armour, Ph.D.
Alberto Espinosa, Ph.D.
William H. Money, Ph.D.

Presented at HICSS-4
January 5, 2015
Grand Hyatt, Kauai, Hawai'ii



Who We Are

Frank Armour, Ph.D.
Program Director – MS Analytics
Kogod School of Business
American University
Washington, DC
farmour@american.edu

Alberto Espinosa, Ph.D.
Professor and Chair
Kogod School of Business
American University
Washington, DC
alberto@american.edu

Stephen H. Kaisler, D.Sc.
Principal
SHK & Associates
Laurel,MD
Skaisler1@comcast.net

William H. Money, Ph.D.
Associate Professor
School of Business Administration
The Citadel
wmoney@citadel.edu

Outline

Topic	Schedule
Big Data: Introduction Overview of Big Data Technology	0900-1015
Break	1015-1045
Analytics Knowledge and Process Framework Big Data Cases and Examples Big Data Challenges	1045-1200

Big Data: Introduction and Challenges

Hmmm!

Big Data: Introduction, Definitions and Representations

- What is Big Data
- Big Data Characteristics
- Where does it come from and how is it used

There's a Wealth of Data out there, but…….

Organizations have access to a wealth of information, but they can't get value out of it because it is sitting in its most raw form or in a semistructured or unstructured format; and as a result, they don't even know whether it's worth keeping (or even able to keep it for that matter).

Did you know that :

- 80 percent of the world's information is unstructured.
- Unstructured information is growing at 15 times the rate of structured information.
- Raw computational power is growing at such an enormous rate that today's off-the-shelf commodity box is starting to display the power that a supercomputer showed half a decade ago.
- Access to information has been democratized: it is (or should be) available for all.

- Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2012

Big Data: What is it?

- The situation where our most difficult problem is not how to store the data, but how to process it in meaningful ways
- Estimated 70-85% of all data is unstructured text, audio and images (and rising!)
  - In the past 50 years, the New York Times produced 3 billion words.
  - Twitter users produce 8 billion words – **every single day**.

  *Source: Kalev Leetaru, University of Illinois
- *More data is not simply more data, but more data is different!*

However, Big Data Is Evolutionary

- A terabyte size (10^{12}) data warehouse used to be Big Data
  - Now, some data warehouses store a petabyte of data
- New analytical platforms have emerged to store and analyze Big Data

- And, its creation and aggregation are accelerating,
- Petabytes → Exabytes → Zettabytes

  - 1,000,000,000,000,000,000,000

  - Total created in a year would fill over 75 billion 16 GByte iPads

Big Data Trivia

- 1 Petabyte = 1000 TeraByte
- 1 Exabyte = 1000 Petabyte
- 1 Zettabyte = 1000 Exabyte
- 1 Yottabyte = 1000 Zettabyte

- 1 Zettabyte = 1 099 511 627 776 Gigabytes
  = 1 Billion **1TB** Disk Drives

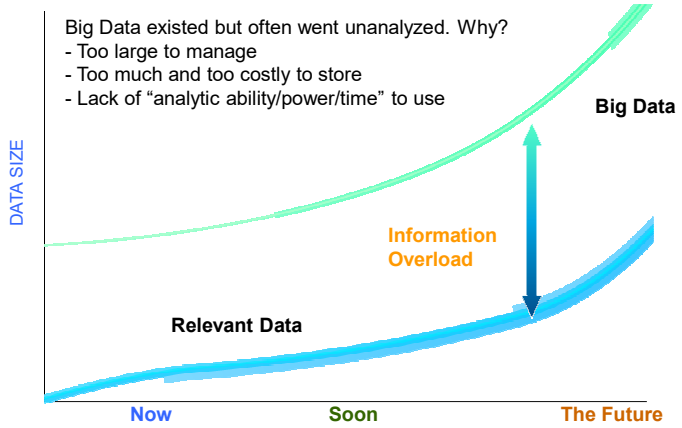
- From the *Cisco Visual Index Global Mobile Data Traffic Forecast Update, 2012*:
  - The projected world population by 2016 would be 7.3 Billion. What is the projected mobile connected devices by then?
    - Answer: 1.4 mobile devices per person
  - Video now accounts for over half of all Mobile device traffic bandwidth (52%). By 2016 over 70%.

Big Data: Where Does It Come From?

- Categories of Big Data include, but are not nearly limited to:
  - Social media activity, Web sites (e.g. Weblogs), Machine generated, RFID, Image, video, and audio, GPS
  - Cameras, Internet search histories, retail transactions,, genomic/biomed research, etc.
- Consider:
  - Tweets: How many copies are retweeted? Are these all saved?
  - Technical Papers: How many are downloaded and saved in personal archives?
  - Email: Backups, Local copies, saved copies, forwards, etc…

A Continuing Challenge

Big Data existed but often went unanalyzed. Why?
- Too large to manage
- Too much and too costly to store
- Lack of "analytic ability/power/time" to use



Big Data: Characteristics

- Doug Laney (Meta Group, now Gartner):
  - *Volume*: Total number of bytes associated with the data.
  - *Velocity*: The pace at which data are to be consumed. As volumes rise, the value of individual data points tend to more rapidly diminish over time.
  - *Variety*: The complexity of the data, which limits the utility of traditional means of analysis.
- Gartner suggests an additional feature:
  - *Variability*: The differing ways in which the data may be interpreted. Differing questions require differing interpretations.
- Kaisler, Armour, Espinosa and Money:
  - *Value*: The relative importance of different data to the decision-making process.

Big Data Definitions

Big Data can be defined as the amount of data *just beyond technology's capability to store, manage and process efficiently*.

