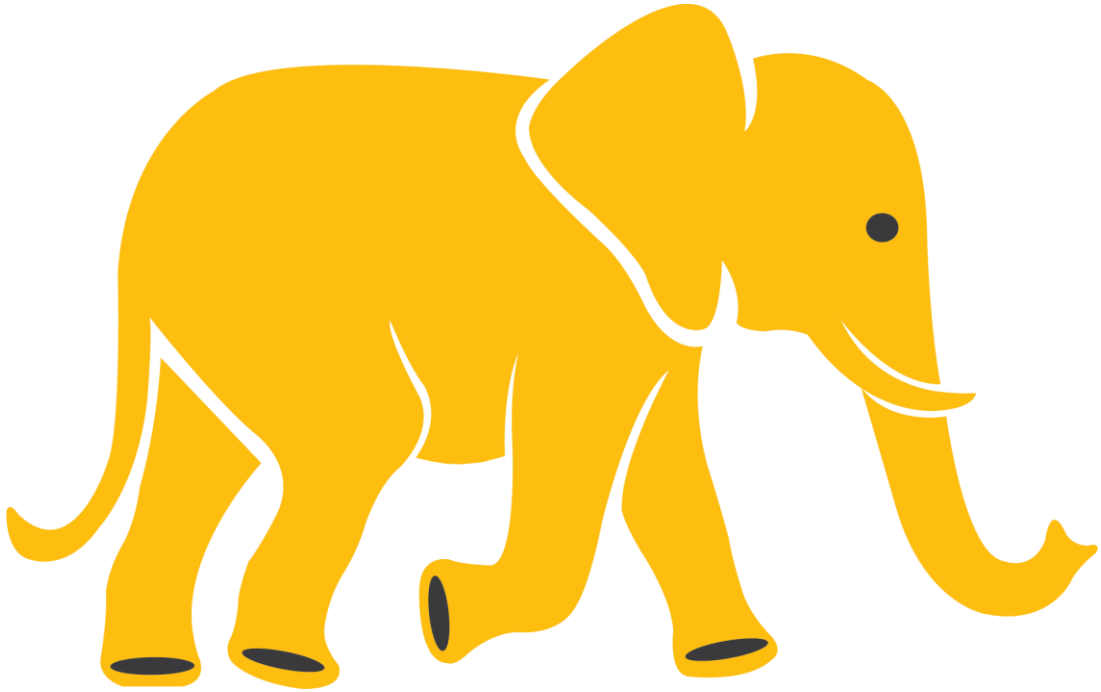


Big Data & Hadoop

Module : 1 Introduction to Big Data and Hadoop

Topics

MODULE 1
BIG DATA AND HADOOP INTRODUCTION

MODULE 2
HDFS ARCHITECTURE, HADOOP CONFIGURATIONS & DATA LOADING

MODULE 3
INTRODUCTION TO MAP REDUCE

MODULE 4
ADVANCED MAP REDUCE CONCEPTS

MODULE 5
INTRODUCTION TO PIG AND ADVANCE PIG

MODULE 6
INTRODUCTION TO HIVE AND ADVANCE HIVE

MODULE 7
INTRODUCTION TO HBASE AND ADVANCED HBASE

MODULE 8
SQOOP AND FLUME

MODULE 9
BASIC OOZIE AND OOZIE CONFIGURATION

MODULE 10
PROJECT DISCUSSIONS

Session Objectives

This Session helps you to understand :

- What is Big Data?
- Use cases and Challenges associated with Big Data
- Difference between Real time and Batch Processing
- Hadoop definition and its characteristics
- Hadoop core component in Hadoop 1.x and Hadoop 2.x
- Hadoop echo system in Hadoop 1.x and Hadoop 2.x

What is Big Data?

- Huge Amount of Data (Terabytes or Petabytes).
- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, storage, search, sharing, transfer, analysis and visualization.

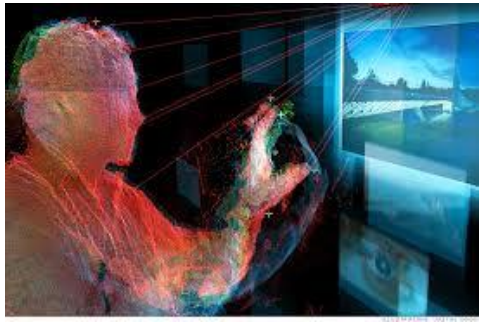
Who is generating Big Data ?

Social Media and Network

ATM

Mobile Devices

Scientific Instruments



Sensor Technology and Networks

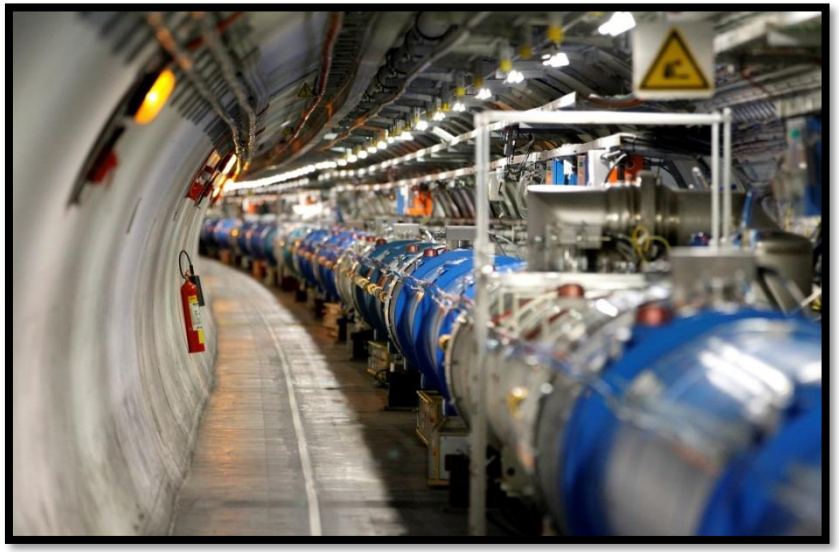
Facebook

- It has about a 1,000,000,000 + (a billion) daily active users
- Generates close to 500,000 GB of data per day, 15,000,000 GB data per month, 180,000,000 GB per year
- Executes 70,000 queries on that data every day

Use Cases of Big Data Technology

- Biggest retail store chain in the world with 11,500 stores in 28 countries around the world under 65 different banners globally
- 1,000,000 (a million) transactions every hour, which is 8,000,000 per day and 240,000,000 per month and 2,880,000,000 transactions per year
- It generates 2,560,000 GB / day, 76,800,000 GB / month and 921,600,000 GB / year
- It generates 2,560 TB / day, 76,800 TB / month and 921,600 TB / year

