


# Big Data Analytics

Sampath Deegalla



7/31/2021

1

## Overview


- What is Big Data?
- Examples
- Characteristics of Big Data
- Challenges and some examples
- Deployment

23/03/2022

2

## What is Big Data?

- Big Data refers to **large data sets** that are computationally analysed to reveal patterns and trends relating to a certain aspect of the data.
- **Elmasri & Navathe** : Big data refers to datasets whose size is beyond the ability typical database software tools to capture, storage, manage and analyze.
- **Edd Dumbill**: Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures.



23/03/2022

3

## Examples



- CLICKSTREAM DATA
- SOCIAL NETWORKS
- SMARTPHONE LOCATION-BASED SERVICES
- WEB SERVER LOGS
- DATA STREAMS FROM INSTRUMENTS
- REAL-TIME TRADING DATA
- BLOGS
- SOCIAL MEDIA SUCH AS TWITTER AND FACEBOOK.

23/03/2022

4

## Characteristics of Big Data

- **Volume:** size of data (Having more data beats out having better models)
  - Sensor data, gene sequencing, remote sensing, environmental monitoring, traffic monitoring, remote monitoring of patients, inventory control using RFIDs
- **Velocity:** speed at which data is created, accumulated, ingested and processed
  - Stock exchanges, real-time and streaming data in Twitter and Facebook

23/03/2022

5

## Characteristics of Big Data (cont.)

- **Variety:** types of sources
  - clickstream and social media, research data (e.g., surveys and industry reports), location, images (e.g., surveillance, satellites and medical scanning), e-mails, supply chain data (e.g., EDI—electronic data interchange, vendor catalogs), signal data (e.g., sensors and RFID devices), and videos (YouTube).
  - Big data includes structured, semi-structured, and unstructured data.
- **Veracity:** credibility of the source and suitability of data
  - Many sources of data is uncertain, incomplete, and inaccurate
  - Requires quality testing and credibility analysis

23/03/2022

6

## Structured, Semi-structured and Unstructured Data

### Unstructured data

The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

### Semi-structured data

```
<University>
<Student ID="1">
  <Name>John</Name>
  <Age>18</Age>
  <Degree>B.Sc.</Degree>
</Student>
<Student ID="2">
  <Name>David</Name>
  <Age>31</Age>
  <Degree>Ph.D. </Degree>
</Student>
....
</University>
```

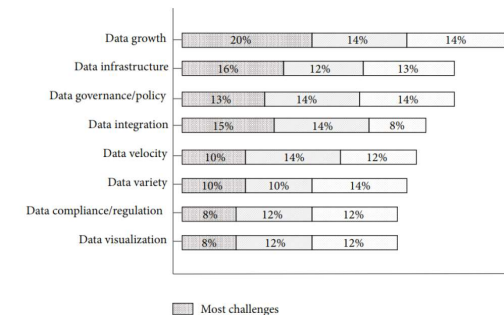
### Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

23/03/2022

7


## Challenges in Big Data



23/03/2022

8

## Rapid growth of unstructured data



Source	Production
YouTube [15]	(i) Users upload 100 hours of new videos per minute (ii) Each month, more than 1 billion unique users access YouTube (iii) Over 6 billion hours of video are watched each month, which corresponds to almost an hour for every person on Earth. This figure is 50% higher than that generated in the previous year
Facebook [16]	(i) Every minute, 34,722 Likes are registered (ii) 100 terabytes (TB) of data are uploaded daily (iii) Currently, the site has 1.4 billion users (iv) The site has been translated into 70 languages
Twitter [17]	(i) The site has over 645 million users (ii) The site generates 175 million tweets per day
Foursquare [18]	(i) This site is used by 45 million people worldwide (ii) This site gets over 5 billion check-ins per day (iii) Every minute, 571 new websites are launched
Google+ [19]	1 billion accounts have been created
Google [20]	The site gets over 2 million search queries per minute Every day, 25 petabytes (PB) are processed
Apple [20]	Approximately 47,000 applications are downloaded per minute
Brands [20]	More than 34,000 Likes are registered per minute
Tumblr [20]	Blog owners publish 27,000 new posts per minute
Instagram [20]	Users share 40 million photos per day
Flickr [20]	Users upload 3,125 new photos per minute
LinkedIn [20]	? 1 million groups have been created

23/03/2022

9

## Facebook

- Current storage = 300 petabytes
- Processed per day = 600 terabytes
- Users per month = 1 billion
- Likes per day = 2.7 billion
- Photos uploaded per day = 300 million