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

- Gartner

Big Data: Value Proposition

- Potential value to U.S. health care per year: $300B
  - More than double the total annual spending on health in Spain
- $600B potential value obtained from using personal geolocation data globally
- 60% possible growth in retailer's operating margins form using big data

Ref: McKinsey Global Institute, Big Data: The Next Frontier for innovation, Competition, and Productivity, May 2011 and others

Mobile Big Data Possibilities

- Collect customer likes, dislikes, purchases
- Track customer location (GPS coordinates)
- Send customer push notification when they reach a specific region
  - Coupon for favorite restaurant
  - Special offer at stores they shop at
- Net results - Increased sales

Shipment Tracking: Big Data Possibilities

- Monitor truck routes via GPS
- Ensure trucks are on schedule
- Alert when truck is not on schedule
- Monitor refrigerator temperature
- Real-time visualization and alerting
- Net result – shipments stay on schedule and products delivered are not spoiled

Measuring the Value of Big Data

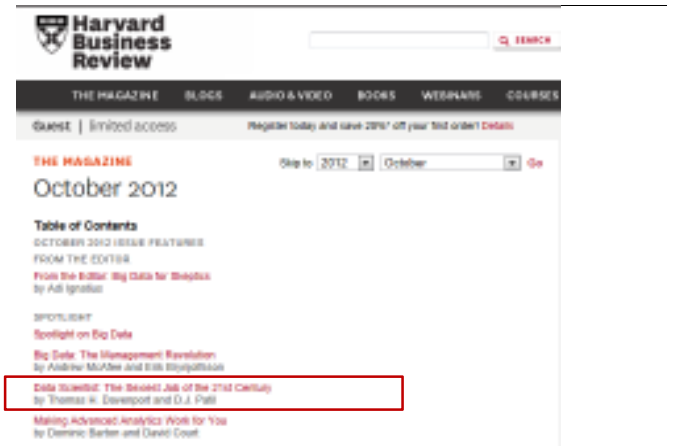
*Top 5 Advantages of Successfully Managing Big Data**

- Improving Overall Agency Efficiency
- Improving Speed/Accuracy of Decision
- Ability to Forecast
- Ease of Identifying Opportunities for Savings
- Greater Understanding of Citizens Needs

* Based on Meritalk Research survey of 151 federal IT professionals – The Big Data Gap, May 2012

Industry is serious about Big Data, for example

- IBM has made over $15B worth of acquisitions since 2005 betting on value migrating from reporting to predictive analytics.
  - Includes SPSS, Cognos, Unica, Softech, ILog, etc.
  - Developed IBM Business Analytics: http://www-01.ibm.com/software/analytics/spss/



Overview of Big Data Technologies and Architecture

- "*Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.*"

  – *Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO,* Philip Carter, IDC, September 2011.

Technologies for Big Data (and Analytics)

- Hadoop/MapReduce
- Columnar/Non-relational databases (HBase, MongoDB, Neo4J, SciDB,…)
- Data warehouses
- Appliances
- Analytical sandboxes (using Virtual Machines)
- In-memory analytics
- In-database analytics
- Graph Mining and Information Network Analysis
- Streaming and Critical Event Processing (CEP) Engines
- Cloud-based services (Software As A Service)

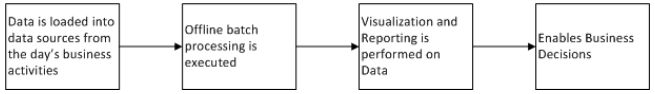
Classes of Big Data Processing

- **Batch Processing – offline**
  - **Hadoop**
  - **Data Mart**
  - **Data Warehouse**
  - **Database**

Wilkipedia, States of Data, 2011

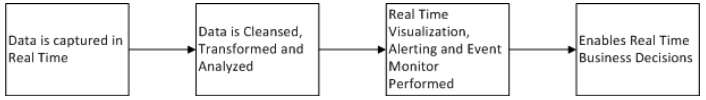
- **Stream Processing – real-time**
  - **In memory (i.e., BigMemory)**
  - **Complex Event Processing (CEP)**
  - **Dashboard visibility (i.e., MashZone)**
  - **Ingesting & analyzing Streams** (IBM ***InfoSphere Streams)***

Batch Processing



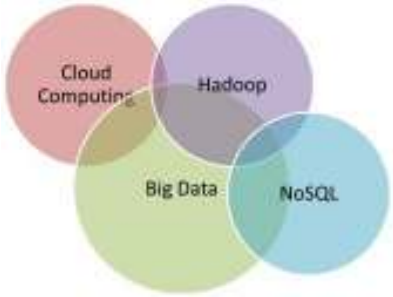
- Data is loaded into data sources from the day's business activities
- Offline batch processing is executed
- Visualization and Reporting is performed on Data
- Enables Business Decisions

Stream Processing



- Data is captured in Real Time
- Data is Cleansed, Transformed and Analyzed
- Real Time Visualization, Alerting and Event Monitor Performed
- Enables Real Time Business Decisions

Big Data and Cloud Computing



What is Hadoop?

- Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers.
- Hadoop is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

The Motivation For Hadoop

- Problems with Traditional Large-Scale Systems
- Traditionally, computation has been processor-bound
  - Relatively small amounts of data
  - Significant amount of complex processing performed on that data
- For decades, the primary push was to increase the computing power of a single machine
- – Faster processor, more RAM
- Need for a New Approach

Hadoop Components

- **Hadoop consists of two core components**
  - The Hadoop Distributed File System (HDFS)
  - MapReduce Software Framework (Yarn)
- **There are many other componants in the Hadoop Ecosystem**
  - Eg Pig, Hive, HBase, Flume, Oozie, Sqoop, etc
- **The Severs running HDFS and MapReduce is known as a Hadoop Cluster**
  - Individual machines are known as nodes
  - A cluster can have as few as one node, as many as several thousands
  - More nodes increased performance

Hadoop is more than just software

Hadoop provides: a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by MapReduce. There are other parts to Hadoop, but these capabilities are its kernel.

White, Tom (2010-10-01). Hadoop: The Definitive Guide: The Definitive Guide (p. 4). OReilly Media - A. Kindle Edition.

Components of Hadoop: MapReduce, HBase, Pig, Hive, HDFS, Zookeeper, Sqoop, YARN

Hadoop Components: HDFS

- **HDFS, the Hadoop Distributed File System, is responsible for storing data on the cluster**
- **Data files are split into blocks and distributed across multiple nodes in the cluster**
- **Each block is replicated multiple times**
  - Default is to replicate each block three times
  - Replicas are stored on different nodes
  - This ensures both reliability and availability

Core Hadoop Concepts

- **Applications are written in high-level code**
  - Developers don't have to worry about network programming, temporal dependencies etc
- **Nodes talk to each other as little as possible**
  - Developers should not write code which communicates between nodes
  - 'Shared nothing' architecture
- **Data is spread among machines in advance**
  - Computation happens where the data is stored, wherever possible
  - Data is replicated multiple times on the system for increased availability and reliability

What Is MapReduce?