Other Examples



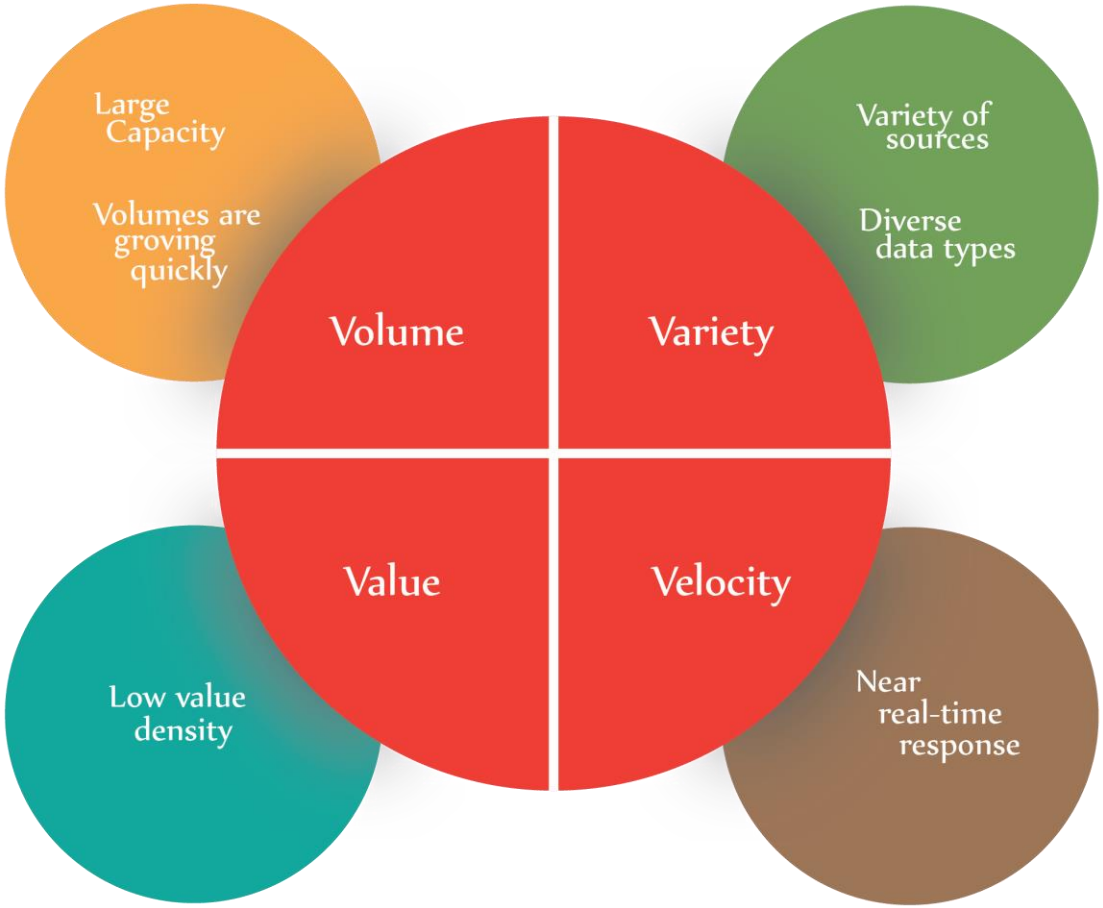
NASA Centre for Climate Simulation (NCSS) stores 32 Petabytes of data comprising of climatic observations

Large Hadron Collider (LHC) produces 30 GB data every day

Other Examples

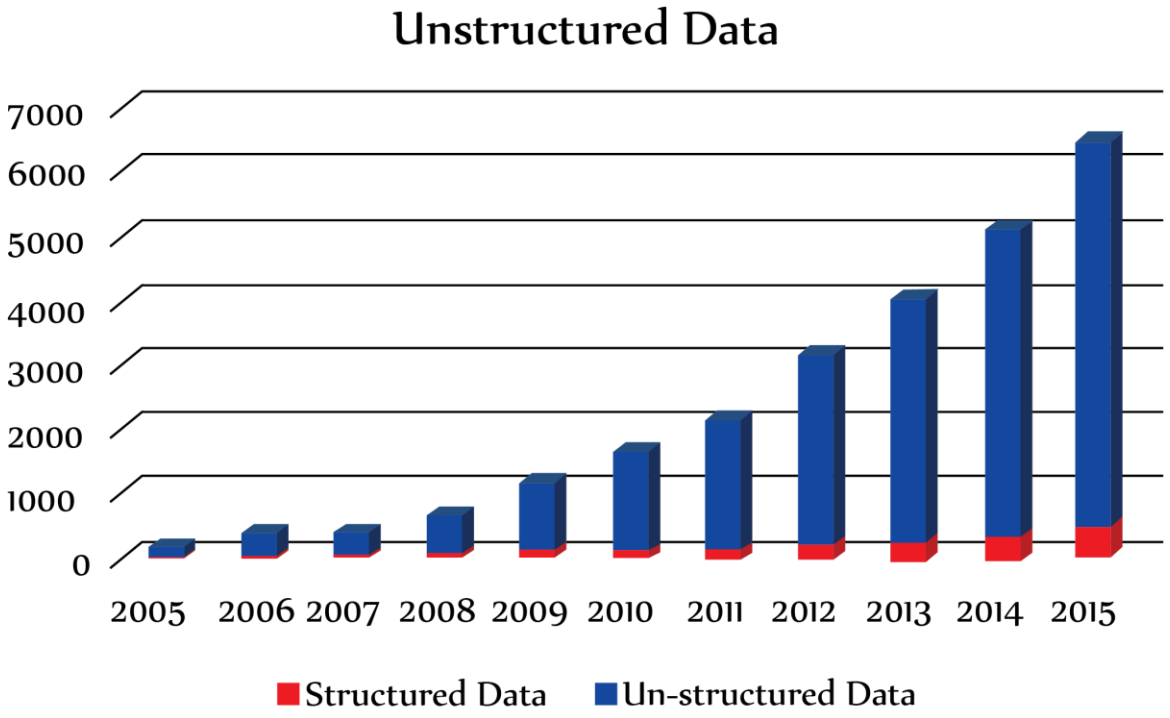
It was also in parts responsible for the BJP and its allies to win a highly successful Indian General Election 2014.

IBM Definition of Big Data



Structured and Unstructured Data

- The volume of digital data will grow 40% to 50% per year
- By 2020, IDC predicts the number will have reached 40,000 EB, or 40 Zeta bytes (ZB)



eBay Case Study

- Ebay had an existing enterprise data warehousing solution based on relational databases.
- It was working well as long as the data (to the tunes of 50TB) was structured
- Data became more and more unstructured as the customer journeys were being saved now
- The Customer Journey data was huge (100s of Peta bytes). So the challenge for ebay was:
  - Either to give structure to this data and loose most of it (Store 1% and loose rest)
  - Or keep everything and make it un manageable
- Cost element was also there with huge cost of software and hardware licensing.
- High end configuration hardware required on the hardware front too. Moreover, the scale up / down had an additional cost

eBay Case Study (Cont'd)

ebay came up with a three pronged approach:

1 Developed a highly efficient traditional Tera data based enterprise data warehousing platform (RDBMS based) which was extremely reliable for it immediate / urgent analytical needs.

This processes 50 TB of data (keeps the data that is needed for urgent analysis, moves the rest to second part), accessed by 7000 analysts with 700 concurrent users (Fairly expensive but highly responsive)

2 Custom data ware house based on Tera data, built on commodity machines.

This database was called Singularity and was saving the entire dataset (structured or Unstructured).

