Big Data Analytics – An Introduction

Chandan Mazumdar
Professor, CSE, JU
chandan.mazumdar@gmail.com

Course Outcomes

The students of this course should be able to

- CO1. Understand the concept and challenge of big data and why traditional technology is inadequate to analyze big data [K1, K2]
- CO2. Collect, manage, store, query, and analyze various forms of big data [K3, K4, K5]
- CO3. Gain experience on analytics methods and tools to solve big data problems [K3, K4, K5]
- CO4. Translate data into clear, actionable insights [K4, K5, K6]

Syllabus

- 1. What is Big data – Properties of Big Data – Volume, Velocity, Variety and Veracity. [CO1]
- 2. Data Mining and Analytics – Types of Analytics, Statistical, Machine learning and Computational Models [CO1]
- 3. Distributed File Systems – Hadoop and other variants – Data Ingestion and Munging [CO1, CO2]

- 4. Map-Reduce mechanism – Its architecture and algorithmic issues [CO2, CO3]
- 5. Similarity Search in very large data sets [CO2, CO3]
- 6. Frequent Item-set Mining in very large data sets [CO2, CO3]

- 7. Clustering very large high dimensional data sets [CO2, CO4]
- 8. Outlier Detection in very large data sets [CO2, CO3]
- 9. Data Stream Analytics [CO2, CO3]

- 10. Applications:

  Advertisement on the Web [CO4]

  Recommendation Systems for Online Stores [CO4]

  Mining very large graphs (social graphs) [CO4]

- 11. Infrastructural Issues:

  Hardware and Software Architectures [CO1, CO2]

  Reliability and Availability Issues [CO1, CO2]

Total 40 hours

Data Explosion: Example

- In a single day 294 billion emails are sent
- 2 million blog posts are written everyday
- 172 million people visit Facebook everyday and more than 250 million photos are uploaded to Facebook everyday
- Twitter serves more than 500 million tweets per day
- Google conducts more than 4 billion searches per day, number of web pages indexed 130 trillion
- Walmart handles more than 1 million customer transactions every hour, which is estimated to contain more than 2.5 petabytes of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.

Motivation: What is the Big Deal?

- Cannot store data @ generation and collection
- Cannot transfer the huge data to where it can be processed
- Data sets are becoming increasingly heterogeneous (type, grain, structure, meaning, …)
- Data sets are unorganized and hence not easily usable
- Very high volume data have high value for a very short time

However,

- The utility of the data is limited only by our ability to interpret it in time

What is Big Data? Definition

Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, curate, manage, and process the data within a tolerable elapsed time

- Wikipedia

What is Analytics?

- The discovery and communication of meaningful patterns or interesting insights from data using
  - Mathematical properties of data
  - Computing for accessing and manipulating data
  - Domain knowledge to increase interpretability of data and results of analysis
  - Statistical techniques for drawing inferences or making predictions on/from data

Why Analytics?

- 3 broad purpopses
- Using observed or measured data from a real-life situation
  - Uncover the characteristics of a data set based on its mathematical properties
  - Answer specific questions from one or more datasets with a given level of certainty
  - Develop a mathematical model for predicting the characteristics or behaviour of yet-unobserved data from the same situation

Analysis vs Analytics

- **Data analysis** refers to the process of compiling and analysing data to support decision making

- **Data Analytics** include Ingestion, Munging, Processing & Analysis, Visualization and Interpretation. Covers the tools and techniques for the same

Types of Analytics

- Descriptive
  - Evolving new descriptions of entities and their properties
- Diagnostic
  - Determine whether something has happened
- Predictive
  - Forecasting events that have not happened
- Prescriptive
  - Finding the best course of action in a given situation
- Cognitive
  - Combining analytics with AI and Machine Learning

Data Mining: The real challenge

Change in approach

- Instead of using data to train a Machine Learning Engine that can extract knowledge from the data,
- Apply the algorithms to the data

Technology changes

- Change the data structure of the data store
- Change the processing structure of the data store
- Change both

How it all started: Google PageRank①

- Intent: Based on search terms, the web pages to be ranked and serviced

- This has to be done on billion+ pages in the web!

① PageRank was invented by Larry Page also founder of Google

Key Problem Domains: Areas of focus

- Similarity Search in very large data sets
- Frequent Item-set Mining in very large data sets
- Clustering very large high dimensional data sets
- Outlier Detection
- Advertisement on the Web
- Recommendation Systems for Online Stores
- Mining very large graphs (social graphs)

