Overview of Big Data Analytics

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

June 02, 2018

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

APPLICATIONS

Management

- Big change in decision making culture
- From HiPPO to Data-driven decisions
- Questions asked
  - What does the data say?
  - Where did the data come from?
  - What is the quality of data?
  - What kind of analyses has been made?
  - What is the confidence of the results of analyses?

E-Commerce

- Advertising for sales promotion

- Targeted advertisements for customer groups

- Personalized promotional offers based on buying pattern, time of transaction and location (use of location data from customers' mobile phones)

- Surveillance Capitalism

Economics

- Drivers:
  - *Data is available in real time*
  - *Data is available at a larger scale*
  - *Data is available on novel types of variables*
  - *Data come with less structure*
- Better Predictive Modeling
- Use of Government Administrative Data for Policy shift, and better and newer citizen services
- Economics of Data Industry

Biology

- Bioinformatics
- Molecular Biology
- Descriptive Ontology
- Evolutionary Developmental Biology
- Gathering huge descriptive data of the object and the environment

Chemistry

- Analytics to look into microbial chemistry and characterize antibiotics and other drug candidates
- Analytical Chemistry
- Computational Chemistry
- Quantum Chemistry
- Medicinal Chemistry

2.5 X 10^{18} bytes (quintillion) / day :: 90% in last 2 years

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.
- IoT, Participatory Sensing will generate huge volumes of data

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

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

How it all started: Google PageRank①

- Intent: Based on search terms, the web pages to be ranked and serviced
  - Term Spam: Web pages had hidden ‘terms’ to push rank
    - PageRank fought with idea of important page based on number of surfers and analyzing the terms in source page near the page link
  - Link Spam: Artificial pages with ‘links’ to push rank
    - TrustRank fought with idea of assigning score based on how many trustworthy page link to a web page
    - Spam Mass [(r-t)/r] closer to 1 indicate probable spam and hence remove from the pages serviced

Note: 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

- Finding Similar Items in very large sets of high-dimensional data
  - Shingling, Minhash Signature, Locality Sensitive Hashing, used in Plagiarism detection, Fingerprint matching
- Frequent Item-set Mining in very large data sets
  - Market Basket Analysis, Association Rule Mining
- Clustering very large high dimensional data sets
  - Discovering clusters in numeric and categorical data sets
- Outlier Detection
  - Finding out anomalous events/items for Intrusion Detection, Fraud Detection

Key Problem Domains: Areas of focus 2

- Advertisement on the Web
  - Funding web applications by advertising and not by subscriptions – Adwords Problem
- Recommendation Systems for Online Stores
  - Content-based filtering, Collaborative filtering
- Mining very large graphs (social graphs)
  - Community detection, CDR analysis

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”

Hadoop Distributed File System (HDFS)

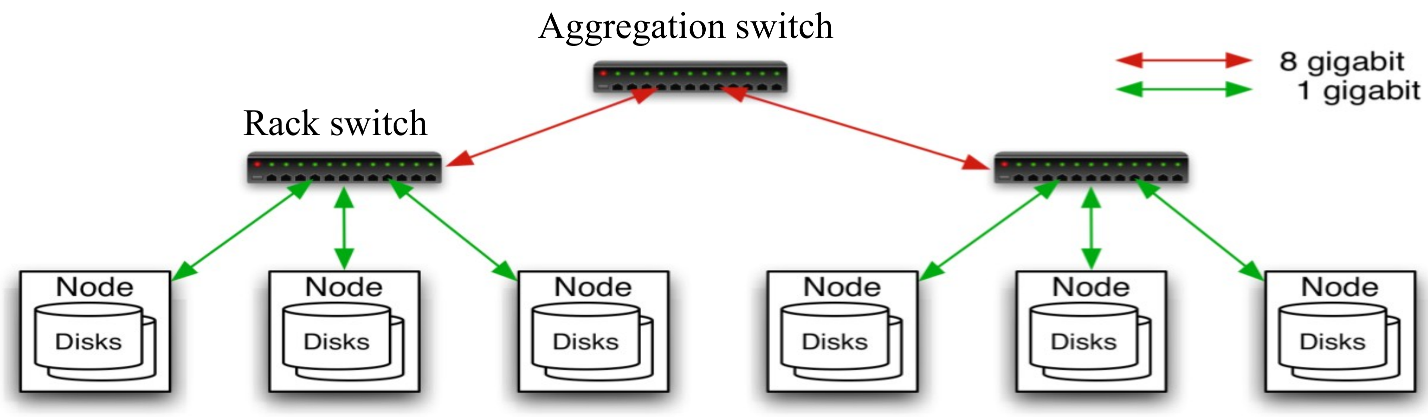
Hadoop - Why ?