This met the requirement of keeping all the data at a cheaper cost

eBay Case Study (Cont'd)

3 long with the above two, eBay also created a Hadoop implementation with 20,000-node cluster with 80PB capacity which provide all tools (read pig, hive, HBase) to analyze output coming out from 1sttwo parts

PayPal Case Study

- Paypal had an existing Market Capitalization.
- It was working well as long as the data (to the tunes of 50TB) .
- Data became more and more as the customer journeys were being saved now.
- The Customer Journey data was huge (1.1 Peta bytes). So the challenge for PayPal was:
  - PayPal lets all the data to go, as it was difficult to catch-all schema types on traditional databases.
  - Or keep everything and make it un manageable
- Cost element was also there with huge cost of software and hardware licensing.
- High end configuration hardware required on the hardware front too.
- Moreover, the scale up / down had an additional cost .

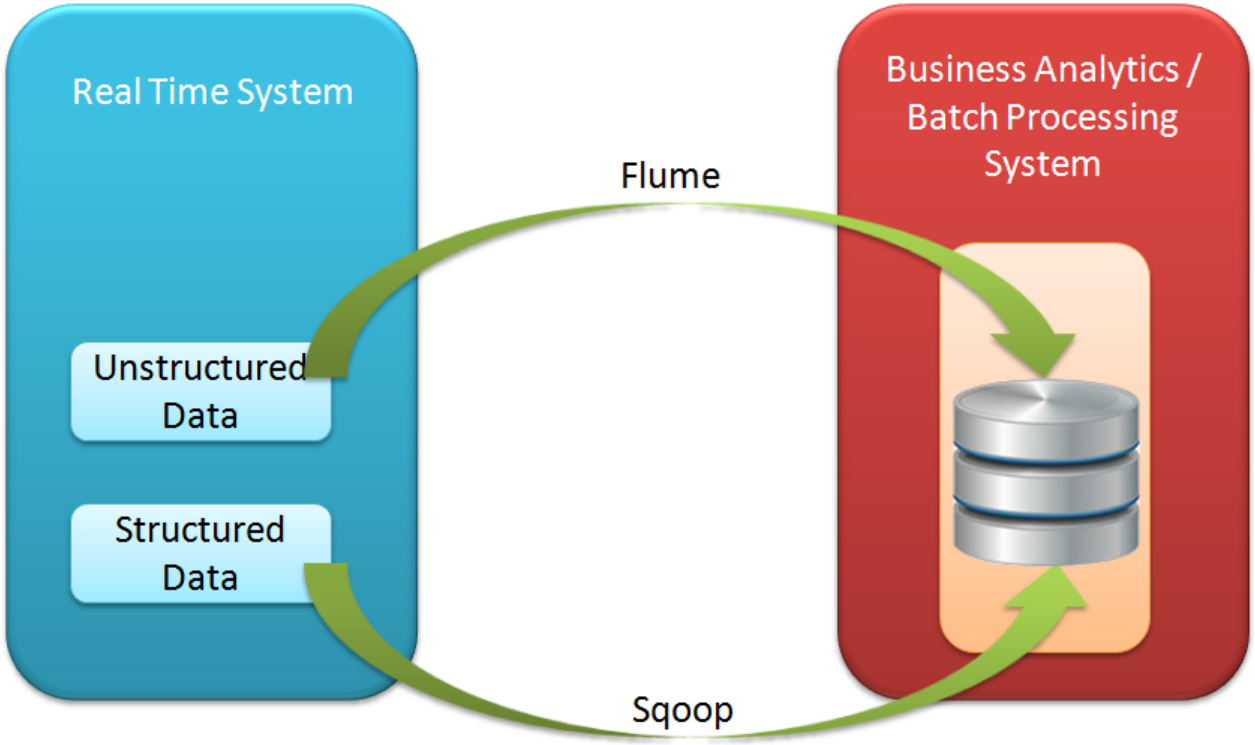
PayPal Case Study (Cont'd)

- PayPal collects more than 20 terabytes of log data every day for sentiment analysis, event analytics, customer segmentation, recommendation engine and sending out real-time location based offers.
- PayPal is building new and modifying the existing fraud analytics system by incorporating various open source technologies like Hadoop and spark, applying machine learning algorithms, online caching and human detectives.
- PayPal expands its Hadoop usage into HBase to leverage HDFS. HDFS also acts as the storage layer for HBase for reading and writing - to large unstructured datasets.

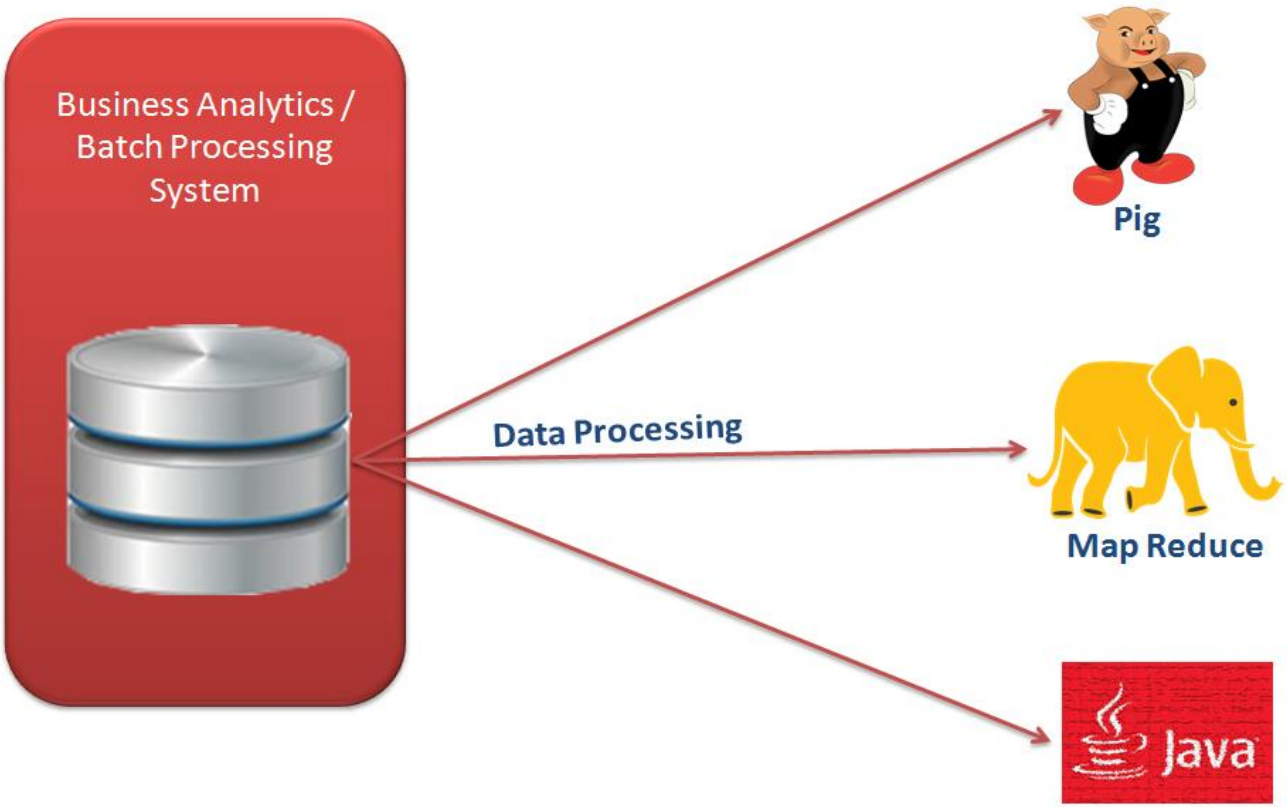
Batch Processing

- Processing large amount of data transactions in a group or batch
- Following three phases are common to batch processing or business analytics project, irrespective of the type of data (structured or unstructured)