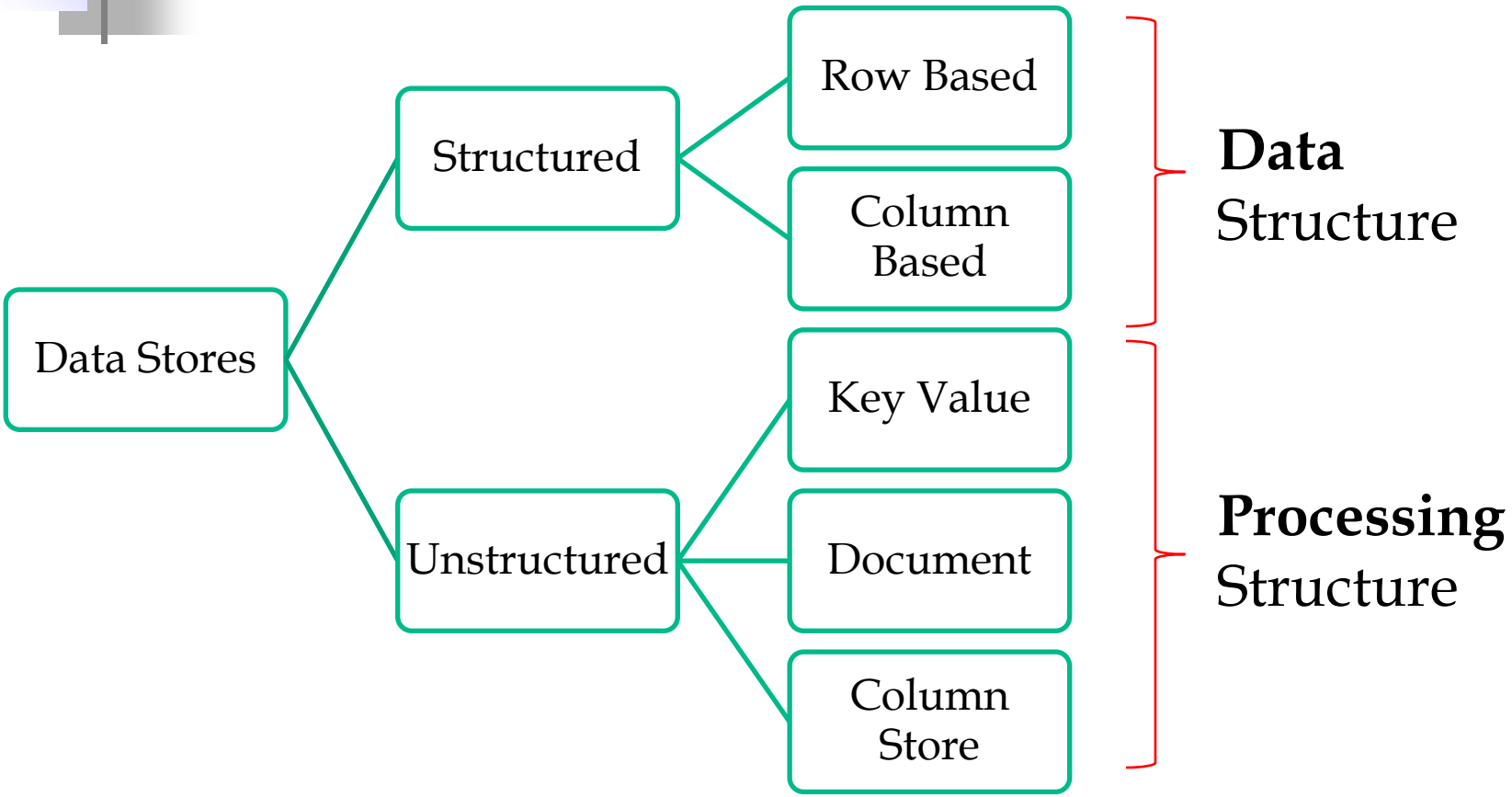
Sanity check of result: Significance

- Bonferroni's Principle
  - As the input set is very large, it is important to make sure that the output is more significant than the general probability applied on random data item
- Matthew Effect
  - "Rich get richer" concept, where page that has links from many page keeps on increasing in "importance"

