Introduction to Big Data and Hadoop

Year 3 – Semester 2

B.Sc. (Hons) in Information Technology Specializing in Software Engineering

Contents

- Introduction to Big Data
- Types of Data
- Big Data V's
- Big Data Technologies
- Introduction to Apache Hadoop
- Hadoop Architecture
- Key features of Hadoop
- Hadoop History
- Hadoop Ecosystem

Introduction to Big Data

Big Data

- Big Data' is also a data but with a huge size.
- "Big Data" is a term used to describe collection of data that is huge in size and yet growing exponentially with time.
- In short, such a data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

From Data to Big Data

Bit: 1or0

Nibble : 4 bits

Byte: 8 bits

KiloByte: 1024 Bytes (2_{10})

MegaByte: 1024 Kb = 1024 x 1024 (2_{20})

GigaByte: 1024MB (2_{30})

TeraByte: 1024GB (2_{40})

PetaByte: 1024TB (2_{50})

ExaByte: 1024PB (2_{60})

ZettaByte: 1024EB (2_{70})

YottaByte: 1024ZB (2_{80})

Types of Data

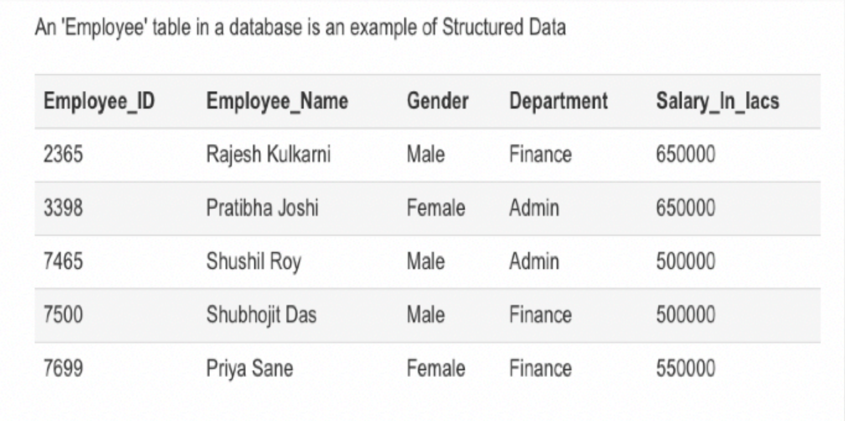
Structuring Big Data

On the basis of the data received from the sources, data can be:

- Structured
- Unstructured
- Semi-structured

Structured data

Structured data can be defined as the data that has defined “repeating pattern”. This pattern makes it easier for any program to sort, read, and process the data.



An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Structuring Big Data

Unstructured Data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Few examples:

- audio
- video
- word/pdf files
- digital images

challenges with unstructured data:

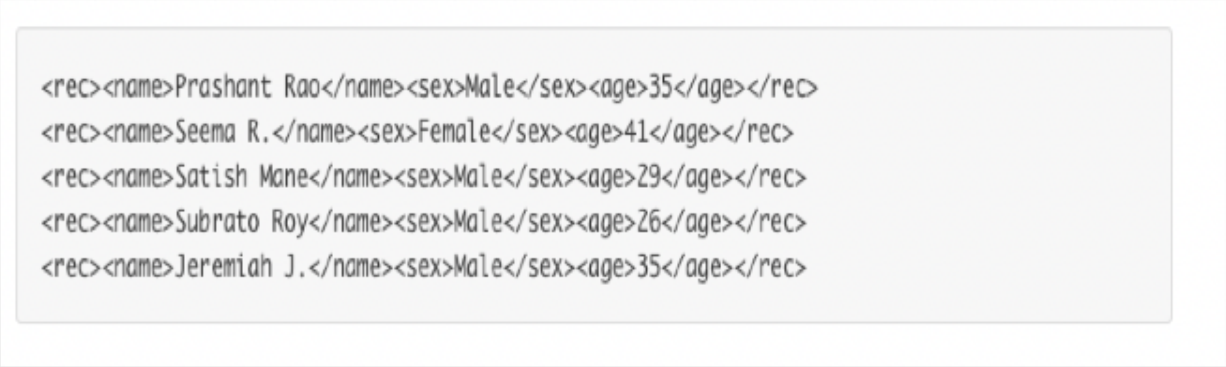
- organization
- processing
- costing in terms of storage space and human resource.

Structuring Big Data

Semi-structured Data

It a self describing structure that contains tags or markup elements in order to separate elements.