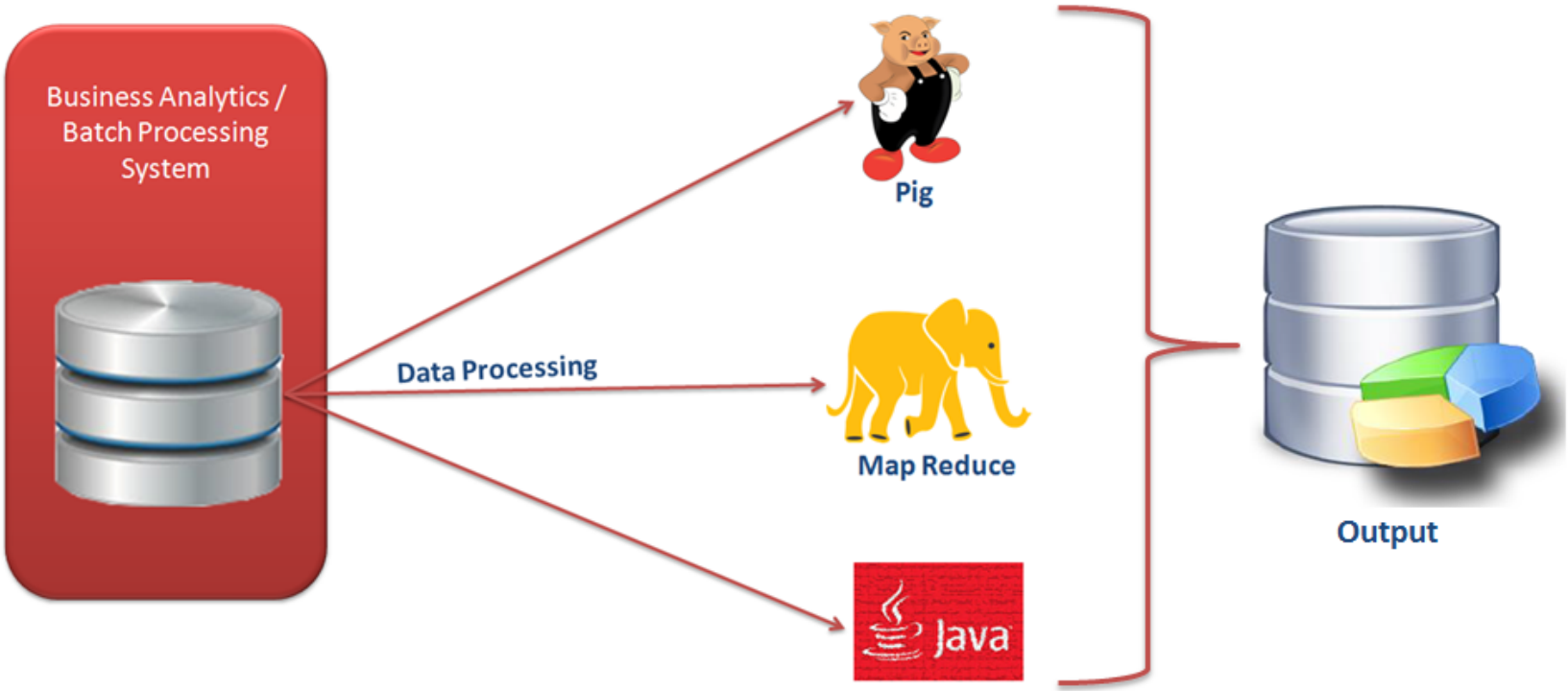
Data Collection



Data Preparation

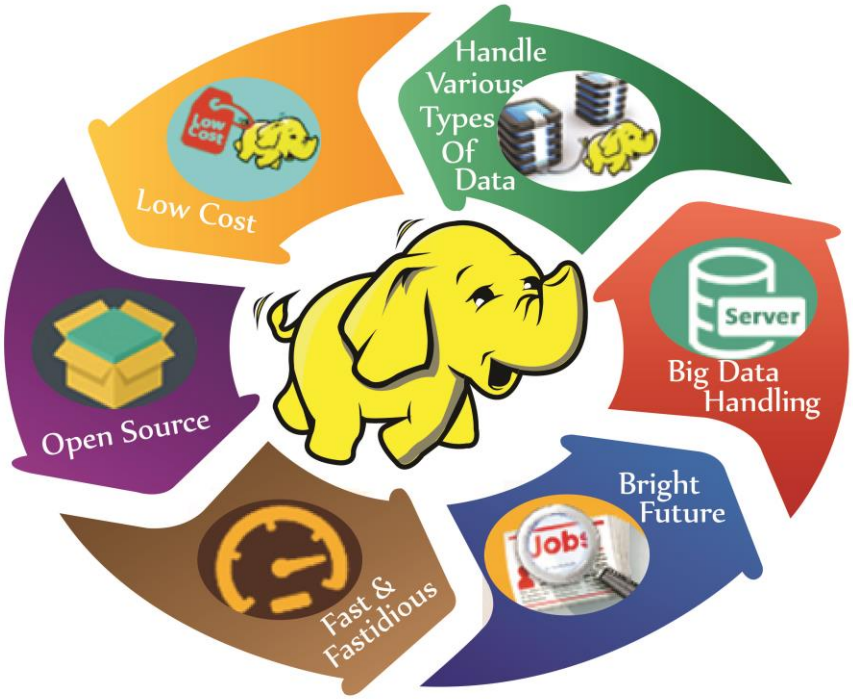


Data Presentation



What is Hadoop?