Map Reduce, Columnar DB

TECHNOLOGIES FOR BIG DATA

Data Store Options



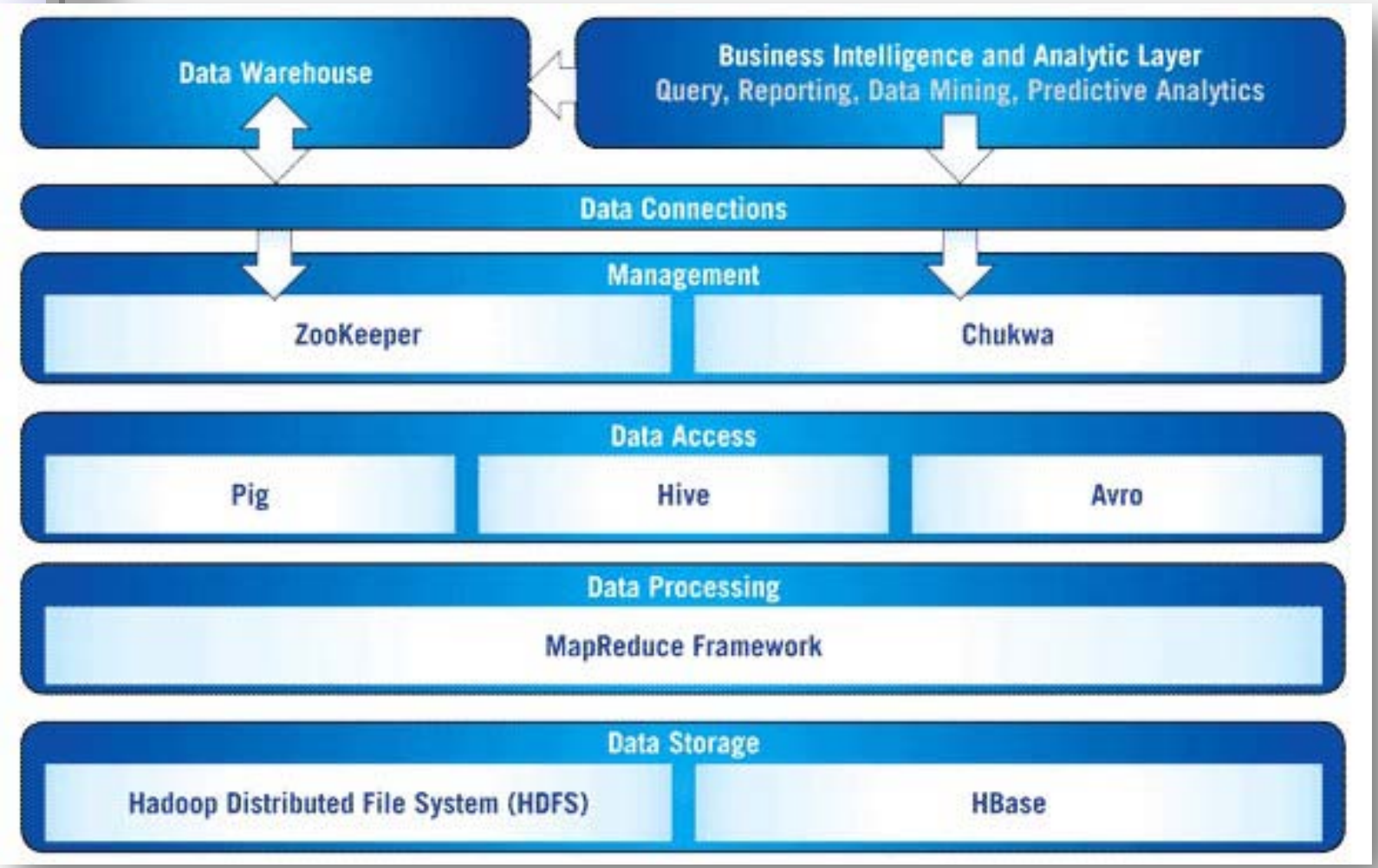
Landscape

The Emerging Database Landscape

This chart gives a quick overview of the strengths, weaknesses and use cases for row-based, columnar and NoSQL databases.

	Row-Based	Columnar	NoSQL—Key Value Store	NoSQL—Document Store	NoSQL—Column Store
Basic Description	Data structured in rows	Data is vertically striped and stored in columns	Data stored usually in memory with some persistent backup	Persistent storage for unstructured or semi-structured data along with some SQL-like querying functionality	Very large data storage, MapReduce support
Common Use Cases	Transaction processing, interactive transactional applications	Historical data analysis, data warehousing, business intelligence	Used as a cache for storing frequently requested data for a web app	Web apps or any app which needs better performance and scalability without having to define columns in an RDBMS	Real-time data logging as in finance or web analytics
Strengths	Capturing and inputting new records. Robust, proven technology.	Fast query support, especially for ad hoc queries on large datasets, compression	Scalability, very fast storage and retrieval of unstructured and partly structured data	Persistent store with scalability features such as sharding built in with and better query support than key-value stores	Very high throughput for Big Data, strong partitioning support, random read-write access
Weaknesses	Scale issues—less suitable for queries, especially against large databases	Not suited for transactions; import and export speed; heavy computing resource utilization	Usually all data must fit into memory, no complex query capabilities	Lack of sophisticated query capabilities	Low-level API, inability to perform complex queries, high latency of response to queries
Typical Database Size Range		Several GBs to 50 TB	Several GBs to several TBs	Few TBs to several PBs	Few TBs to several PBs
Key Players	MySQL, Oracle, SQL Sever, Sybase ASE	Infobright, Aster Data, Sybase IQ, Vertica, ParAccel	MemCached, Amazon S3, Redis, Voldemort	MongoDb, Couchdb, SimpleDb	HBase, Big Table, Cassandra