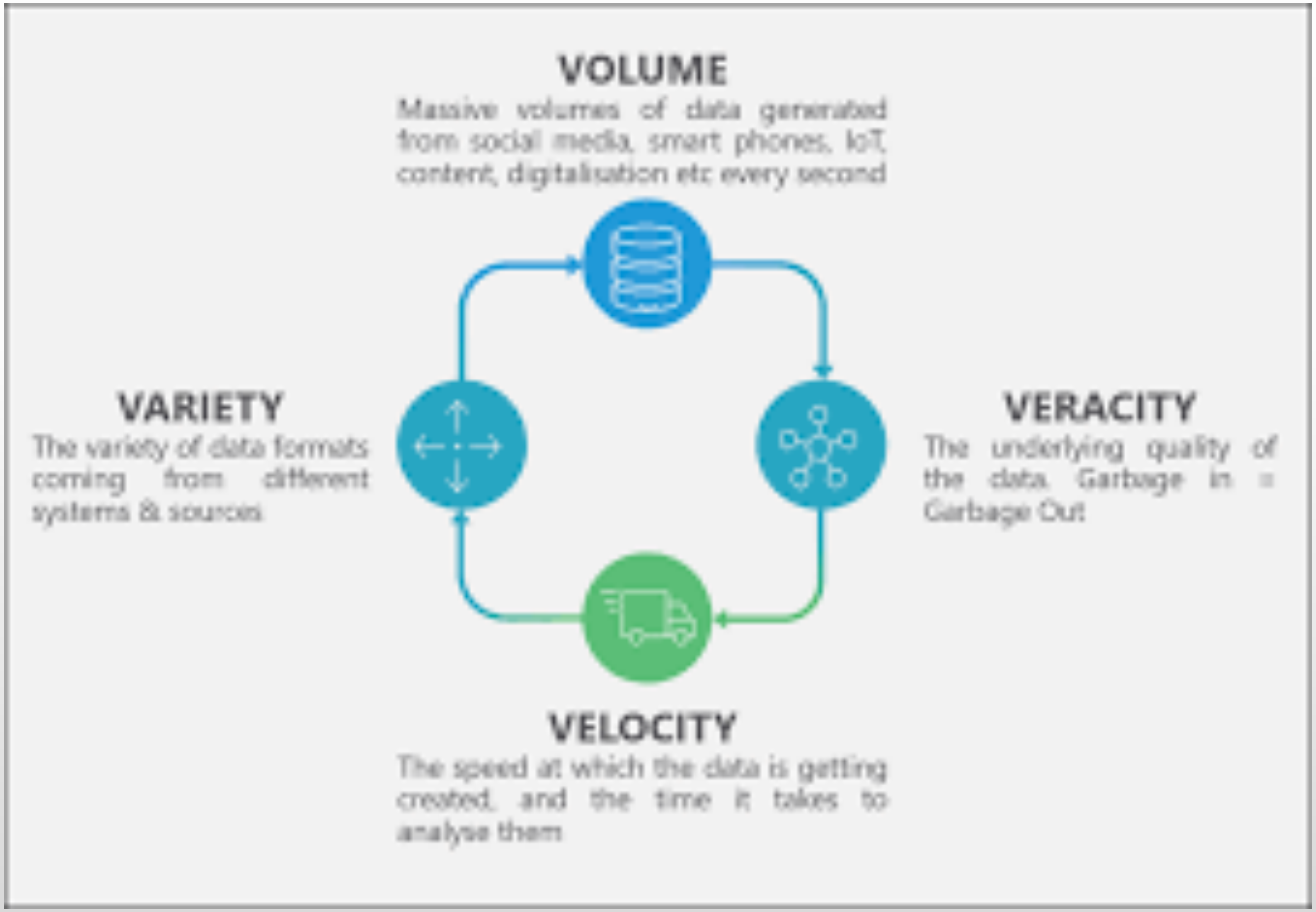
- web data in the form of cookies
- data exchange formats such as JSON data.



```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Big Data 4 V’s

4 V’s



VOLUME

Massive volumes of data generated from social media, smart phones, IoT, content, digitalisation etc every second

VARIETY

The variety of data formats coming from different systems & sources

VERACITY

The underlying quality of the data. Garbage in = Garbage Out

VELOCITY

The speed at which the data is getting created, and the time it takes to analyse them

Big Data Technologies

Big Data Technologies

To handle Big data, several technologies are developed. The most effective innovations are in the field of distributed and parallel computing.

Hadoop:
Open source platform that provides analytical technologies and computational power required to work with large volume of data.

IMC:
In IMC technology, The RAM is used for analyzing data. The volume related issues are handled by using IMC, and the diversity of big data is taken care by NoSQL.

Big data cloud:
Companies like Google and Amazon provides cloud based services to contribute in the field of Big data processing. The key feature of cloud is "elasticity".

Introduction to Apache Hadoop

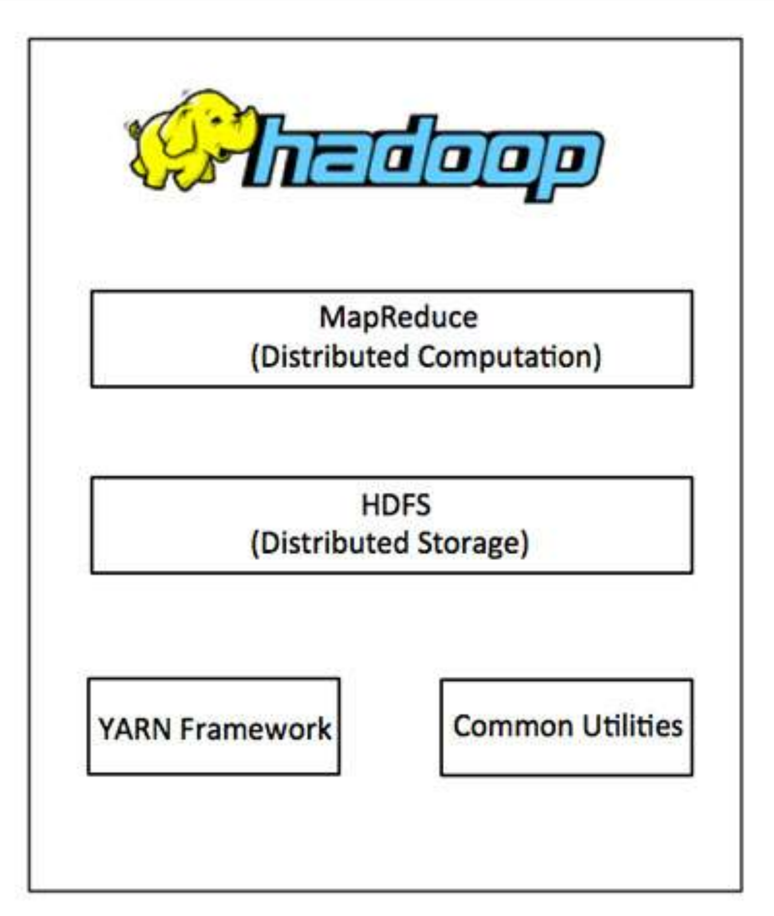