- Need to process huge datasets on large clusters of computers
- Very expensive to build reliability into each application
- Nodes fail every day
  - Failure is expected, rather than exceptional
  - The number of nodes in a cluster is not constant
- Need a common infrastructure
  - Efficient, reliable, easy to use
  - Open Source, Apache Licence

Who uses Hadoop?

- Amazon/A9
- Facebook
- Google
- New York Times
- Yahoo!
- Oracle
- …. many more

Commodity Hardware



- Typically in 2 level architecture
  - Nodes are commodity PCs
  - 30-40 nodes/rack
  - Uplink from rack is 8 gigabit
  - Rack-internal is 1 gigabit

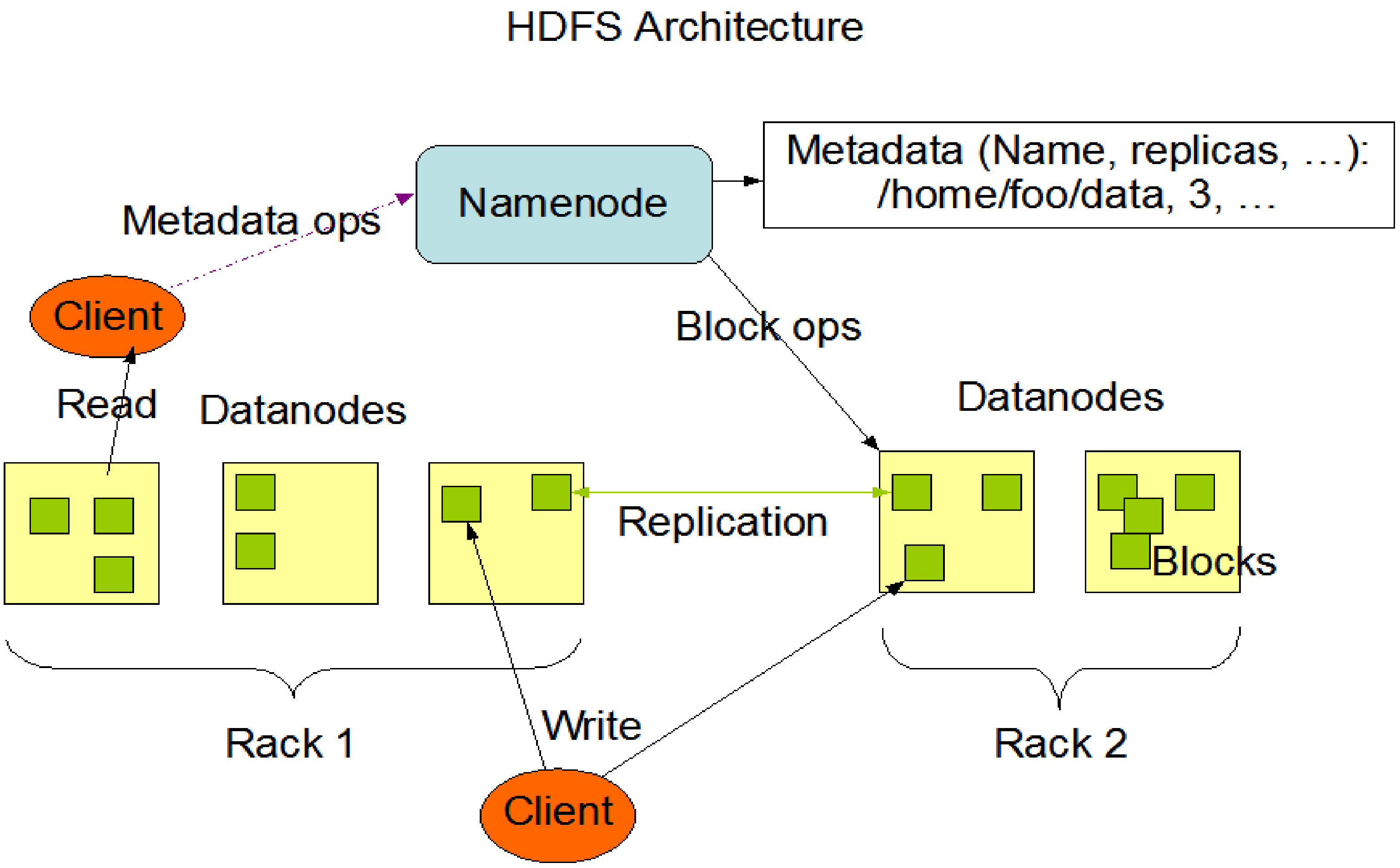
Goals of HDFS

- Very Large Distributed File System
  - 10K nodes, 100 million files, 10PB
- Assumes Commodity Hardware
  - Files are replicated to handle hardware failure
  - Detect failures and recover from them
- Optimized for Batch Processing
  - Data locations exposed so that computations can move to where data resides
  - Provides very high aggregate bandwidth

Hadoop Distributed File System

- Single Namespace for entire cluster
- Data Coherency
  - Write-once-read-many access model
  - Client can only append to existing files
- Files are broken up into blocks
  - Typically 64MB block size
  - Each block replicated on multiple DataNodes
- Intelligent Client
  - Client can find location of blocks
  - Client accesses data directly from DataNode