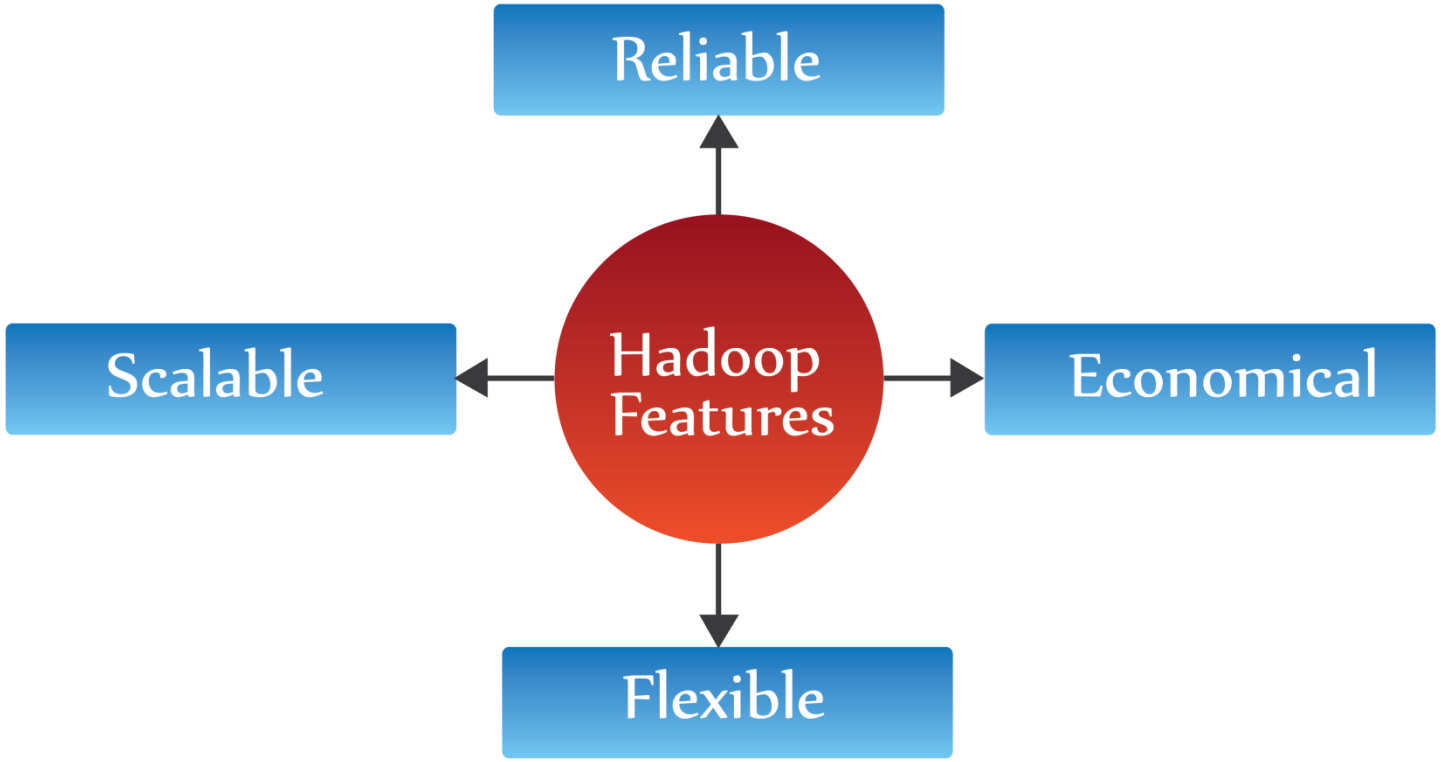
- Apache Hadoop is a framework that allows to store and processing of large data sets across clusters of commodity computers using a simple programming mode.
- It is an Open-source Data Management with scale-out storage and distributed processing



What Hadoop does?

- Hadoop provides a distributed file system and a framework for processing large data sets.
- Hadoop is designed to run on large number of machine that doesn't share any memory or disk.
- HDFS(Hadoop component) divides files into many small blocks.
- HDFS creates multiple replicas of data blocks for reliability , placing them on compute nodes around the cluster.
- MapReduce then process the data in it's location.
- Hadoop's target is to run on cluster of the order of 10,000-nodes.

Key Characteristics of Hadoop



Hadoop in the Industry

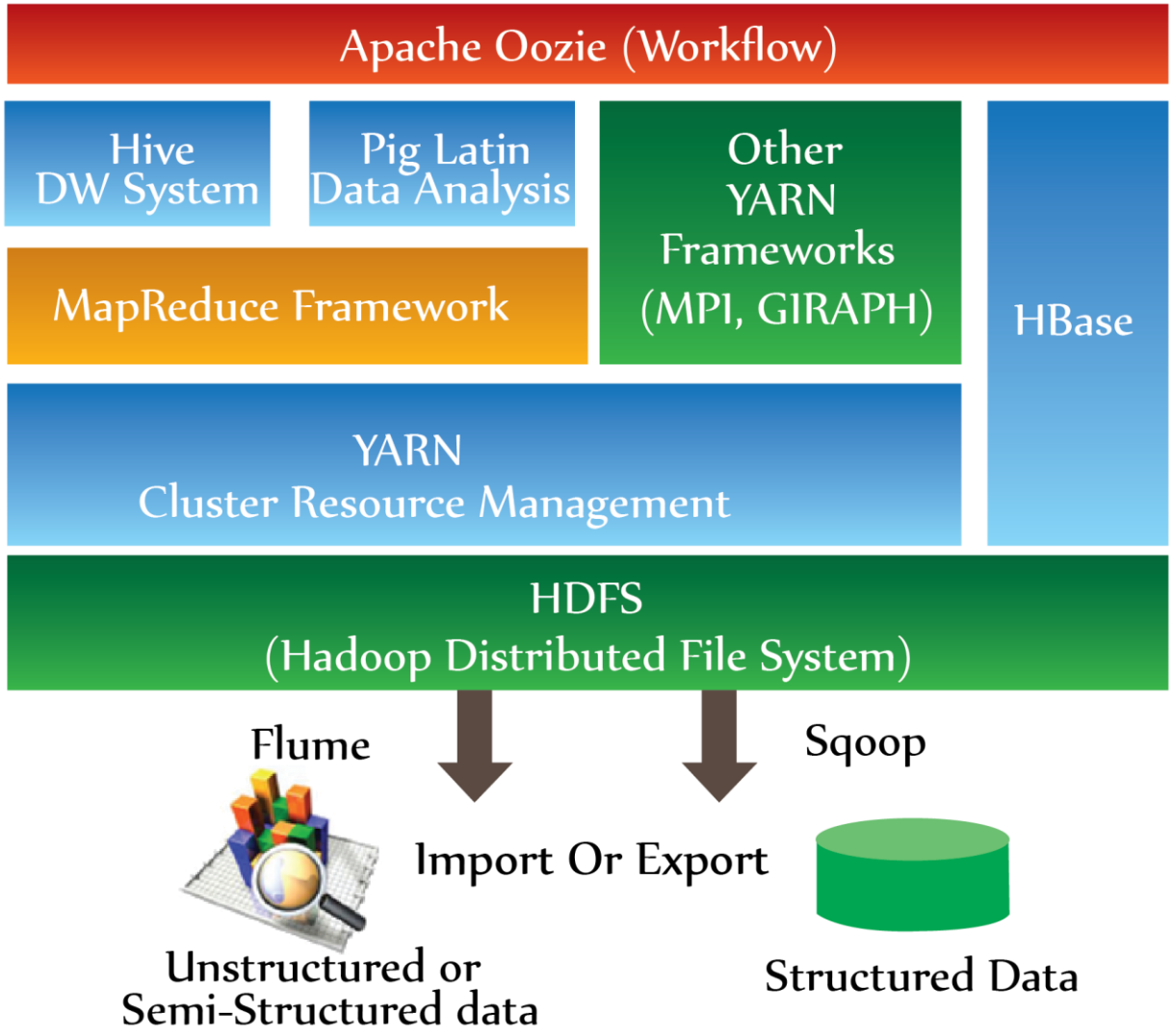
Nokia

Nokia collects and analyzes vast amounts of data from mobile phones

Problem: Dealing with 100TB of structured data and 500TB+ of semi-structured data. How to save it a cost efficient manner.