Hadoop Architecture



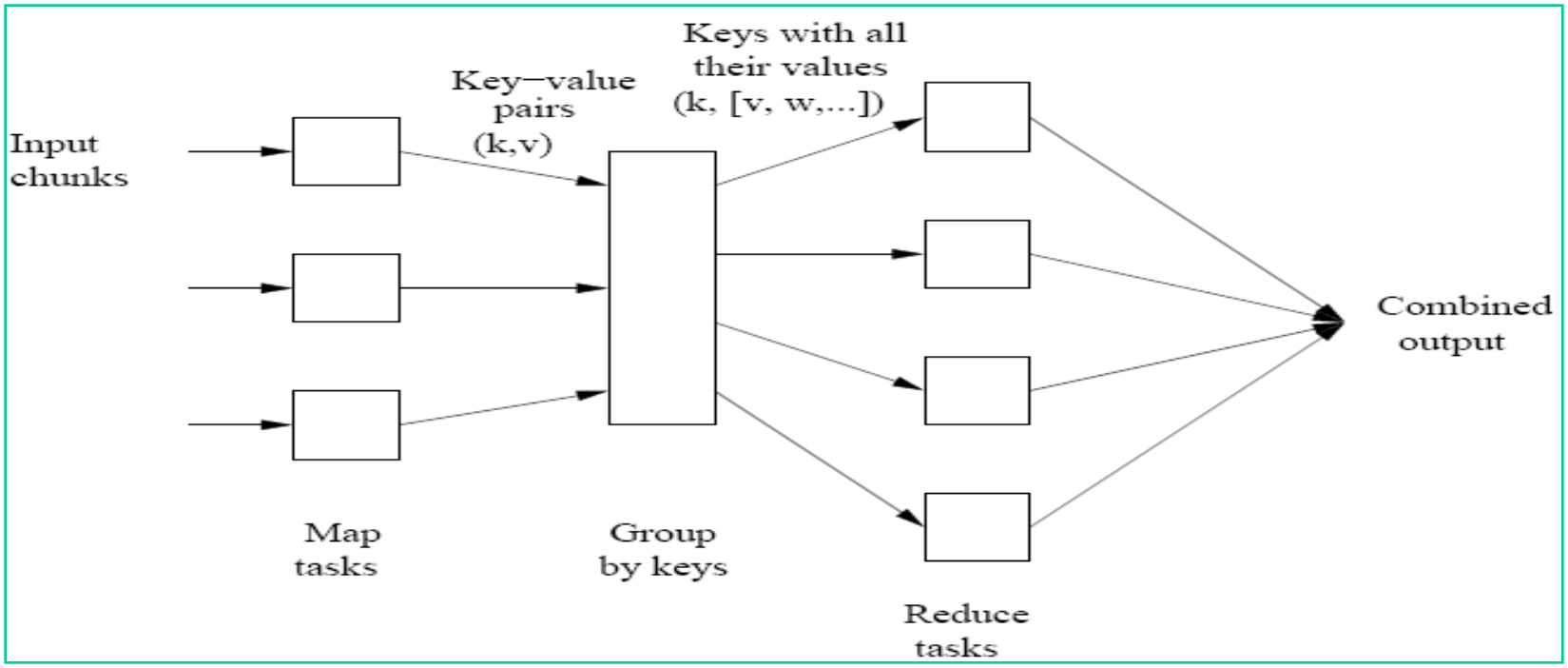
Source: codemphasis.wordpress.com

Big Data Processing Architecture

MAP REDUCE

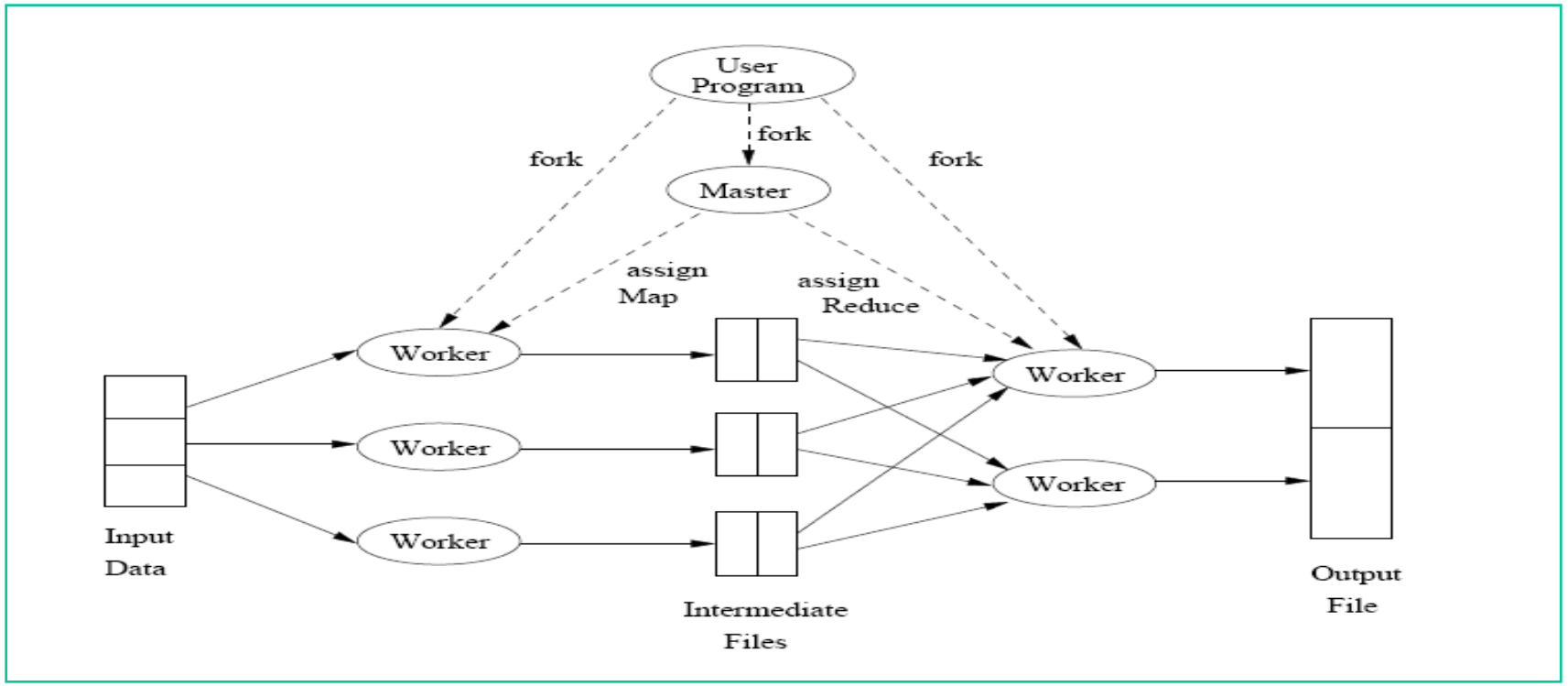
Map Reduce: Google's Invention

- Map: User program that processes input to generate (key, value) pairs
- Reduce: User programs that act on the data sorted on 'key' of Map to generate the output

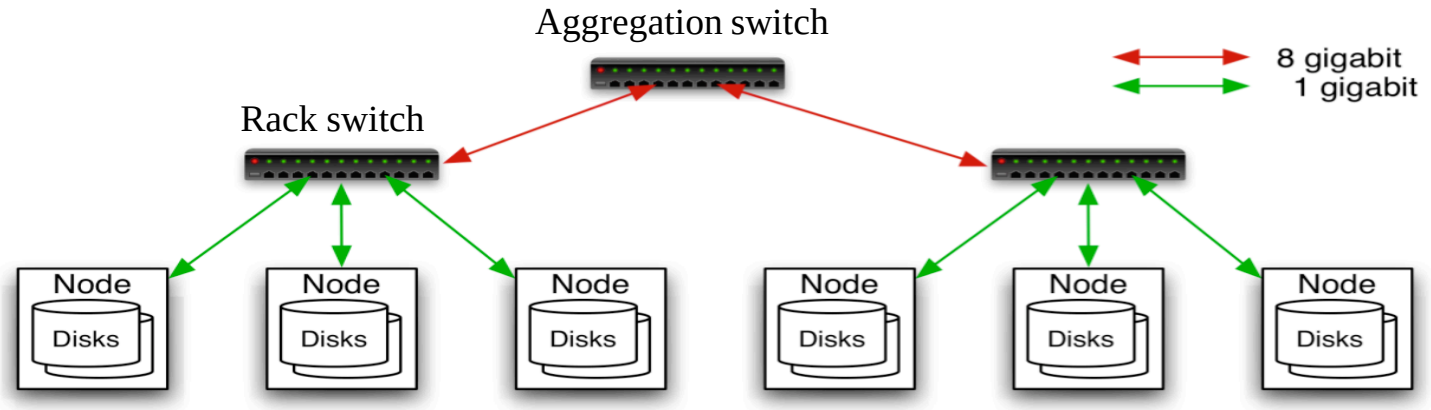