- **MapReduce is a method for distributing a task across multiple nodes**
- **Each node processes data stored on that node**
  - Where possible
- **Consists of two key phases:**
  - Map
  - Reduce

History of MapReduce, HDFS

- "MapReduce" White paper, written by Jeffrey Dean and Sanjay Ghemawat, google researchers in 2004 - http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf
- "BigTable" White paper, written by google reserachers in 2006 - http://static.googleusercontent.com/media/research.google.com/en/us/archive/bigtable-osdi06.pdf
- Apache Foundation took on a clone of Google's MapReduce, now found at: http://hadoop.apache.org
- MapReduce 2.0 (YARN)
  http://hadoop.apache.org/releases.html#15+October%2C+2013%3A+Release+2.2.0+available

Spark

What is Apache Spark

- Apache Spark originally developed in 2009 at UC Berkeley and open sourced in 2010 as an Apache project.
- Spark provides a comprehensive, unified framework to manage big data processing requirements
- Can deal with a variety of data sets that are diverse in nature (text data, graph data etc) as well as
- Provides both batch v. real-time streaming data capabilities (unlike MapReduce, that is batch only)

Spark runs on top of Hadoop

- Spark runs on top of existing Hadoop Distributed File System (HDFS)
- Provides support for deploying Spark applications in an existing Hadoop cluster
- Spark is an alternative to MapReduce rather than a replacement to Hadoop.

Why Spark?

- Spark can enable applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk.
- MapReduce is a great solution for one-pass computations, but not very efficient for use cases that require multi-pass computations and algorithms.
- Spark lets you quickly write applications in Java, Scala, or Python. It comes with a built-in set of over 80 high-level operators. And you can use it interactively to query data within the shell.
- In addition to Map and Reduce operations, it supports SQL queries, streaming data, machine learning and graph data processing.
- Developers can use these capabilities stand-alone or combine them to run in a single data pipeline use case.

Spark EcoSystem

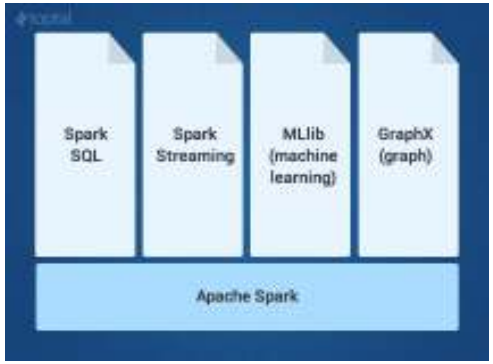
Spark Language Support

- Spark is written in the Scala Programming Language and runs on Java Virtual Machine (JVM) environment. It
- Currently supports the following languages:
  - Scala
  - Java
  - Python
  - R
  - Clojure

Spark and Hadoop EcoSystem

- Integrates well with the Hadoop ecosystem and data sources (HDFS, Amazon S3, Hive, HBase, Cassandra, etc.)
- Can run on clusters managed by Hadoop YARN or Apache Mesos, and can also run standalone
- The Spark core is complemented by a set of powerful, higher-level libraries which can be used in the same application.
- These libraries currently include SparkSQL, Spark Streaming, MLlib (for machine learning), and GraphX

Spark EcoSystem



Spark Core

- Spark Core is the base engine for large-scale parallel and distributed data processing.
- It is responsible for:
  - Memory management and fault recovery
  - Scheduling, distributing and monitoring jobs on a cluster
  - Interacting with storage systems

SparkSQL

- SparkSQL the Spark SQL component
- It originated as the Apache Hive port to run on top of Spark (in place of MapReduce
- Supports querying data either via SQL or via the Hive Query Language.

MLlib

- MLlib is a machine learning library that provides algorithms designed to scale out on a cluster
- Current includes classification, regression, clustering, collaborative filtering, etc
- Some algorithms also work with streaming data, (eg linear regression, k-means clustering.

GraphX

- GraphX is a library for manipulating graphs and performing graph-parallel operations.
- Provides tool for ETL, exploratory analysis and iterative graph computations.
- Has built-in operations for graph manipulation,
- Also provides a library of common graph algorithms such (eg PageRank).

Introduction to NoSQL Basic Concepts

from “Geek and Poke”

What is NOSQL?

- NoSQL is a class of database management system identified by its non-adherence to the widely used relational database management system (RDBMS) model with its structured query language (SQL).
- NOSQL has evolved to mean “Not Only” SQL
- NOSQL has become prominent with the advent of web scale data and systems created by Google, Facebook, Amazon, Twitter and others to manage data for which SQL was not the best fit.

NoSQL Definition

From www.nosql-database.org:

- Next Generation Databases mostly addressing some of the points: being **non-relational, distributed, open-source** and **horizontal scalable**.
- The original intention has been **modern web-scale databases**.
- The movement began early 2009 and is growing rapidly.
- Often more characteristics apply as: **schema-free, easy replication support, simple API, eventually consistent** / **BASE** (not ACID), a **huge data amount**, and more.

NoSQL vs Relational Databases:

RDBMS: Pros and Cons

Pros
- Many programmers are already familiar with it.
- Transactions and ACID make development easy.
- Lots of tools to use.
- Rigorous Design (Schema)

Cons
- Impedance mismatch.
  - Object Relational Mapping doesn't work quite well.
- Rigid schema design.
- Harder to scale.
- Replication.
- Joins across multiple nodes? Hard.
- How does RDMS handle data growth? (eg scaling) Hard.
- Need for a DBA.

- SQL and NoSQL have different strengths and weaknesses

NoSQL involves more Programming and Less Database Design

Alternative to traditional relational DBMS

- Flexible schema
- Quicker/cheaper to set up
- Massive scalability
- Relaxed consistency → higher performance & availability
- No declarative query language (SQL) → more programming
- Relaxed consistency → fewer guarantees

Transactions – ACID Properties (Relational Databases do this very well)

- ***A**tomic* – All of the work in a transaction completes (commit) or none of it completes
- ***C**onsistent* – A transaction transforms the database from one consistent state to another consistent state. Consistency is defined in terms of constraints.
- ***I**solated* – The results of any changes made during a transaction are not visible until the transaction has committed.
- ***D**urable* – The results of a committed transaction survive failures

What is BASE?