What is Apache Hadoop

It is an open source platform that provides analytical technologies and computational power required to work with Big Data. Some key points about Hadoop:

- Solution for Big data
    - Deal with complexities of big data such as Volume, Variety, Veracity and Velocity.
- Set of open source projects. It is an ecosystem.
- Provides improved programming model used to create and run distributed system quickly and efficiently.
- Transforms commodity software into a service that:
    - that stored PB of data reliably.
    - Allows huge distributed computations.
- Key attributes:
    - Redundant and Reliable
    - Full access of data.
    - Batch processing centric
    - Easy to program
    - Runs on commodity hardware

Hadoop Architecture

Hadoop Architecture



hadoop

MapReduce
(Distributed Computation)

HDFS
(Distributed Storage)

YARN Framework

Common Utilities

Key features of Hadoop

- Easy programming. User does not have to worry about:
  - Where the file is located.
  - How to manage failures.
  - How to break computation in pieces.
  - How to program for scaling
- It follows client-server architecture.
- Efficiency of Hadoop remains up to the mark both in cases of complex and
- large data.
- Add/remove nodes dynamically.
- Ability to adjust change without causing any interrupts in the system.
- It is highly reliable. Reliability is obtained by replacing the data on multiple host(default replication is 3).
- The processing capability is always linear.

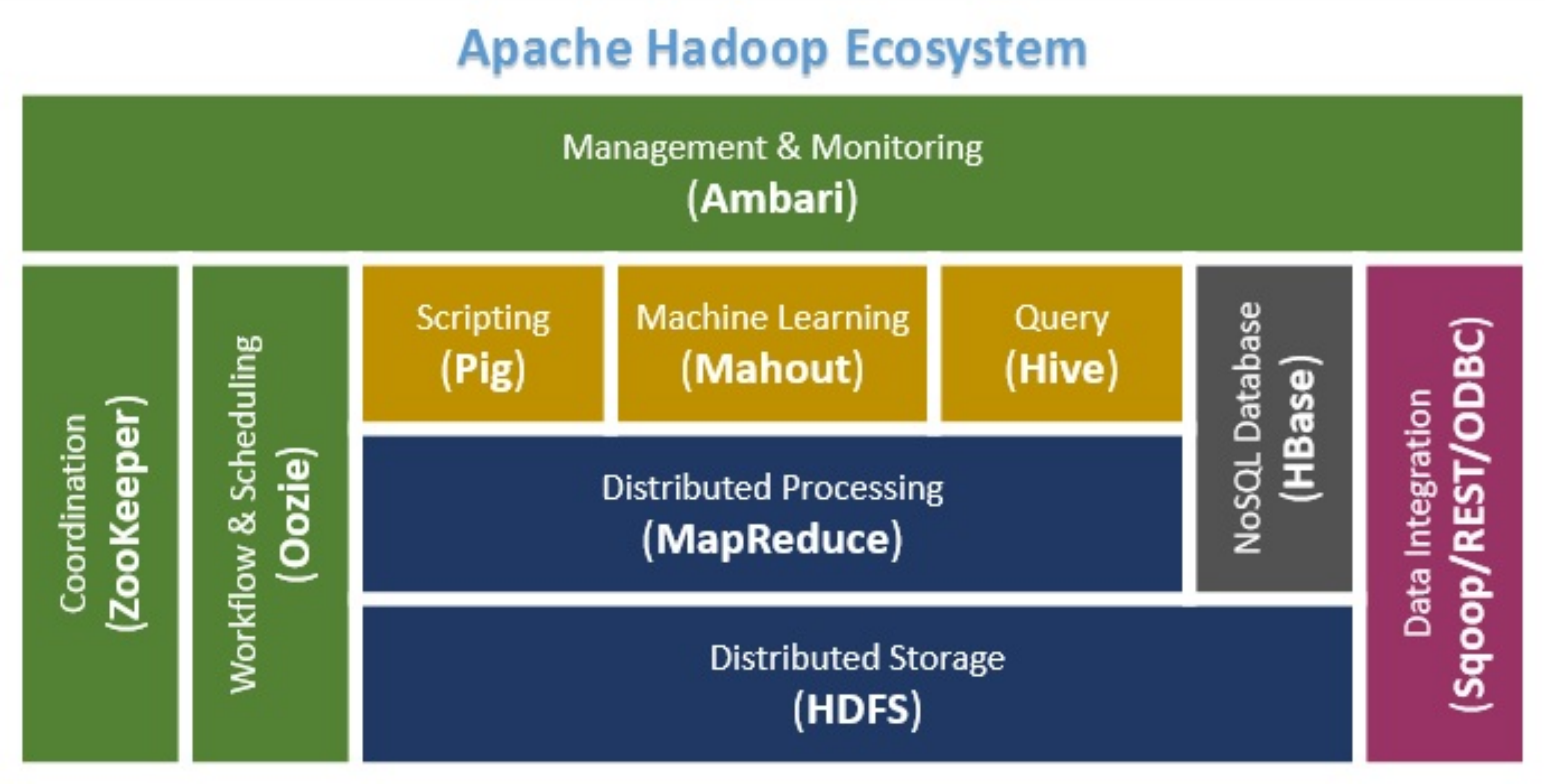
Hadoop History

- Hadoop was created by Doug Cutting, the creator of Apache Lucene, the widely used text search library. Hadoop has its origins in Apache Nutch, an open source web search engine, itself a part of the Lucene project.
- The name Hadoop is not an acronym; it’s a made-up name. The project’s creator, Doug Cutting, explains how the name came about:

> The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid’s term.

Hadoop Ecosystem

Hadoop Ecosystem



Apache Hadoop Ecosystem

Management & Monitoring
(Ambari)

Coordination
(ZooKeeper)

Workflow & Scheduling
(Oozie)

Scripting
(Pig)

Machine Learning
(Mahout)

Query
(Hive)

NoSQL Database
(HBase)

Distributed Processing
(MapReduce)

Distributed Storage
(HDFS)

Data Integration
(Sqoop/REST/ODBC)

Hadoop Ecosystem

Following are the components that collectively form a Hadoop ecosystem:

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE:** Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib:** Machine Learning algorithm libraries
- **Solar, Lucene:** Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling

Thank you …