Map Reduce: Physical Architecture

- Worker Node: Can run on commodity hardware
- Master Node: Normal server scale hardware
- Connectivity: Gigabit per second throughput essential

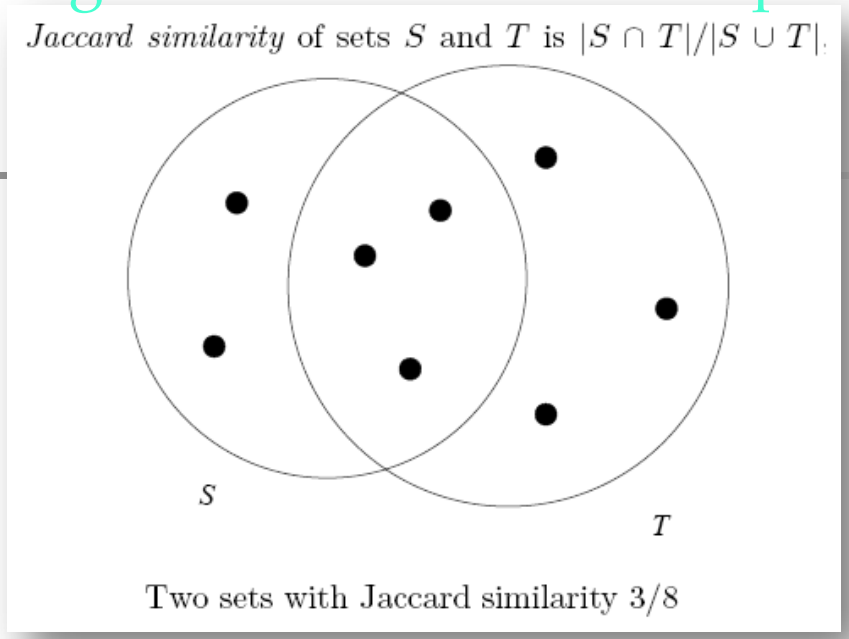


Commodity Hardware



- Typically in 2 level architecture
  - Nodes are commodity PCs
  - 30-40 nodes/rack
  - Uplink from rack is 3-4 gigabit
  - Rack-internal is 1 gigabit

Finding Items with similar Properties (e.g. Plagiarism)