HDFS Architecture

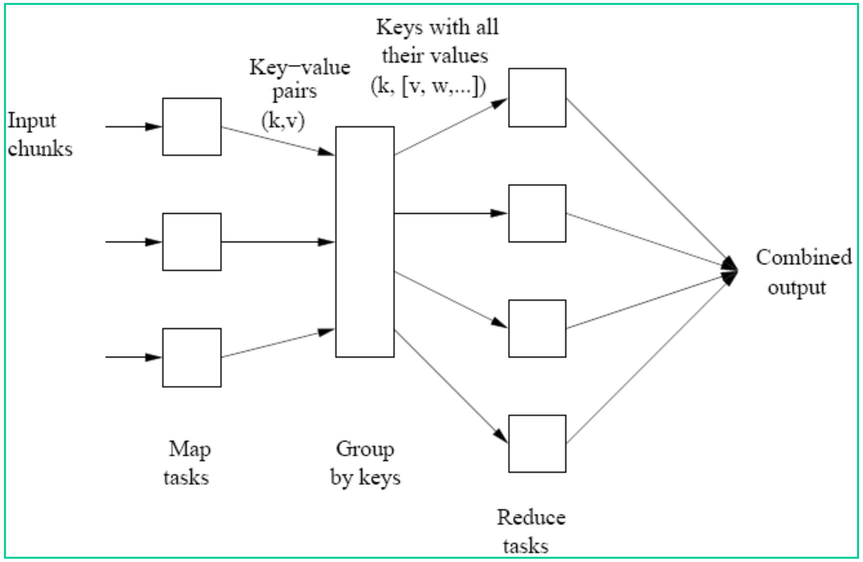


Big Data Processing Architecture

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

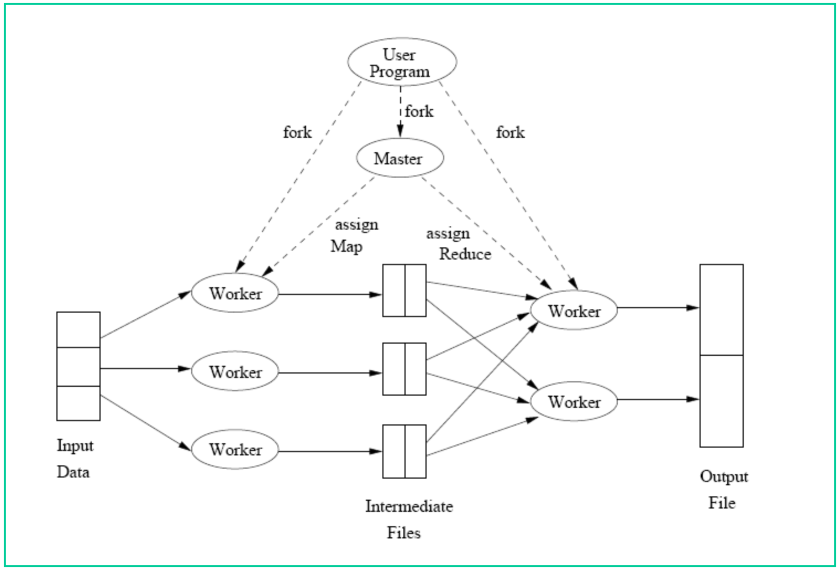


- Matrix Multiplication: MR is ideal for executing very large matrix multiplication[②]
- Relational Algebra supported:
  - Selection, Projection
  - Union, Intersection, Difference.
  - Natural Join.
  - Grouping and Aggregation
- Recursion / Cascading supported

[②] Even if the matrix cannot fit the main memory

Map Reduce: Physical Architecture

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



- Data Block: 64 MB
- Input Data: Replicated across nodes so fault tolerant
- Tasks assigned to Worker: If a task fails need to redo only that task
- No memory between tasks
- Data with same 'Key' to be processed in same Reduce node

Map Reduce: Complexity

