Sohail Akhtar
CS Department
Bahria University, Islamabad Campus

Provided by:
Dr Muhammad Rashid Hussain

Agenda

- What is Hadoop
- Brief History on Hadoop
- Hadoop EcoSystem
- Hadoop Vendors
- Intro of each Hadoop Ecosystem

Hadoop ???

History of hadoop

- Hadoop was created by Doug Cutting who had created the Apache Lucene (Text Search),which is origin in Apache Nutch (Open source search Engine).Hadoop is a part of Apache Lucene Project.Actually Apache Nutch was started in 2002 for working crawler and search
- In January 2008, Hadoop was made its own top-level project at Apache for, confirming success ,By this time, Hadoop was being used by many other companies such as Yahoo!, Facebook, etc.
- In April 2008, Hadoop broke a world record to become the fastest system to sort a terabyte of data.
- Yahoo take test in which To process 1TB of data (1024 columns)

  oracle – 3 ½ day

  teradata – 4 ½ day

  netezza – 2 hour 50 min

  hadoop - 3.4 min

Hadoop ???

WHAT IS HADOOP

- Hadoop is the product of Apach ,it is the type of distributed system, it is framework for big data
- Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.
- Some of the characteristics:
  - Open source
  - Distributed processing
  - Distributed storage
  - Reliable
  - Economical
  - Flexible

Hadoop ???

Hadoop Framework Modules

The base Apache Hadoop framework is composed of the following modules:

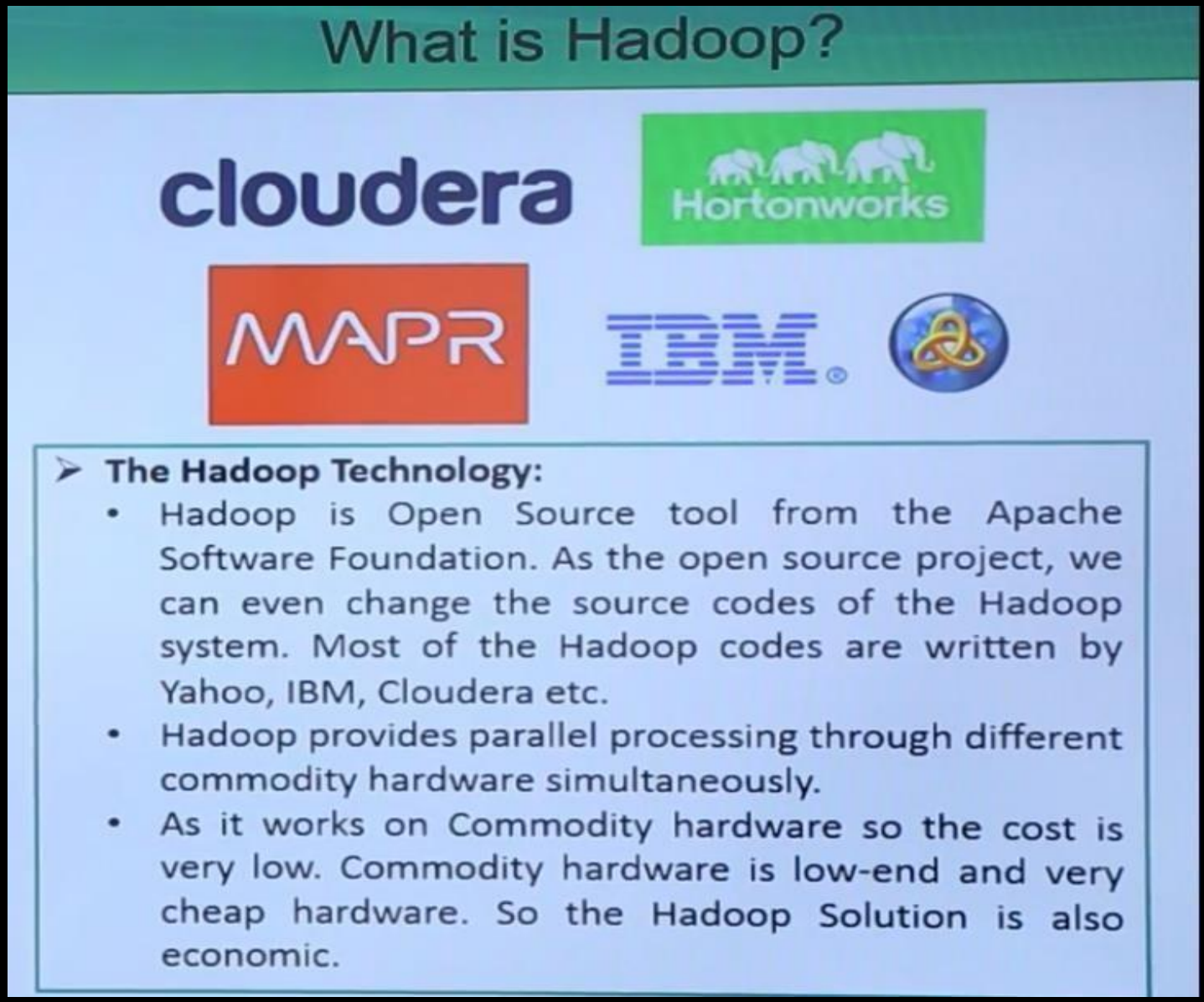
- **Hadoop Common :**– contains libraries and utilities needed by other Hadoop modules
- **Hadoop Distributed File System (HDFS) :**– a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster
- **Hadoop YARN:**– a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications
- **Hadoop MapReduce:**– an implementation of the MapReduce programming model for large scale data processing.