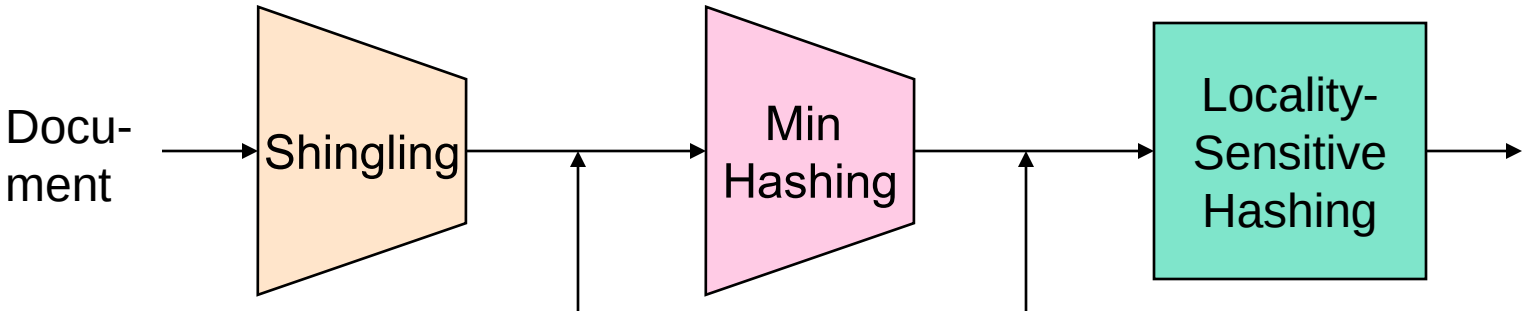
Jaccard Similarity, Shingling, Minhash Signature, Locality Sensitive Hashing, Distances (Eucleadian, non-Eucleadian)

PROBLEM #1 : FINDING SIMILAR ITEMS

Similarity Strategy: Sets with large intersection

- Convert the 'things' to sets of 'items'
  - Documents to k-Shingles (sequence of k-words)
- Find similar large sets that has high intersection
  - Employ minhash to compress large sets to small identifier and use it to evaluate similarity
  - Reduce the number of candidate-pairs to compare for finding similarity of sets using Locality Sensitive Hashing
- Find similar large set that does not lend to intersection of sets using concept of distance measure in arbitrary space

The Big Picture



The set of strings of length ***k*** that appear in the doc-ument

Signatures: short integer vectors that represent the sets, and reflect their similarity

Candidate pairs: those pairs of signatures that we need to test for similarity

Distances

- Eucledian Distance
- Jaccard Distance
- Cosine Distance
- Edit Distance
  - smallest number of insert / delete needed to convert x to y string

$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$[x_1, x_2, \ldots, x_n].[y_1, y_2, \ldots, y_n] = \sum_{i=1}^{n} x_i y_i$$

Distance is used to fine tune within the band the precision of finding the similarity of sets

SIMILARITY = Finding Items with similar Properties (Baskets)
FREQUENT ITEM SET = Finding Baskets with similar subset of Items

Market Basket, Association Rule, A-Priori Algorithm

PROBLEM #2 : FREQUENT ITEM SET

Context: Assumption, Definition

- If a set of items is found in the transaction of baskets more than certain threshold, the set of items is called Frequent Item Set

- Very interesting solution are available following the monotonicity property – A priori algorithm

- Challenge is for very large no. of transactions

Methods for discovering clusters in high dimension very large data

PROBLEM #3 : CLUSTERING

Improve efficiency

- Basic clustering of N points is
  - Step 1: $O(N^2)$
  - Subsequent steps: $O(N^3)$
- Using priority queues
  - Step 1: $O(N^2)$
  - Subsequent steps: $O(N^2 \log N)$
- 2 Families :
  - Hierarchical Clustering (Merge/Stop) and
  - K-Means (point assignment)
- Algorithms
  - Bradley, Fayyad, and Reina (BFR) – variation of K-Means
  - CURE (Clustering Using REpresentatives)
  - GRGPF (V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell, and J. French)
- Using MR to process the clustering
  - One Reduce per cluster

Methods for finding out anomalous events/items in high dimension very large data

PROBLEM #4 : OUTLIER DETECTION

Introduction

- An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism
- An outlier is a pattern in the data that does not conform to the expected behavior
- Outliers are frequently referred to as anomalies.
- Outlier Detection has various real-world applications such as cyber intrusions and credit card fraud