- BASE is an alternative to ACID
  - **B**asically **A**vailable
  - **S**oft state
  - **E**ventual consistency
- Weak consistency
- Availability first
- Approximate answers
- Faster

BASE Transactions Characteristics

- Weak but Eventual consistency – stale data OK
- Availability first
- Best effort
- Approximate answers OK
- Aggressive (optimistic)
- Simpler and faster

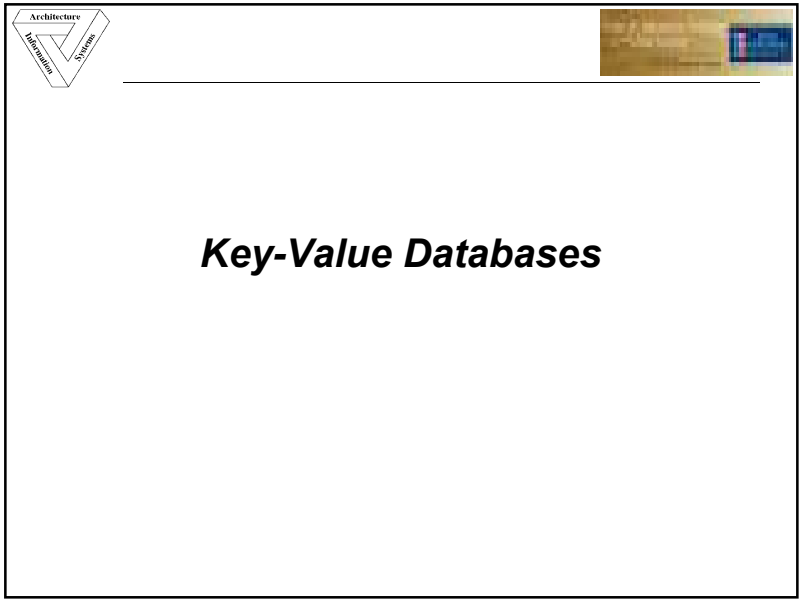
NoSQL: What Forms (Types) do these Databases Take

NoSQL Categories

- Relational data model is best visualized as a set of tables, rather like a page of a spreadsheet.
- NoSQL is a move away from the relational model.
- Four categories widely referred to as NoSQL:
  - Key-value
  - Columnar
  - Document
  - Graph.

Categories of NoSQL storages

- Key-Value
  SimpleDB, BigTable
- Column Family
  - Cassandra
- Document
  - MongoDB, CouchBase,
- Graph
  - Neo4j, Titan



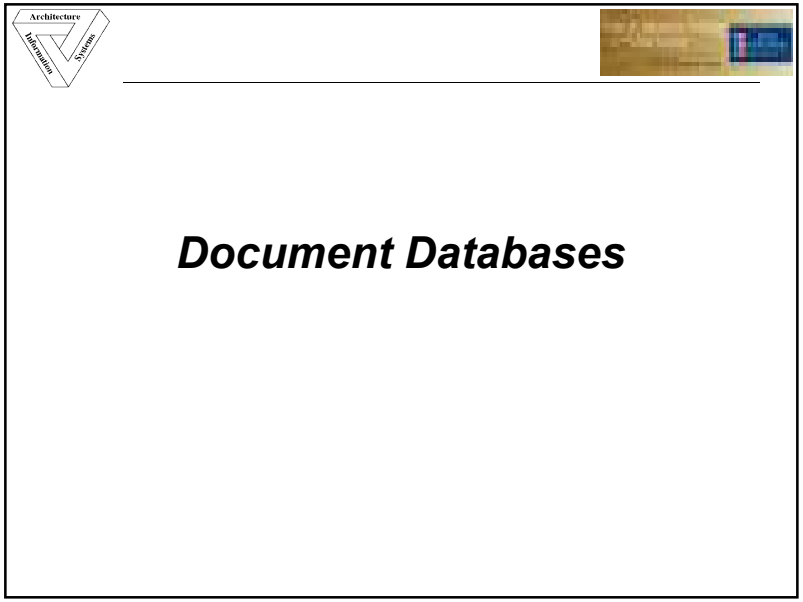
Key-Value NoSQL Databases

Extremely simple interface

- Data model: (key, value) pairs
- Value can be complex structure
- Example Operations: Insert(key, value), Fetch(key), Update(key), Delete(key)

Implementation: efficiency, scalability, fault-tolerance

- Records distributed to nodes based on key
- Replication
- Single-record transactions, “eventual consistency”
- Example: Amazon SimpleDB



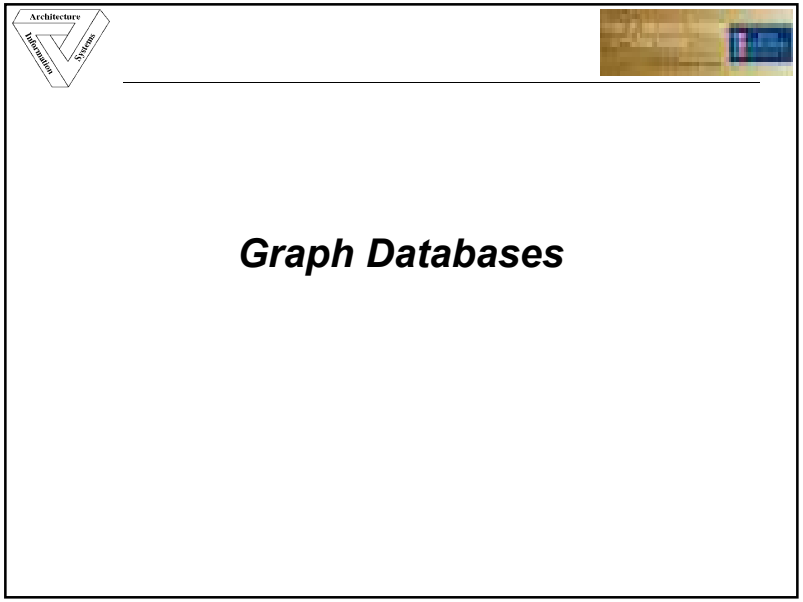
Document Databases

- Like Key-Value Stores except value is document
  - Data model: (key, document) pairs
  - Document: JSON, XML, other semistructured formats
  - Basic operations: Insert(key,document), Fetch(key), Update(key), Delete(key)
- Also Fetch based on document contents
- Example systems
  - MongoDB, SimpleDB

Columnar Databases

Columnar Databases

- In Relational databases data is stored in the disk row by row.
- Where in Columnar databases data is stored in the disk column by column
- Can be significantly faster than row stores for some applications
  - Fetch only required columns for a query
- But can be slower for other applications
  - OLTP with many row inserts



What is a Graph Database?