23/03/2022

10

## Google

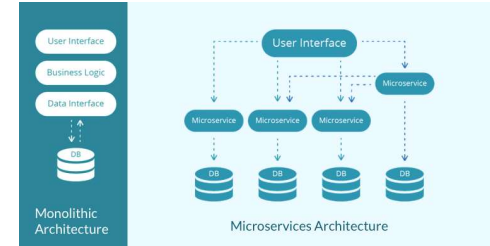
- Current storage = 15 exabytes
- Processed per day = 100 petabytes
- Number of pages indexed = 60 trillion
- Unique search users per month > 1 billion
- Searches per second = 2.3 million

23/03/2022

11

## Systems

- **Monolithic system:** single machine
- **Distributed system:** multiple machines, multiple processors



23/03/2022

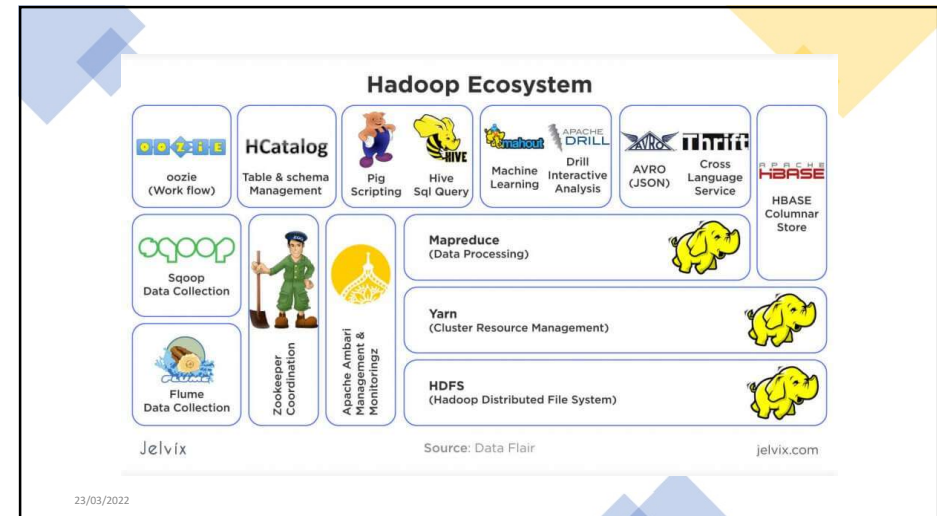
12

## What is Apache Hadoop?

- Hadoop brings the ability to cheaply process large amounts of data (10-100 gigabytes and above)
- How is this different from what went before?
  - Existing enterprise data warehouses and relational databases excel at processing structured data
  - They can store massive amounts of data though at a cost
  - Enterprise data warehouses/relational databases unsuited for agile exploration of massive heterogeneous data
  - The amount of effort required to warehouse data often means that valuable data sources in organizations are never mined

23/03/2022

13



23/03/2022

16

## The Core of Hadoop: MapReduce

- Created at Google in response to the problem of creating web search indexes
- MapReduce takes a query over a dataset, divide it, and run it in parallel over multiple nodes
- Distributing the computation solves the issue of data too large to fit onto a single machine
- Hadoop is an open source MapReduce implementation
- The name “Hadoop” has come to represent this entire ecosystem.

23/03/2022

17

## Hadoop's Lower Levels: HDFS and MapReduce

- MapReduce can be used to distribute computation over multiple servers
- For that, each server must have access to the data. This is the role of **HDFS**, the **Hadoop Distributed File System**. It is a java based file system that provides scalable, fault tolerance, reliable and cost efficient data storage for Big data.
- Servers in a Hadoop cluster can fail and not abort the computation process
- HDFS ensures data is replicated with redundancy across the cluster
- There are no restrictions on the data that HDFS stores. Data may be unstructured and schemeless
- Programming Hadoop at the MapReduce level is a case of working with the Java APIs, and manually loading data files into HDFS

23/03/2022

18

## YARN (Yet Another Resource Negotiator)

- It is a Hadoop ecosystem component that provides the resource management
- YARN is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads
- It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.

Yarn  
(Cluster Resource Management)



23/03/2022

19

## Improving Programmability: Pig and Hive

- Working directly with Java APIs can be tedious and error prone.
  - **Pig** is a programming language that simplifies the common tasks of working with Hadoop. It's built-in operations can make sense of semi-structured data, such as log files.
  - **Hive** enables Hadoop to operate as a data warehouse. Hive do three main functions: data summarization, query, and analysis. It superimposes structure on data in HDFS. It permits queries over the data using a familiar SQL-like syntax.



23/03/2022

20

## Pig and Hive (cont.)

- Pig and Hive provide a higher level interface for working with the Hadoop framework.
- Hive is more suitable for data warehousing tasks. Hive provides an SQL interface on top of MapReduce. Hive's SQL support includes most of the SQL-92 features and many of the advanced analytics features from later SQL standards.
- Pig gives the developer more agility for the exploration of large datasets. It is a thinner layer over Hadoop than Hive, and its main advantage is to drastically cut the amount of code needed compared to direct use of Hadoop's Java APIs.