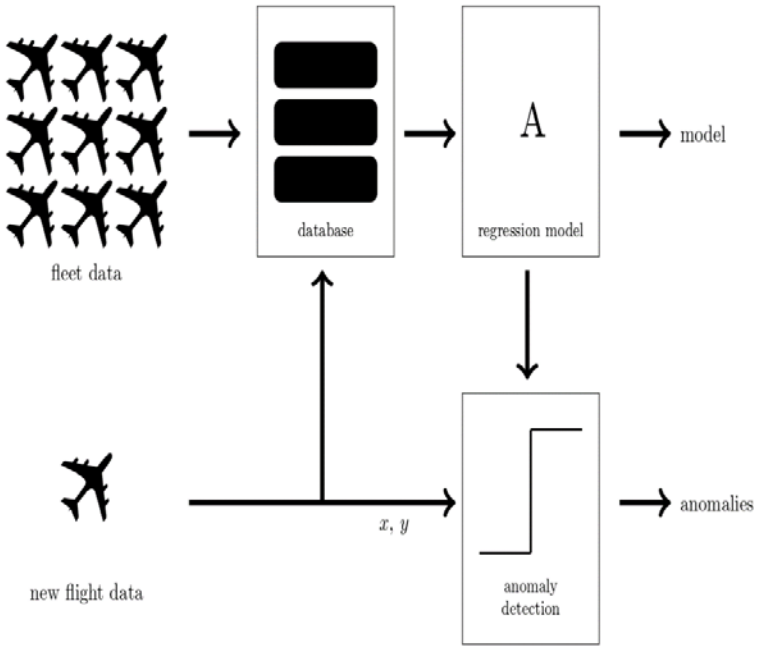
Applications of Outlier Detection

Industrial Damage Detection

- Industrial damage detection refers to detection of different faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
- Example: Aircraft Safety
  - Anomalous Aircraft (Engine) / Fleet Usage
  - Anomalies in engine combustion data
  - Total aircraft health and usage management

Challenges

- Data is extremely huge, noisy and unlabeled
- Detecting anomalous events typically require immediate intervention



Fleet Data Analysis and Anomaly Detection

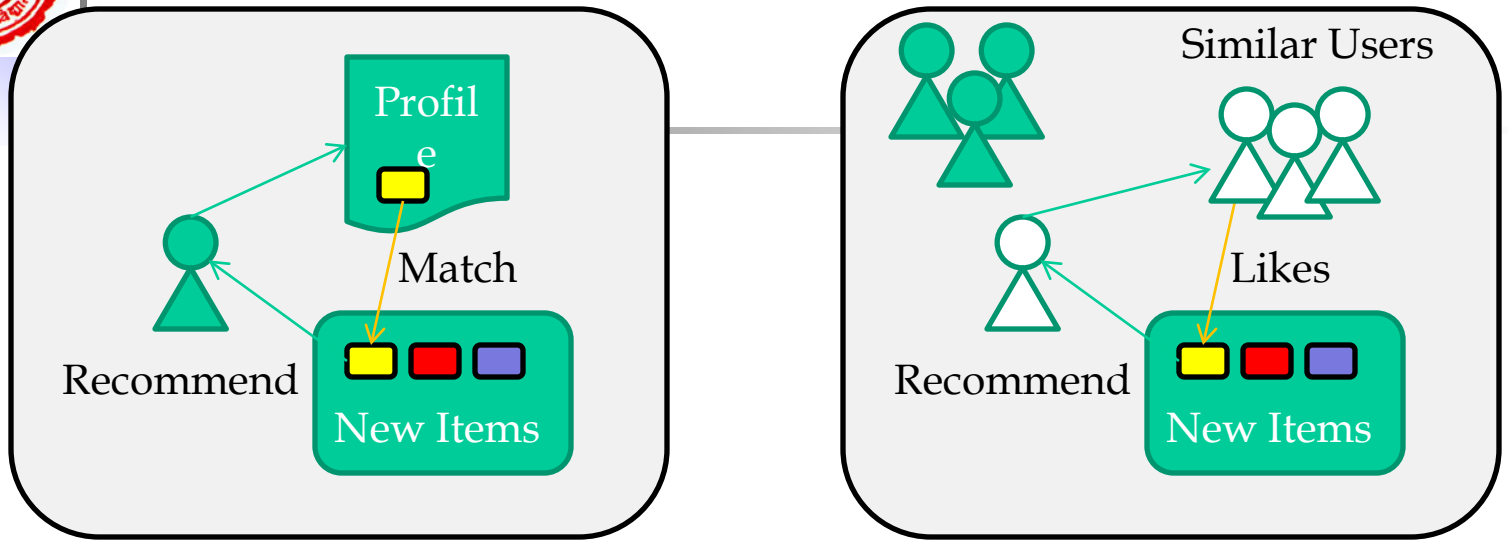
Many web applications today fund themselves through advertisement rather than subscription