- A **graph database** is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.
- A graph database is any storage system that provides index-free adjacency.
- Every element contains a direct pointer to its adjacent elements and no index lookups are necessary.

- https://en.wikipedia.org/wiki/Graph_database

Analytics –

Knowledge and Process Framework

Why Analytics and Big Data ?

- There is little value in just storing and having data
- Business value is created when the data is analyzed, resulting in better decision-making and problem-solving
- Big Data improves decision-making because there is more data to analyze and better data management and analysis environments

What is Analytics?

"It is the scientific process of transforming data into insight for making better decisions" (INFORMS)

It is the use of:

- structured and/or unstructured data,
- information technology,
- statistical analysis,
- quantitative and/or qualitative methods, and
- mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions.

Business Analytics Examples

- Management of customer relationships (free WiFi)
- Financial and marketing activities (credit card drop)
- Supply chain management (find bottlenecks)
- Human resource planning (predict manpower)
- Pricing decisions (best price for a new product-Starbucks)
- Sport team game strategies (Moneyball)

Analytics using Big Data

- Large amounts of both structured and/or unstructured data, either at rest or in a streaming state,
- The use of advanced information technology to support the analysis and modeling of data
- Statistical methods to provide rigor to the analysis
- Visual analysis to help discover patterns and trends in the data and present the results to key decision makers
- Other quantitative and/or qualitative methods, and mathematical or computer-based models

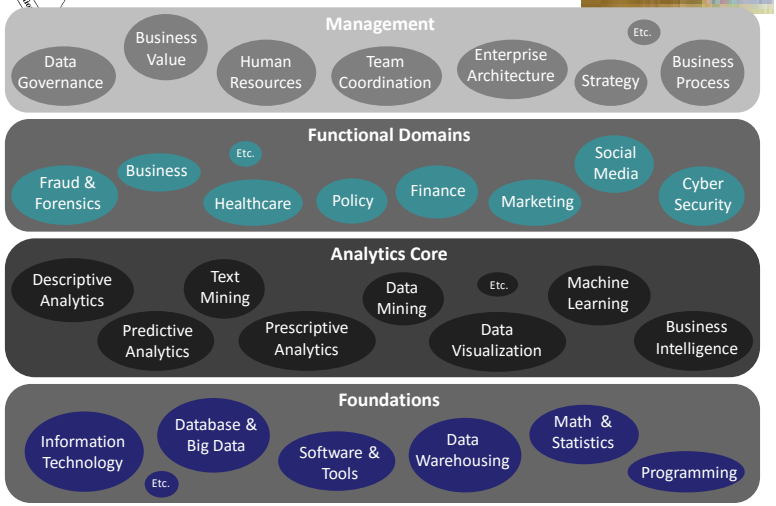
Another way to distinguish Big Data analytics environment in which the analysis is conducted

- Can do standard analytics work with Big Data by downloading the necessary data from large data warehouses and applying conventional analytics methods and tools.
- Big data analytics is also the practice of conducting analytics directly in the Big Data environments.
- This often requires programming skills, such as Java, Python, or R, and technologies like “in-memory” analytics, which are rapidly becoming mainstream.
- However, in some cases, when working with Big Data, analytics is not conducted directly on the Big Data per se.
- Instead, the necessary data is extracted from traditional and nontraditional sources into a structured data warehouse or database for further manipulation and analysis.

Old Stuff? New Stuff?

	Big Data	Analytics	Business Intelligence
Old Stuff	• Relational Databases • SQL • Data warehousing • Structured	• Data mining • Quantitative analysis • Statistics, regression • Decision models • Descriptive, predictive, prescriptive • Optimization • Machine learning	• Decision Support Systems • Executive Information Systems • OLAP
New Stuff	• V's • Multi-format • Unstructured data • Big data technologies (NoSQL, MapReduce, Hadoop, etc.)	• Unstructured data • Multi-source data • Multi-format data • Data in motion • Text analytics • Visual analytics • Social media analytics	• Smarter software • Better data visualization • More dynamic tools • Ease of use

Analytics Body of Knowledge



The Analytics Cycle

Step 1 – Problem: Formulate analytics question

Step 2 – Data: Identify, gather and prepare the data
Data identification, collection, cleansing, formatting, pre-processing, etc.

Step 3 – Model: Select the appropriate analytic methods
Descriptive – familiarize with and analyze the data to identify patterns: unsupervised machine learning, descriptive statistics, correlation, clustering, principal components, data mining
Predictive – use existing data to predict other data: develop and test prediction hypotheses; classification (e.g., classification trees, logistic regression) vs. value prediction (e.g., regression trees, linear regression); numeric (e.g., linear regression) vs. categorical (e.g., logistic regression, classification trees)
Prescriptive – decision models to inform on best courses of actions: optimization; linear programming;

Step 4 – Analysis: Apply the models to the data:
Explanation vs. prediction; variance vs. bias; parsimony vs. over-identification; accuracy testing.

Step 5 – Reporting: Prepare results report

Types of Analytics

- **Descriptive:** getting meaning from the data – e.g., BlueFin Technologies → did viewers liked a particular TV show last night, based on Tweets?
- **Predictive:** using some variables in the data to predict the outcome of others – e.g., Target → which product purchases are the best predictors that the customer is expecting a baby?
- **Prescriptive:** using the data to recommend what to do to achieve outcomes – e.g., Starbucks → what's the optimal price of a new coffee brand to maximize sales.