Solution: HDFS data warehouse allows storing all the semi/multi structured data and offers processing data at peta byte scale

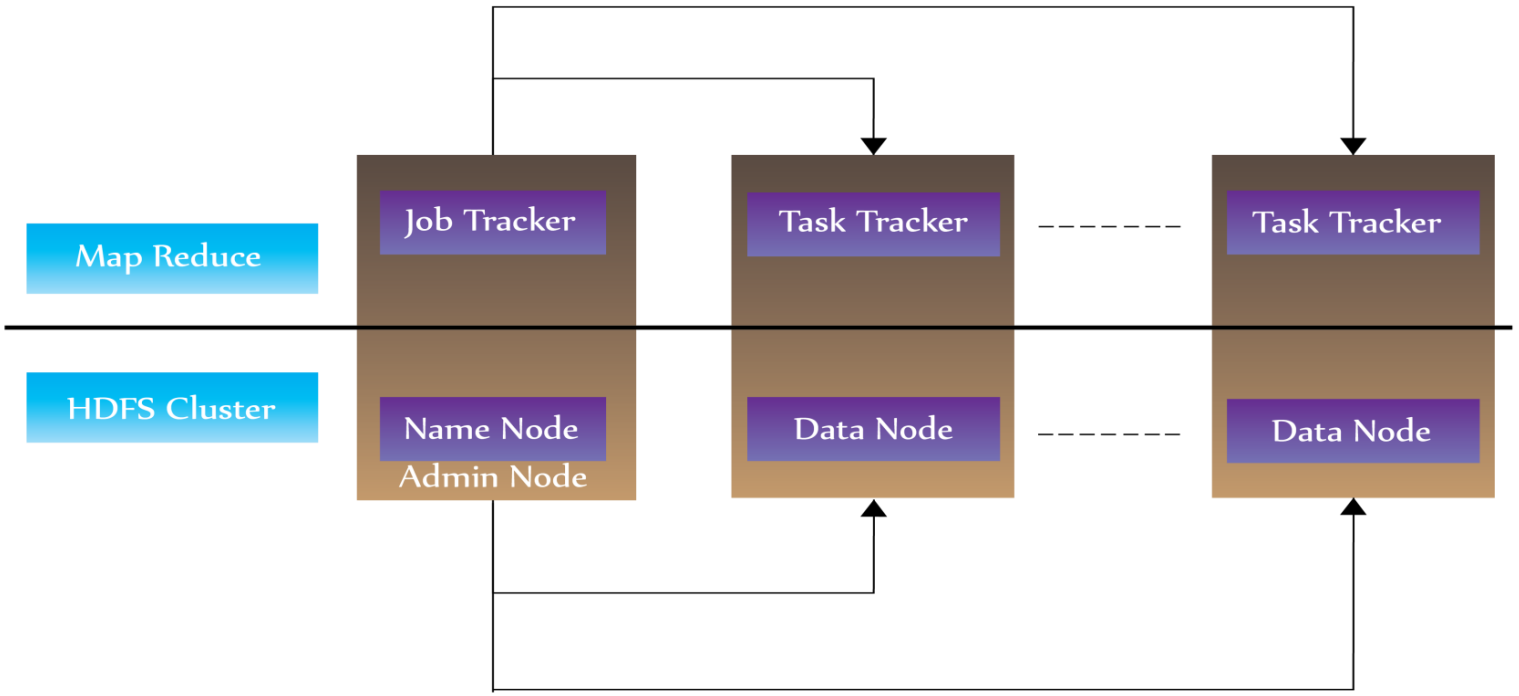
Hadoop Ecosystem



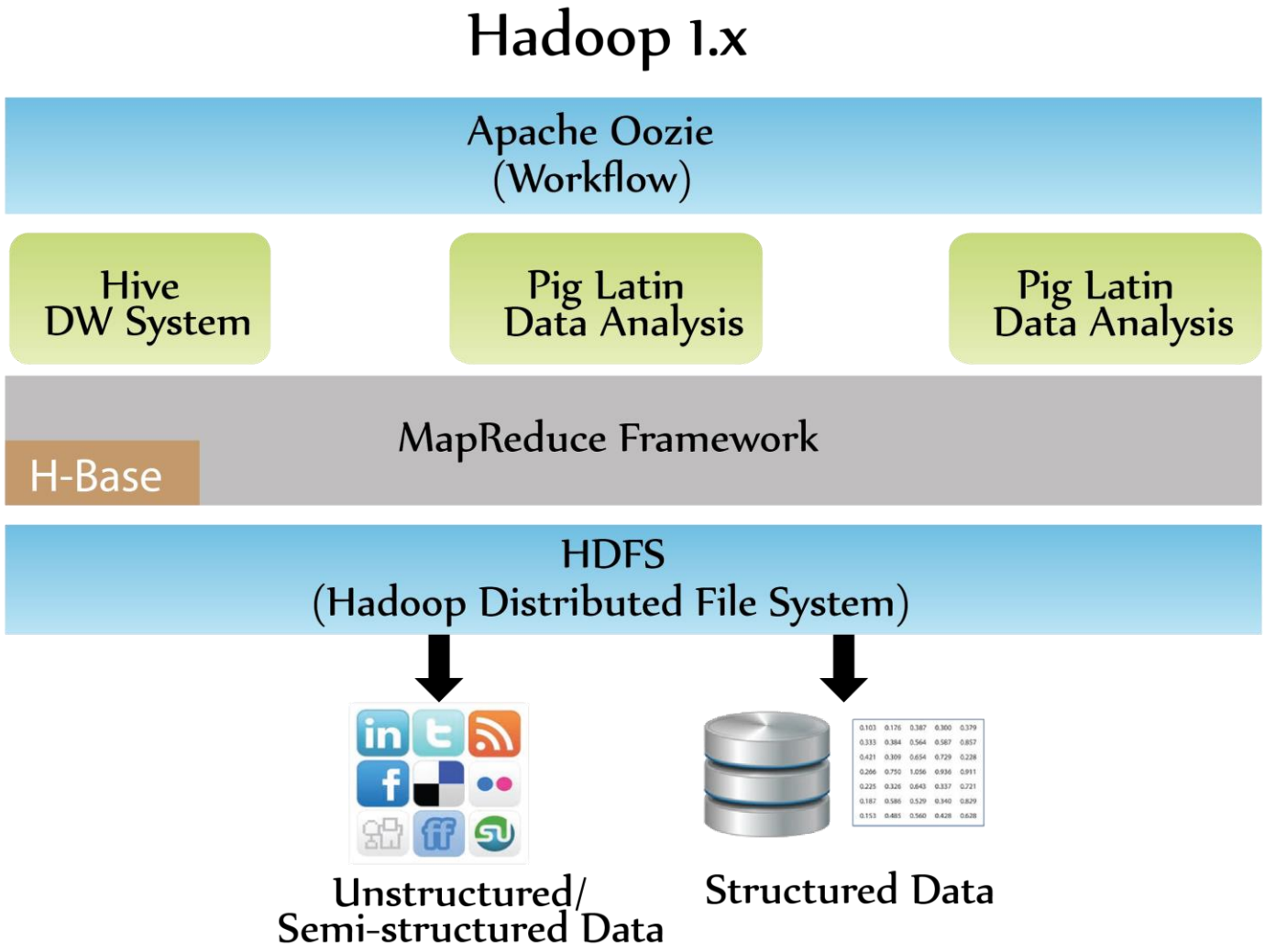
Hadoop 1.0 Core Components

Hadoop has two main components:

- **HDFS** : - The storage layer (Hadoop Distributed File System).
- **MapReduce** : - The programming model or brain of Hadoop .