PROBLEM #5 : ADVT ON THE WEB

Novelty of the problem: Being online …

- Advertising based on the online behaviour of the user

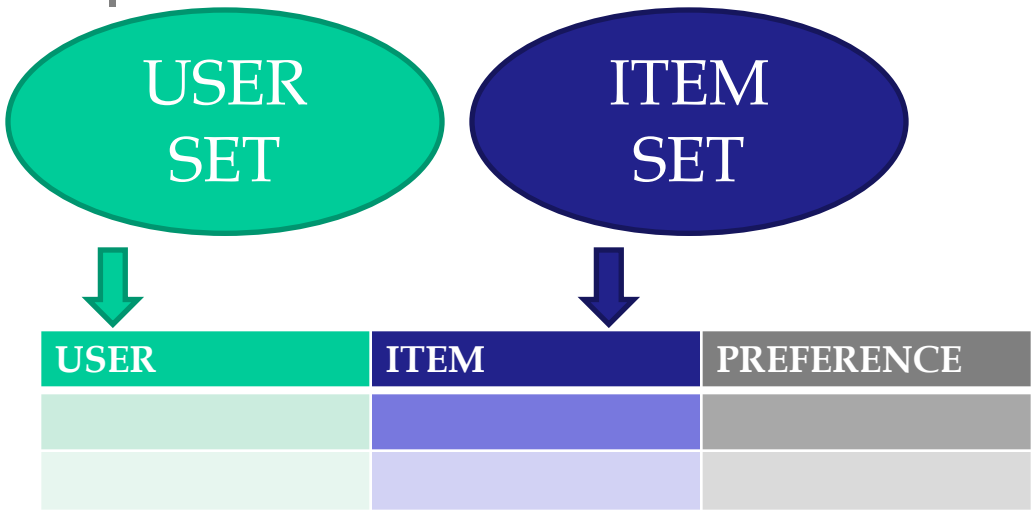
More >>

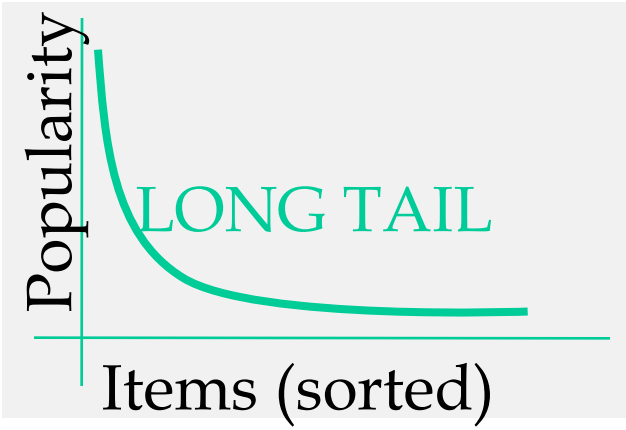


Content Based System, Collaborative Filtering

PROBLEM #6 : RECOMMENDATION SYSTEMS

Utility Matrix: Base Tool





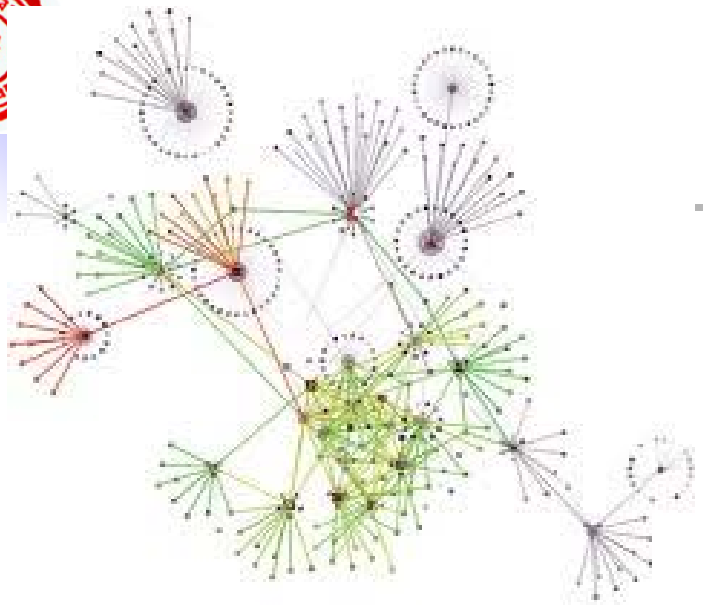
Movie Buffs

Movies

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

HP : Harry Porter **TW** : Twilight **SW** : Star War

Online stores can recommend items in the long tail even if these are lower in popularity charts; physical stores cannot do this



Social Graphs, Betweenness, Bipartite Graph, Partitioning, Laplacian Matrix, Neighborhood

PROBLEM #7 : LARGE GRAPHS

Social Graphs: Key Characteristics

- Large
- Exhibit "locality" or "community" factor
  - Very high density of edges for small subsets of nodes compared to the average density of edges
- Communities are not same as Clusters
  - Individuals belong to many Communities
- Key Challenge: Finding the communities

The art of correlation of information

FROM INFORMATION TO INTELLIGENCE

Analysis / Predictions

- Behavioral
- Social
- Financial
- Medical
- Scientific
- Astronomical

Data Streams: Mining data from the flow

- Challenges / Techniques
  - **Sampling**, without loss of characteristics
  - **Filtering**, selecting the elements that **belong to a set** and discarding the rest
  - **Distinct Elements**, using statistical functions to arrive at counts of distinct elements
  - **Standing Queries**, to "collect" the answers in the fly
  - **Decaying Time Windows**, to weight the properties in the past as a weight of time

Freely available test data

- http://labrosa.ee.columbia.edu/millionsong/
  - The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks

- http://www.tpc.org/tpch/default.asp
  - Transaction Processing Performance Council (TPC) is a non-profit organization founded in 1988 to define transaction processing and database benchmarks and to disseminate objective, verifiable TPC performance data to the industry. TPC benchmarks are used in evaluating the performance of computer systems; the results are published on the TPC web site

Data Science

- CS Background
- Statistics
- Programming
- Machine Learning
- Text Mining / NLP
- Visualization
- Big Data
- Data Ingestion
- Data Munging

References

- 1. Understanding Big data by Zikopou Los, Eaton, deRoos, Deutsch & Lapis, McGrawHill, 2012
- 2. Mining of Massive Data Sets by Rajaraman, Leskovec, Ullman, Stanford University, 2013

THANK YOU !!!