Analytics Approaches

- **Quantitative:** regression analysis; logistic regressions, classification trees; correlation; data reduction; etc.
- **Qualitative/Text:** text mining; natural language process analysis; content pattern analysis; adding structure to unstructured data; etc.
- **Visual:** infographics; heat maps; trend charts; Tableau graphics; social network diagramming; etc. (see IBM's Many Eyes www-969.ibm.com).

Analytics Methods

- **Association:** correlation among variables, analysis of variance, regression models, which variables co-vary with which? – e.g., how much does annual income increase with each year of additional university education?
- **Classification (and probability estimation):** in which class does a case belong (predicting the probability that a new case will fall in a given class); chi-square analysis, logistic regression models – e.g., patient tested positive (or negative) for a disease → what are the probabilities of testing positive for a disease?
- **Others:** clustering, similarity matching, co-occurrence grouping, profiling, link (strength) prediction, data reduction (factor analysis), causal modeling, etc.

Machine Learning

•**Machine learning** refers to analytics models that learn from the data as the data changes

- Training data: data used to build the models by associating predictors (or rules) with outcomes; e.g., spam filtering
- Test data: data used to test the models by evaluating if the models predicted new data outcomes correctly
- Unsupervised learning: no target group specified; e.g., clustering; co-occurrence grouping; profiling; density estimates; pattern discovery; customer categorization.
- Supervised learning: specific target specified – e.g., which loan clients are more likely to default on their loan? Regression; predictions.

Example Software Tools for Analytics

- SAS: powerful suite of statistical programs, some are visual and easier to use (e.g., SAS Enterprise Guide, SAS Enterprise Miner)
- SPSS: powerful suite of statistical programs, some are visual and easier to use (e.g., SPSS Modeler)
- R: powerful open source object-oriented programming language for statistical work, with literally thousands of publicly available libraries for most statistical work; R Studio is a popular free software tool to manage R projects
- Tableau: intuitive and popular visual analytics program – it can run R code and draw the resulting graphics

Big Data Analytics in Use Cases

Big Data: The Reward

- Prof. Brynjolfsson (MIT) studied 179 large companies and found that those adopting "data-driven decision making" achieved productivity gains that were 5 -6% higher
- "Competing on Analytics" by Tom Davenport – "Some companies have built their very businesses on their ability to collect, analyze and **act** on data."

Big Data Interoperability Framework

- The government of the US and other major countries are actively supporting the big data industry with standardization organizations focusing on Big data issues
- To help US Government agencies comply with big data requirements, the National Institute of Standards and Technology (NIST) has released a draft of its Big Data Interoperability Framework,
- Standardize on a vendor-neutral, technology- and infrastructure-agnostic reference architecture.
- Includes defining data analytics-related phrases, outlining management templates and describing common use cases for large data sets and other large amounts of information traditional data architectures can't efficiently handle.

- http://bigdatawg.nist.gov/V1_output_docs.php

Fraud Detection: Big Data possibilities

- Every credit card transaction runs through fraud detection real-time
  - Keep black-listed cards in memory
- Real-time pattern matching for event prediction
- Net result - saves millions of dollars in fraud

FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide

Tool for the early detection of fraudulent activity on credit and debit cards. This system predicts the probability of fraud on an account by comparing current transactions (normal cardholder activity) to current fraud trends.

- Real-time, Transaction-Based
- scoring with neural network
- models
- Real-Time Decisioning
- Global profiles
- Adaptive models.

Big Data in Government Fraud Detection

- The US Social Security Administration (SSA) is using big data to analyze massive amounts of unstructured data in disability claims.
  - The SSA is now able to process medical results and expected diagnoses more rapidly and efficiently to recreate the decision making process, better identifying suspected fraudulent claims.
- The Securities Exchange Commission (SEC) is applying big data strategies to monitor financial market activity.
- They are using natural language processors and network analytics to help identify illegal trading activity.

Big Data and Government Oversight

- The **Notice and Comment project** provides instant access to over 4 million government documents, from federal regulations published by the Federal Register to local public notices.
- The project is using advanced analytics and natural language processing to ingest government documents and track changes in policies, laws or regulations.
- Users can comment or vote on pending federal regulations or local public notices with ease.
- The site's data updates daily for real-time monitoring of proposed actions and emerging trends.
- Users can gain support for their views using integrated social media and the web's best writing tips to effectively advocate before proposals become law.
- http://www.noticeandcomment.com/

Big Data and Health-Related Research

- The US Food and Drug Administration (FDA) is deploying big data technologies across many labs involved in testing to study patterns of foodborne illness.
- The database, is part of the agency's Technology Transfer program,
- Allows the FDA to respond more quickly to contaminated products that enter the food supply and contribute to the 325,000 hospitalizations and 3000 deaths related to foodborne illness every year.

Starbucks

- Concern over the taste of a new coffee product
- Social media was followed
- Taste was fine but price was too high
- Price was reduced by mid afternoon

Big Data and Crime Fighting

- In response to the Boston Marathon bombing, big data technologies enabled the rapid analysis of more than **480,000 images.**
- Images represent unstructured data.
- Good descriptions of the suspects allowed analysts to write code and algorithms to quickly analyze the images, looking for anomalies and certain patterns.
- Automation of the screening of sensor information for criminal behavior enables real-time analysis, reduces the time to decision.

Big Data in Environmental Protection

- National Aeronautics and Space Administration (NASA) and the U. S. Forest Service have a big data strategy to improve interoperability and integrated research efforts which enable them to better predict weather, ground conditions, and forest fire risks.
- This effort took a lot of coordination in data requirements and data governance.

Using Big Data to Revamp Boston, Massachusetts Bus Transit

- Pop up bus transportation system that adapts in real-time to ridership needs
- use a network of express shuttles that offer efficient and flexible trips that are dynamic
- Analyzes between two and three billion data points to understand how Boston moves.
- Has about 19 different data streams," including municipal data, census data and social media data

95

Using Big Data to revamp Boston, MA bus transit II

- Has cut some commute times in half by strategically offering bus service in the city.
  - For example, a ride from Coolidge Corner to Kendall Square, would likely 42 to 55 minutes on the Massachusetts Bay Transportation Authority, has taken 15-18 minutes on his company's buses.
- Starts by targeting neighborhoods it considers commuter pain points.
- Over the next 18 months set-up pilot programs around the United States.