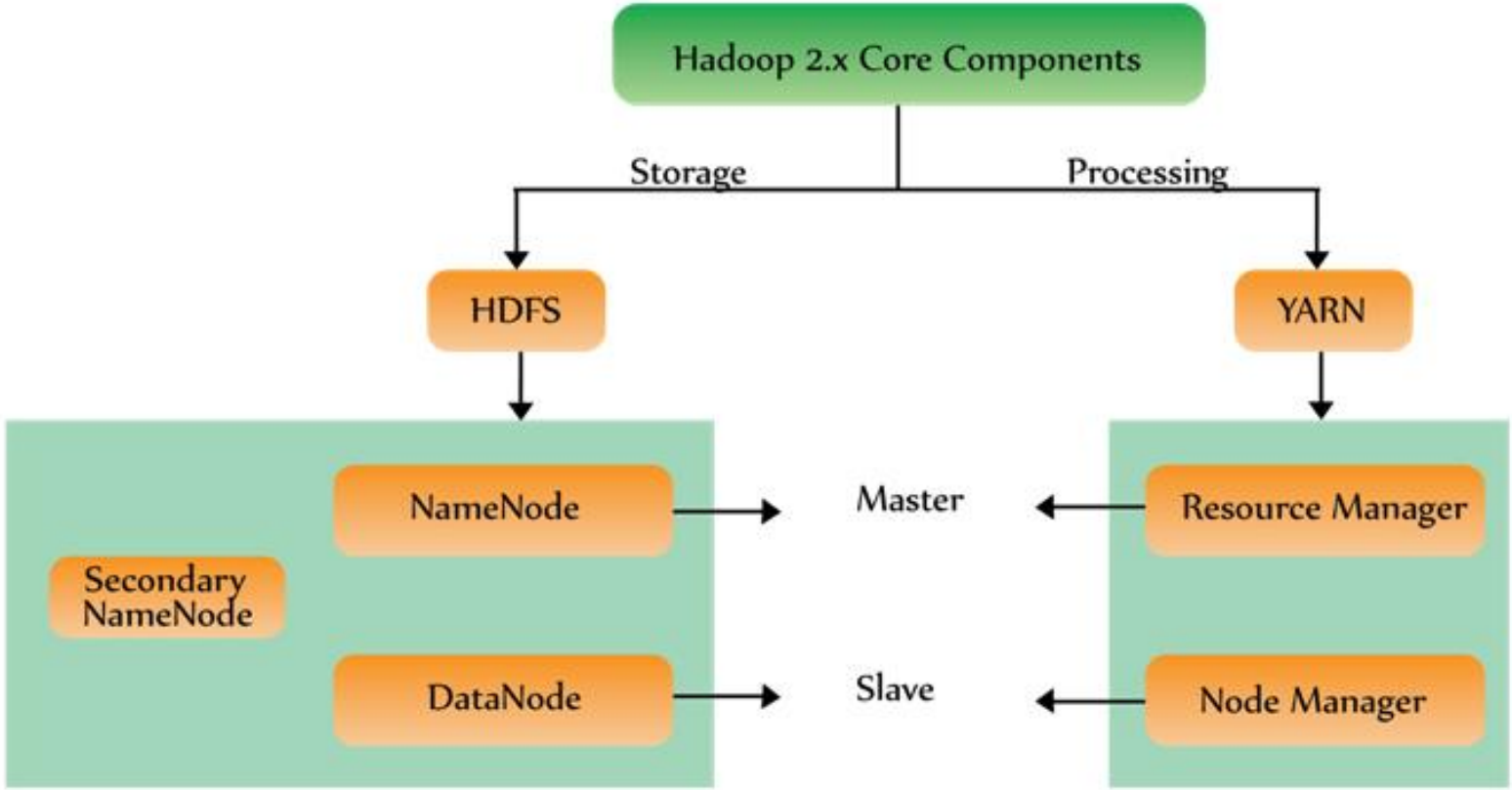
Hadoop Ecosystem for 1.x



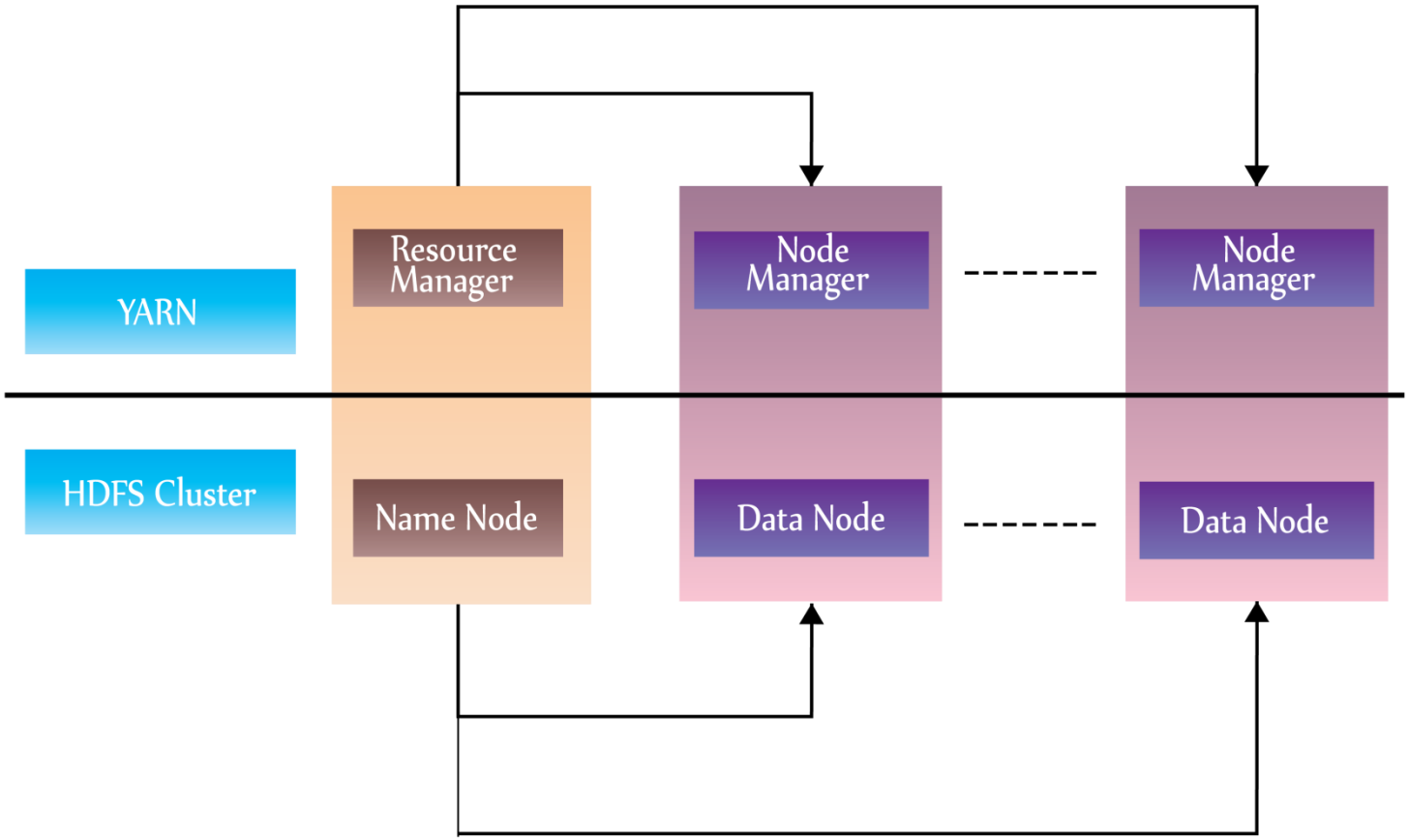
Hadoop 2.x Core Components

- Hadoop infrastructure is divided into 2 parts -storage and processing.
- Hadoop Distributed File System (HDFS) provides the storage mechanism. Yet Another Resource Negotiator (YARN) provides the processing part.
- In hadoop total 5 Daemons -3 for HDFS and 2 for YARN. They work in a master and slave mode.
- Name Node, Secondary Name Node (Masters) and Data Nodes (Slaves) for HDFS and Resource Manager (Master) and Node Managers (Slaves) for YARN

Hadoop 2.x Core Components(Cont'd)

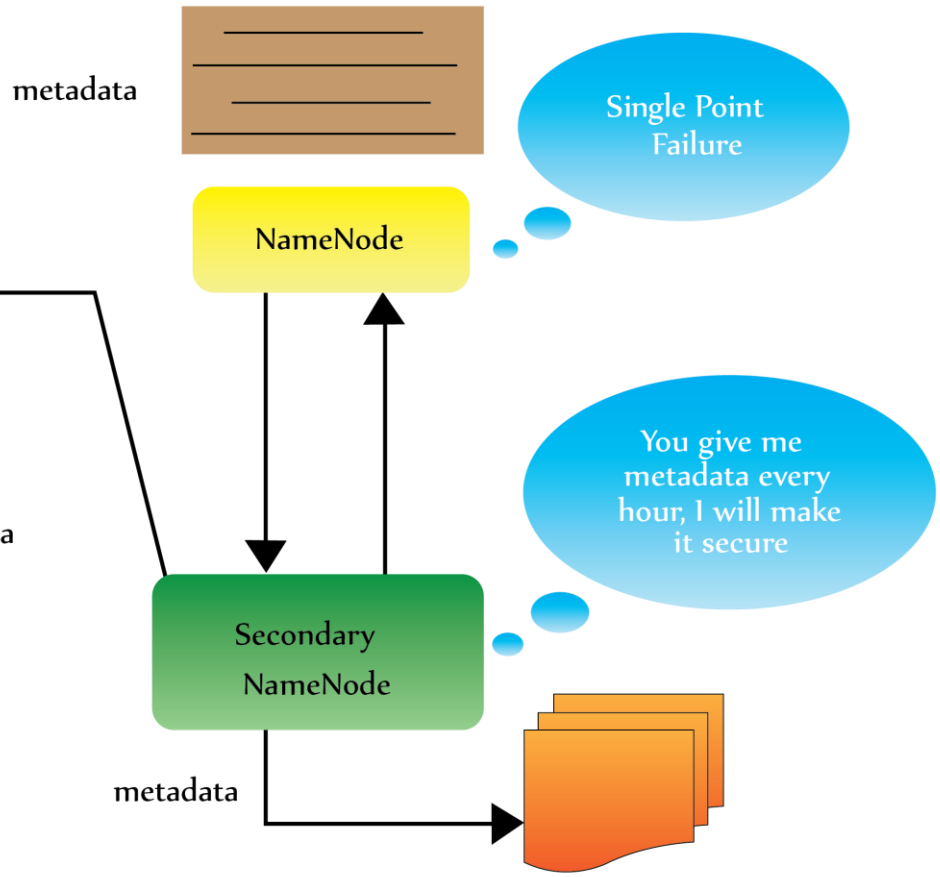


Hadoop 2.x Core Components (Cont'd)



Secondary Name Node

- Secondary NameNode:
  - Not a hot standby for the NameNode
  - Connects to nameNode every hour
  - Housekeeping, backup of NameNode metadata
  - Saved metadata can build a failed nameNode



Quiz - 1

How does Hadoop process large volumes of data?

A - Hadoop uses a lot of machines in parallel. This optimizes data processing.

B - Hadoop was specifically designed to process large amount of data by taking advantage of MPP hardware.

C - Hadoop ships the code to the data instead of sending the data to the code.

D - Hadoop uses sophisticated caching techniques on name node to speed processing of data.

Quiz - Solution

How does Hadoop process large volumes of data?

A - Hadoop uses a lot of machines in parallel. This optimizes data processing.

B - Hadoop was specifically designed to process large amount of data by taking advantage of MPP hardware.

✓ **C** - Hadoop ships the code to the data instead of sending the data to the code.

D - Hadoop uses sophisticated caching techniques on name node to speed processing of data.

Quiz - 2

Which one of the following stores data?

A - Name node

B - Data node

C - Master node

D - None of these

Quiz - Solution

Which one of the following stores data?

A - Name node

✓ **B** - Data node

C - Master node

D - None of these

Quiz - 3

The main role of the secondary namenode is to

A - Copy the filesystem metadata from primary namenode.

B - Copy the filesystem metadata from NFS stored by primary namenode

C - Monitor if the primary namenode is up and running.