23/03/2022

21

## Improving Data Access: HBase, Sqoop, and Flume

- Hadoop is a batch-oriented system. Data are loaded into HDFS, processed, and then retrieved. In Big data analytics interactive data access is needed
- **HBase**, a column-oriented database that runs on top of HDFS. It can host billions of rows of data for rapid access.
- **Sqoop** is a tool designed to import data from relational databases into Hadoop, either directly into HDFS or into Hive.
- **Flume** is designed to import streaming flows of log data directly into HDFS.



23/03/2022

22

## HCatalog



- It is a table and storage management layer for Hadoop
- HCatalog supports different components available in Hadoop ecosystems like MapReduce, Hive, and Pig to easily read and write data from the cluster.
- HCatalog is a key component of Hive that enables the user to store their data in any format and structure.
- HCatalog supports RFile, CSV, JSON, sequenceFile and ORC file formats.

23/03/2022

23

## AVRO and Thrift

- Avro provides data serialization and data exchange services for Hadoop.
- Thrift is an interface definition language for RPC(Remote procedure call) communication.
- Hadoop does a lot of RPC calls so there is a possibility of using Thrift for performance.



23/03/2022

24

## Apache Drill

- The drill is the first distributed SQL query engine that has a schema-free model.
- Features
  - Extensibility
  - Flexibility
  - Dynamic schema discovery
  - Drill decentralized metadata



23/03/2022

25

## Ambari

- Ambari is a management platform for provisioning, managing, monitoring and securing apache Hadoop cluster.
- Hadoop management gets simpler as Ambari provide consistent, secure platform for operational control.

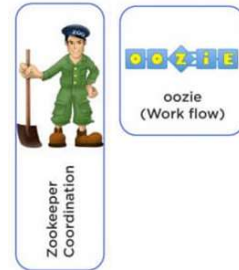


23/03/2022

27

## Coordination and Workflow: Zookeeper and Oozie

- **Zookeeper** manages and coordinates a large cluster of machines. As computing nodes can come and go, members of the cluster need to synchronize with each other, know where to access services, and know how they should be configured.
- **Oozie**: This is a service for scheduling and running workflows of Jobs; individual steps can be MapReduce jobs, Hive queries, Pig scripts, and so on.



23/03/2022

28

## Management and Deployment: Ambari and Whirr

- Ambari is intended to help system administrators deploy and configure Hadoop, upgrade clusters, and monitor services. Through an API, it may be integrated with other system management tools.
- Whirr is a highly complementary component. It offers a way of running services, including Hadoop, on cloud platforms. Whirr is cloud neutral and currently supports the Amazon EC2 and Rackspace services.

23/03/2022

29

## Machine Learning: Mahout

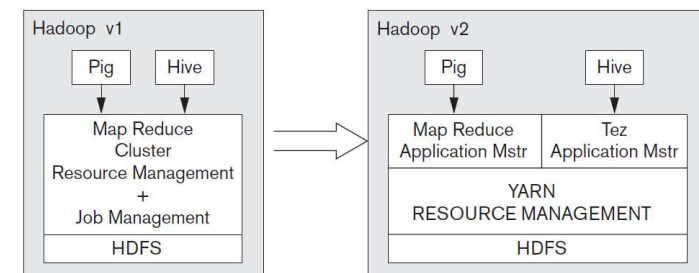


- Mahout is open source framework for creating scalable machine learning algorithm and data mining library.
- Once data is stored in Hadoop HDFS, mahout provides the data science tools to automatically find meaningful patterns in those big data sets.
- The Mahout project is a library of Hadoop implementations of common analytical computations. Use cases include user collaborative filtering, user recommendations, clustering, and classification.

23/03/2022

30

## Hadoop V1 vs Hadoop V2



23/03/2022

31

## Deployment: Dimensions to consider

- Cloud or in-house: three forms of solutions: software, appliance or cloud-based. Many prefer hybrid solution: OnDemand cloud resources to supplement in-house deployment
- Big data is big: too big to process conventionally and too big to transport anywhere
- Big data is messy: 80% of the effort involved in cleaning the data
- Culture: organizational willingness to understand and use data for advantage
- Know where you want to go: decide what problem you want to solve

23/03/2022

## References

- S. Pyne, B.L.S.P. Rao, and S.B. Rao: "Big Data Analytics: Methods and Applications", Springer, 2016.
- Big Data Now, O'Reilly Media, 2012.
- B. Schmarzo, "Big Data MBA", Wiley, 2015
- <https://www.kdnuggets.com/2015/02/how-big-data-pieces-technology-fit-together.html>
- <http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf>

23/03/2022