96

Retail Analytics

- Walmart handles more than 1 million customer transactions every hour (2.5 petabytes of data)
- Analytics are not new for retailers:
  - Doing analysis of point of sale transactions and other business data for years
  - Bar code data first appeared in 1970s

    (Pack of Wrigley's Chewing Gum scanned via UPC in Marsh Supermarket, in Troy, OH in 1974)
- Types:
  - Customer analytics
  - Merchandising analytics
  - Store operations
  - Marketing analytics
  - Return, fraud and Loss Prevention analytics

Target Stores

- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
- Target assigns every customer a guest ID tied to their credit card
- Stores a history of everything they bought and any demographic info
- Andrew Pole looked at historical data for all women signed up for Target's baby registry
- From Pole's analysis, some key patterns emerged
- For example, they noticed that women started buying larger quantities of unscented lotion around the second trimester
- And, in the first 20 weeks, pregnant women loaded up on mineral supplements and large cotton balls, hand sanitizers and wash clothes
- They identified about 25 products that predicted pregnancy and could do so within a narrow window
- So, Target started sending out coupons for baby items
- Which started arriving at this teen's house and caused her father to wonder what was going on

Some Challenges with Big Data and Analytics

Big Data: The Opportunity

- Information Gap – The difference between the information managers view as important and the information that their organizations can provide for them. The gap on information about customer needs and preferences is estimated to average 80%.
-
  - (Source: Pricewaterhouse Coopers, "12th Annual Global CEO Survey, Redefining Success," PricewaterhouseCoopers, 2009)

Analytics: Why?

"In 2014, 30% of analytic applications will use proactive, predictive and forecasting capabilities"

"The market for BI and analytics is undergoing gradual evolution."
Gartner. Feb 1st, 2011

"In 2011, the use of analytics as a competitive differentiator in selected industries will explode"

"The roles of marketing, sales, human resources, IT management, and finance will continue to be transformed by the use of analytics"
International Institute for Analytics. Dec 3rd, 2010

"In 2014, global market for Analytics software will grow to $34 Billion"

IDC. Nov 9th, 2010

Lack of Analytical Talent

By 2018, the US faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions

- McKinsey Global Institute, 2011

U.S. Government Concerns

- 4 out of 10 agencies are lacking staff and infrastructure resources.
- 1 in 3 agencies are having trouble finding expertise
- Need for more robust analytics tools
- Sheer volume of data
- Overall cost

•Federal Government Survey by Unisys Corp. (http://www.unisys.com/big-data).

Top three types of professionals Government agencies are looking to hire

- Business Analyst - A business analyst is someone who analyzes an organization or business and documents its business or processes or systems, assessing the business model or its integration with technology.
  - The business analyst should have good analytic skills
- Data Analyst - A data analyst is someone who identifies, cleans, transforms, validates and models the data with the purpose of understanding or making conclusions from the data for decision making purposes.
- Director/Manager of Analytics

*Federal Government Survey by Unisys Corp. (http://www.unisys.com/big-data).

Growth and Resource Challenges Over the Next Decade

- Servers (Physical/VM): 10x
- Data/Information: 50x
- #Files 75x
- IT Professionals <1.5x

- Source: Gantz, John and Reinsel, David, "Extracting Value from Chaos", IDC IVIEW, June 2011, page 4

Ability to capture digital data has exceeded ability to analyze

- Key Point #1: Our ability to digitize materials at the large scale has outstripped our methods for analyzing them.
- Key Point #2: We haven't figured out how to take advantage of the large scale. (i.e. What NEW things can we see in data that we can't see in small data?)
- Key Point #3: Simply having digital data isn't enough.
- Libraries, archives, governments & other data holders need to work with researchers to determine how to make data available
- Key Point #4: The problems we wish to address are inherently interdisciplinary and international in scope.

Key Challenges Today for Tomorrow

- Big Data Scientists or lack thereof
- Potential to end up like “data mining”
- “80% of the effort is in extracting, moving cleaning, and preparing the data, not actually analyzing it.”
- Don’t disregard Traditional Analytics - Traditional & Big Analytics will beside by side for years

High Expectations for Big Data

- 85% of organizations reported that they have Big Data initiatives planned or in progress.
- 70% report that these initiatives are enterprise-driven.
- 85% of the initiatives are sponsored by a C-level executive or the head of a line of business.
- 75% expect an impact across multiple lines of business.
- 80% believe that initiatives will cross multiple lines of business or functions.
  - *Who's Really Using Big Data*, Paul Barth and Randy Bean, Harvard Business Review Blog Network, September 12, 2012,

Capabilities gap

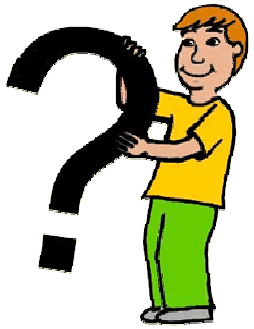
- Only 15% of respondents ranked their access to data today as adequate or world-class.
- Only 21% of respondents ranked their analytic capabilities as adequate or world-class.
- Only 17% of respondents ranked their ability to use data and analytics to transform their business as more than more than adequate or world-class.

– *Who's Really Using Big Data*, Paul Barth and Randy Bean, Harvard Business Review Blog Network, September 12, 2012.

Big Data: Example Benefits

- Creating visibility into operations and processes
- Enabling experimentation to discover needs
  - Simulation/Computational Science if the third leg of research
- Explore variability and diversity
  - Understand the causes across multiple disciplines
- Improve Performance
  - Greater granularity provides deeper insight
- Population Segmentation for Targeted Actions
- Augment/Replace human decision-making with automated algorithms
  - Minimize risk; deeper insight; explore more alternatives
- Create new business models, innovative processes and products, and services

Questions



Thank You!!