D - Periodically merge the namespace image with the edit log.

Quiz - Solution

The main role of the secondary namenode is to

A - Copy the filesystem metadata from primary namenode.

B - Copy the filesystem metadata from NFS stored by primary namenode

C - Monitor if the primary namenode is up and running.

✓ **D** - Periodically merge the namespace image with the edit log.

Quiz - 4

What is are true about HDFS?

A - HDFS filesystem can be mounted on a local client's Filesystem using NFS.

B - HDFS filesystem can never be mounted on a local client's Filesystem.

C - You can edit a existing record in HDFS file which is already mounted using NFS.

D - You cannot append to a HDFS file which is mounted using NFS.

Quiz - Solution

What is are true about HDFS?

✓ **A** - HDFS filesystem can be mounted on a local client's Filesystem using NFS.

B - HDFS filesystem can never be mounted on a local client's Filesystem.

C - You can edit a existing record in HDFS file which is already mounted using NFS.

D - You cannot append to a HDFS file which is mounted using NFS.

Quiz - 5

In the secondary namenode the amount of memory needed is

A - Similar to that of primary node

B - Should be at least half of the primary node

C - Must be double of that of primary node

D - Depends only on the number of data nodes it is going to handle

Quiz - Solution

In the secondary namenode the amount of memory needed is

✓ **A** - Similar to that of primary node

B - Should be at least half of the primary node

C - Must be double of that of primary node

D - Depends only on the number of data nodes it is going to handle

Any Doubts?

Summary

- Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.
- Big data can be characterized by 3Vs - the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed.
- Batch data processing is an efficient way of processing high volumes of data is where a group of transactions is collected over a period of time
- Real time data processing involves a continual input,process and output of data
- Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment.
- Hadoop has two main components:

  1 HDFS for storage and

  2 Map Reduce for processing

Kratoes
for the quick learner

Thank You!