## Schedule of Course Activities: Session 27

## *[Cloud 519: Introduction to Cloud Computing Online-Based]*

## *[Instructor: John C. Chan]*

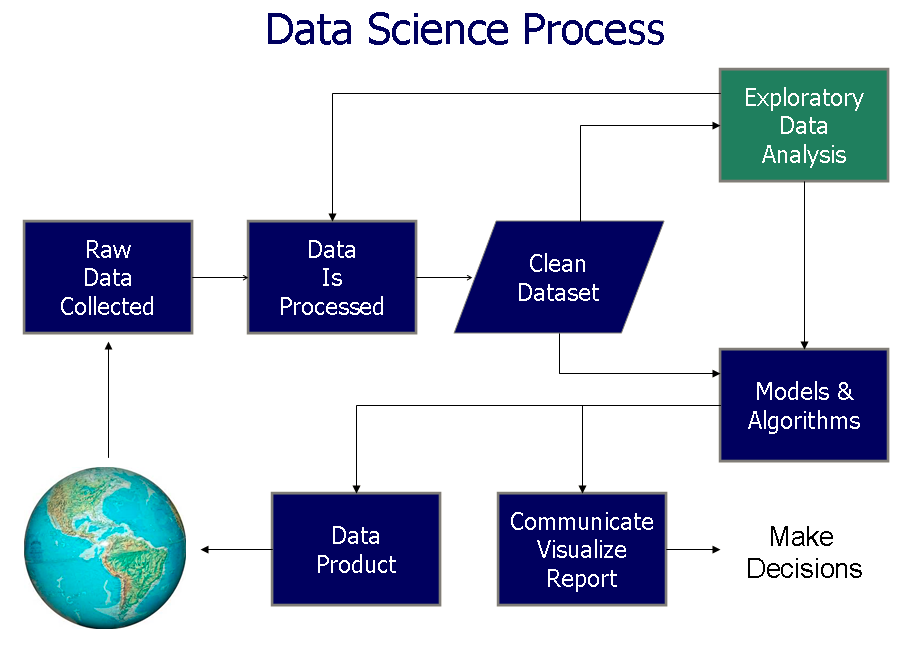
|  |  |
| --- | --- |
| **Overview of Session** |  |
| We will answer the following questions: | 1. Data Analytics 2. Major features and importance. 3. … |

# **Data analysis**

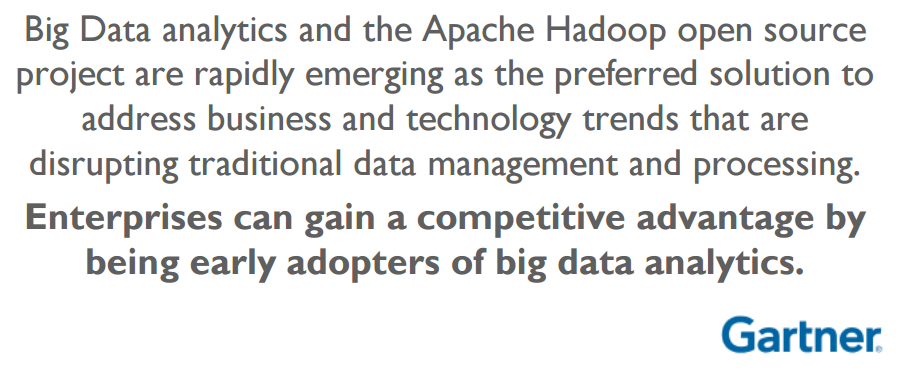
**Analysis of data** is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

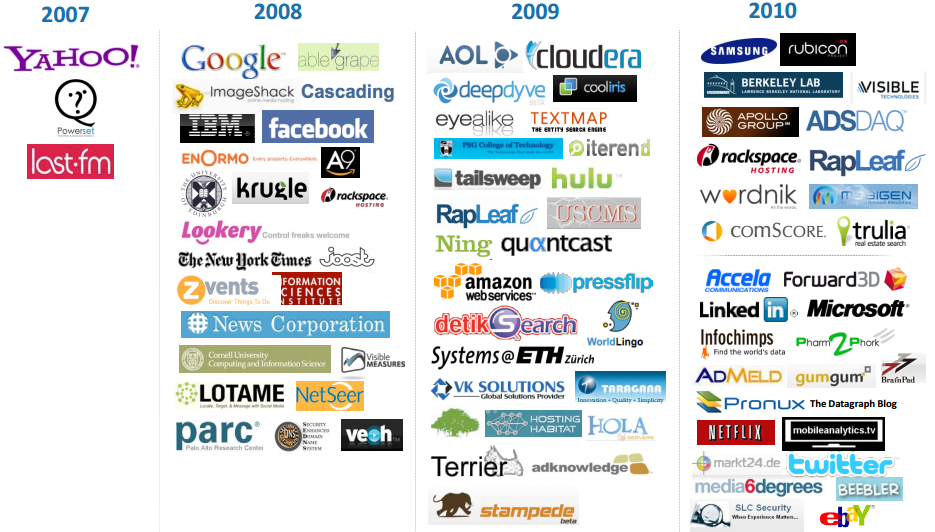
A high level data processing flow chart:



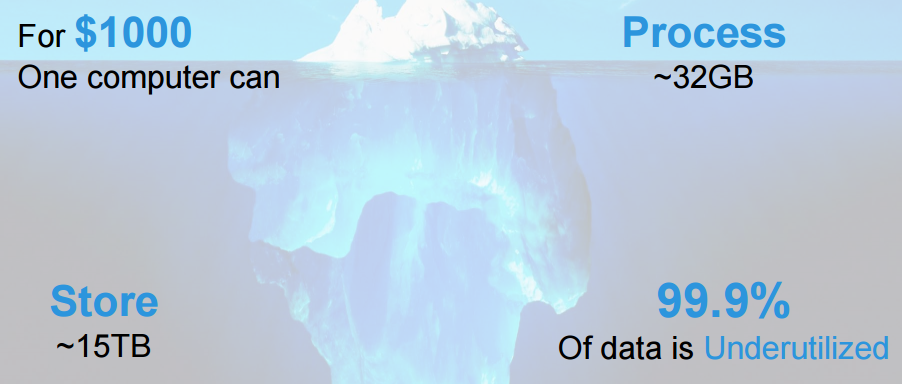
**Why Use Hadoop in Big Data Analysis?**



**Hadoop Adoption for Big Data Analytics (Source: SNIA),**

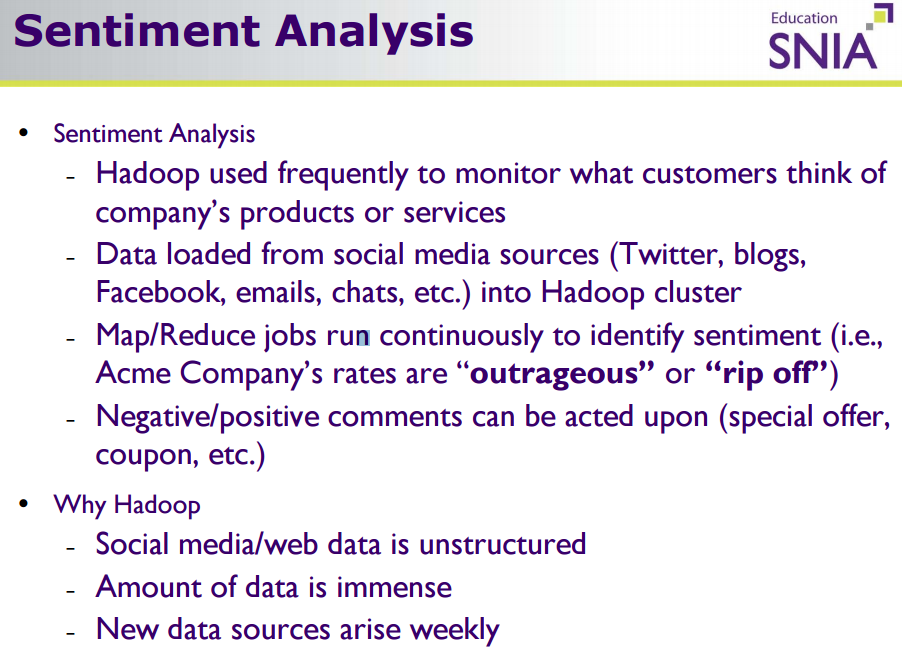


**Our Collected Data is Underutilized (An Example):**

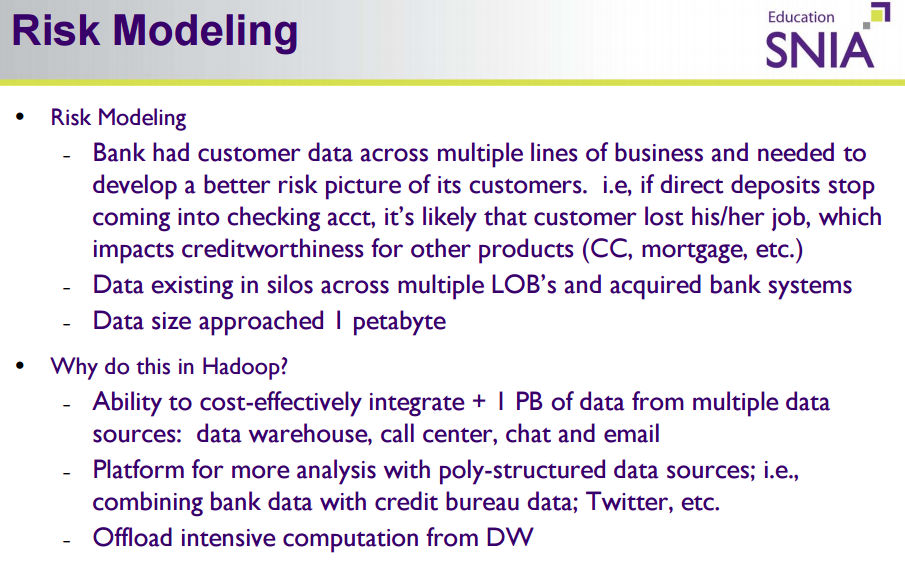


(Furthermore, we have more CPU processing power, than the concurrent availability of data, for it to process. That is, if you can bring the data to the CPU as needed input, on-time.)

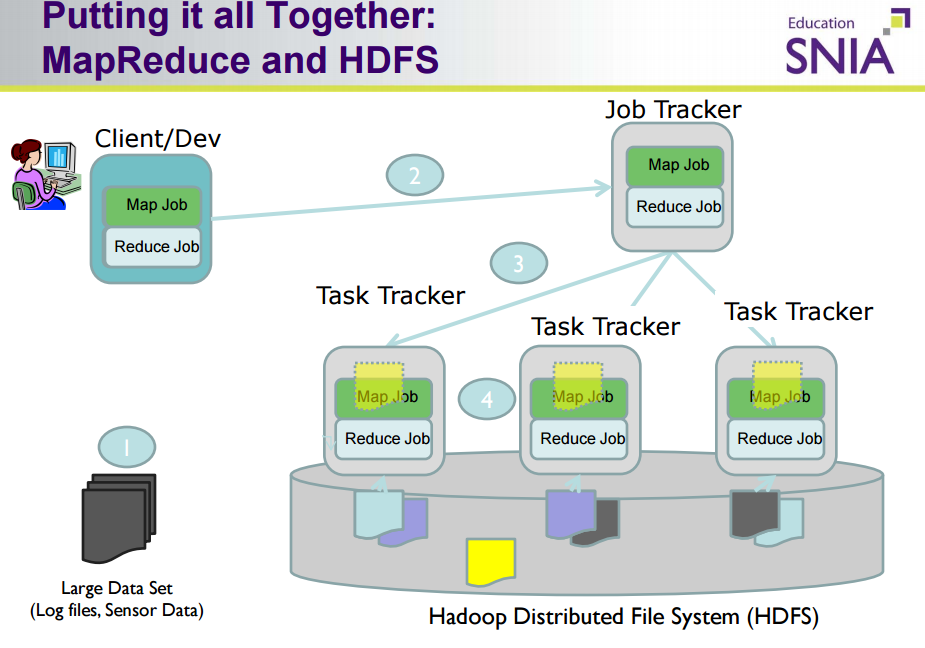
**Hadoop and Big Data Analytics Usage, Example #1:**



**Hadoop and Big Data Analytics Usage, Example #2:**



**How Hadoop does its job behind the scene:**



Let’s get a 1st hand usage example of Hadoop in Big Data Analysis by watching this video (It has practical commercial significance, Enjoy!):

<https://www.youtube.com/watch?v=weJI6Lp9Vw0>

Key Take-Away:

* 3 Steps: Load Data, Refine, and Visualize.
* Multi-year data, multi-data sources. It needs to be coalesced.
* Ability to drag/drop the coalesced file onto your PC for data analysis.
* Import all the columns from the Hadoop/Processed file into Excel.
* Visualize: Many companies find a niche in this business. E.g. Tableau.

Let’s wrap up tis class module, gaining the overall picture of data analysis, by watching this video:

<https://www.youtube.com/watch?v=c4BwefH5Ve8>

Key Take-Away:

* We have huge amount of data now. a) Worker/Employee generated. b) User generated; c) Machine generated data.
* Bring CPU to the data that it needs to be processed, instead of the other way around. (We learned earlier in the class, that bringing data to the CPU is an expensive, impractical operation…too much data to transfer on limited network bandwidth.)
* Hadoop MapReduce=Table\_of\_Content in the distributed file system.
* Think bigger about the internet, think bigger about life…

End-of-Class Module.

Questions? Please email to me, or post it on Blackboard.

Thank you.