Who We Are

Frank Armour, Ph.D

Kogod School of Business/American University

farmour@american.edu

Dr Armour is an assistant professor of information technology at the Kogod School of Business, American University. He is the faculty program director for the MS in Analytics degree program. He received his PhD from the Volgenau School of Engineering at George Mason University. He is also an independent senior IT consultant and has over 25 years of extensive experience in both the practical and academic aspects applying advanced information technology. He has led initiatives on, and performed research in: Business analytics, Big Data, Enterprise architectures, business and requirements analysis, Agile System Development Cycle Development (SDLC), and object-oriented development. He is the coauthor of the book, Advanced Use Case Modeling, Addison Wesley and he is the author or coauthor of over 30 papers in the Information Technology discipline. He is primary co-chair for the enterprise architecture minitracks at both the HICSS and AMCIS conferences.

J. Alberto Espinosa, Ph.D

Kogod School of Business/American University

alberto@american.edu

Dr. Espinosa is currently Professor of Information Technology at the Kogod School of Business, American University. He holds a Ph.D. and Master of Science degrees in Information Systems from Carnegie Mellon University, Graduate School of Industrial Administration; a Masters degree in Business Administration from Texas Tech University; and a Mechanical Engineering degree from Universidad Catolica, Peru. His research focuses on coordination and performance in global technical projects across global boundaries, particularly distance and time separation (e.g. time zones). His work has been published in leading scholarly journals, including: Management Science; Organization Science; Information Systems Research; the Journal of Management Information Systems; Communications of the ACM; Information, Technology and People; and Software Process: Improvement and Practice. He is also a frequent presenter in leading academic conferences.

Who We Are

Stephen H. Kaisler, D.Sc.

Principal/SHK & Associates

Laurel, MD

skaisler1@comcast.net

Dr. Stephen Kaisler is currently a Research Scientist at a small company and a Principle in SHK & ASsociates.. He has previously worked for DARPA, the U.S. Senate and a number of small businesses. Dr. Kaisler has previously worked with big data, MapReduce technology, and advanced analytics in support of the ODNI CATALYST program. He has been an Adjunct Professor of Engineering since 2002 in the Department of Computer Science at George Washington University. Recently, he has also taught enterprise architecture and information security in the GWU Business School. He earned a D.Sc. (Computer Science) from George Washington University, an M.S. (Computer Science) and B.S. (Physics) from the University of Maryland at College Park. He has written four books and published over 35 technical papers.

William H. Money, Ph.D.

School of Business Administration

The Citadel

wmoney@citadel.edu

William Money joined the The Citadel as Associate Professor of Business Administration in 2014. Previously, he was with George Washington University School of Business faculty September 1992 after acquiring over 12 years of management experience in the design, development, installation, and support of management information systems (1980-92). His publications and recent research interests focus on information system development tools and agile software engineering methodologies, collaborative solutions to complex business problems, program management, business process engineering, and individual learning. He developed teaching and facilitation techniques that prepare students to use collaboration tools in complex organizations and dynamic work environments experiencing significant change. Dr. Money's has a Ph.D., Organizational Behavior 1977, Northwestern University, Graduate School of Management; the M.B.A., Management, 1969, Indiana University; and a B.A., Political Science, 1968, University of Richmond.

Data Science

"Is a set of fundamental principles that guide the extraction of knowledge from data." (Provost & Fawcett, Data Science for Business)

A data scientist is has deep knowledge on analytics, all aspects of data, the discipline in which analysis is conducted (marketing, healthcare, social media, etc.), and the underlying core disciplines (e.g., statistics, mathematics, database, etc.)

Data Mining

"Is the computational process of discovering trends in data" (ACM)

- It incorporates machine learning, statistics and database systems.
- It is about extracting patterns and knowledge and identifying previously unknown relationships in the data.
- e.g., Cluster analysis; anomaly detections; associations
- Good for hypotheses development

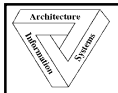
Business Intelligence

Business Intelligence refers to the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help its users make better decisions
(Wixom and Watson, Teradata University Network 2012)

Gartner's 2012 predictions for business intelligence focus on the challenges around cloud, mobility, alignment with business metrics and a balanced organizational model between centralized and scattered
(CIO Magazine, April 2012)

Big Data in Financial Market Analysis

- The US Federal Housing Authority (FHA) has over 23 years of experience in leveraging analytics to manage a positive cash-flow fund.
- The FHA is the only sub-prime mortgage insurance fund that did not need a bail out during the housing bubble burst.
- They apply big data analytics to help forecast default rates, repayment rates, claim rates.



Resources

- American University Analytics links (http://auapps.american.edu/~alberto/analytics)
- R (http://www.r-project.org/); R Studio (http://www.rstudio.com/); RSeek (http://www.rseek.org/)
- SAS for academics (http://support.sas.com/learn/ap/index.html); SAS Enterprise Guide (http://support.sas.com/software/products/guide/index.html); SAS Enterprise Miner (http://www.sas.com/en_us/software/analytics/enterprise-miner.html)
- SPSS for academics (http://www-01.ibm.com/software/analytics/spss/academic/); SPSS Modeler (http://www-01.ibm.com/software/analytics/spss/products/modeler/)
- Visual analytics:
  - Tableu for academics (http://www.tableausoftware.com/academic)
  - Inforgraphics (https://infogr.am/)
  - IBM's Many Eyes (http://www-969.ibm.com/software/analytics/manyeyes/)
  - Other free tools (http://www.computerworld.com/article/2506820/)
- Teradata University Network (http://www.teradatauniversitynetwork.com/)
- Books:
  - Moneyball (Lewis) (http://www.amazon.com/dp/B0076XKTAA/)
  - Introductory R (http://www.amazon.com/dp/B00BU34QTM/)
  - R for Everyone (www.amazon.com/dp/B00HFULELW/)
  - Machine Learning with R (http://www.amazon.com/dp/B00G9581JM/)
  - Social Media Mining with R (www.amazon.com/dp/B00J9B1IOI/)
  - Competing on Analytics (Davenport) (www.amazon.com/dp/B004OC072Q/)
  - Taming the Big Data Tidal Wave (Franks) (www.amazon.com/dp/B007NUQH4S/)
  - Big Data: A Revolution …. (www.amazon.com/dp/B009N08NKW/)