- MR Complexity = Processing Cost + Communication Cost
  - Each task is very simple task so Communication Cost dominates
  - Communication Cost is the cost of transporting data from where it is created to where it is used.
- For MR, efficiency of an algorithm is estimated by calculating the sum of the sizes of the inputs to all the tasks

Mathematics

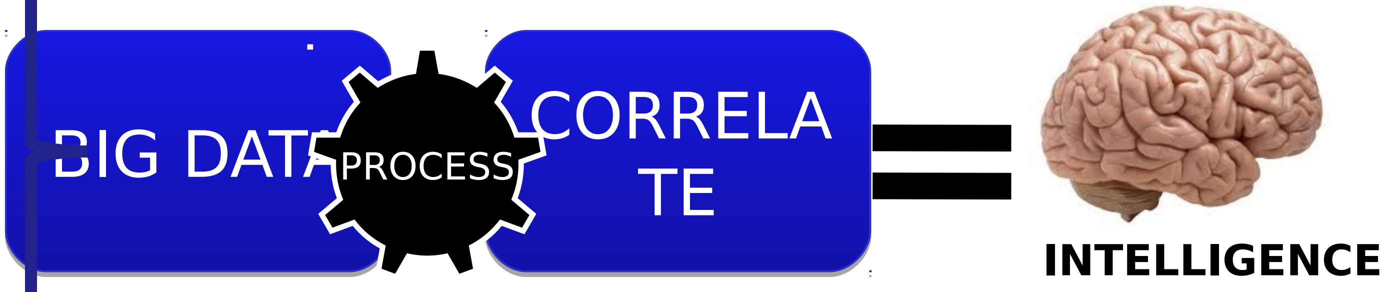
- Mathematics of Big Data
- Big Data ROI = Total Insights / Total Discovery Cost
- Topical Data Analysis - Topological Organization of large data sets to identify areas of persistence and thus relevance
- Persistent homology

The art of correlation of information

FROM INFORMATION TO INTELLIGENCE

Analysis / Predictions

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



Data Streams, Test Data

POSTSCRIPT

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

THANK YOU !!!