Hadoop ???

What is Hadoop?

- **What is the Apache Hadoop?**
  - The Apache Hadoop is an open source framework. Hadoop can easily handle a large amount of data on a low cost, simple hardware cluster. Hadoop is also a scalable and Fault-Tolerant framework.
  - The Hadoop is not only a storage system. Data can be processed using this framework.
  - The Hadoop system is basically written in Java.

Hadoop - Technology



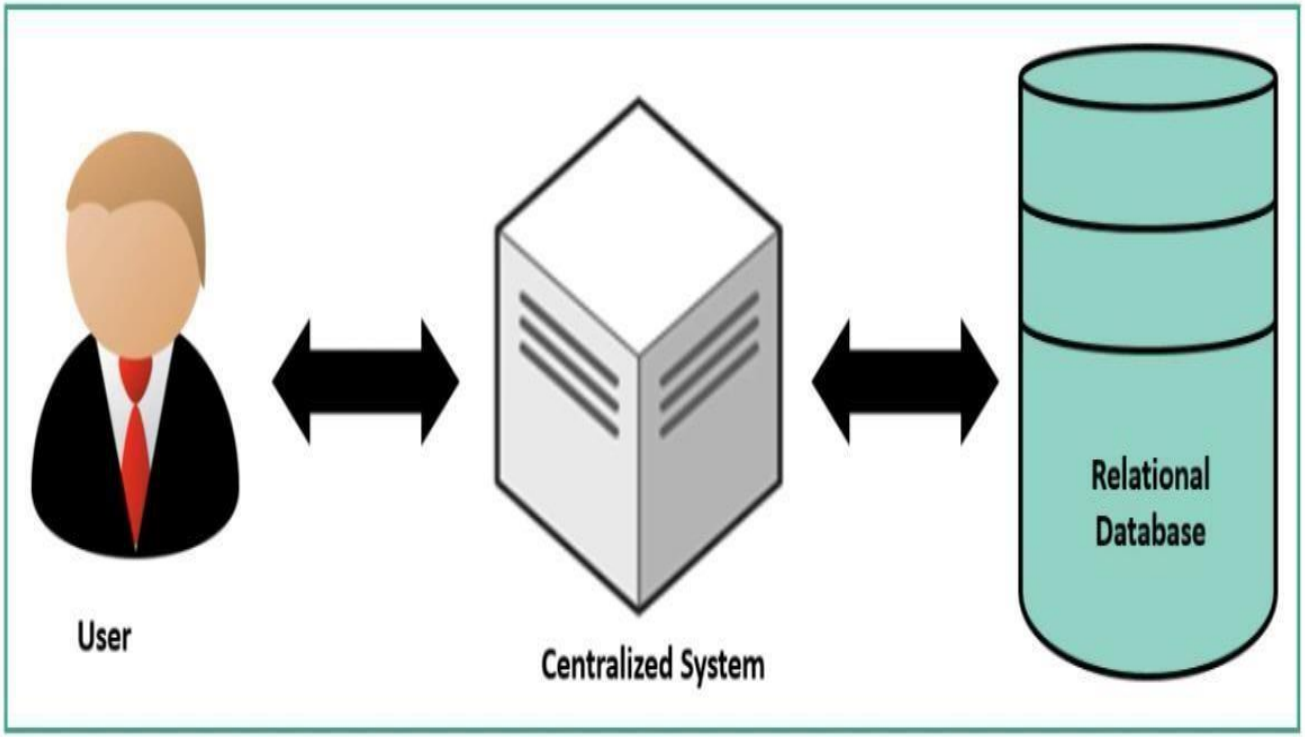
Need of Hadoop

What is Hadoop?

- **Why we should use Hadoop?**
  - The Hadoop solution is very popular. It has captured at least 90% of Big data market.
  - Hadoop has some unique features that make this solution very popular.
  - Hadoop is Scalable. So we can increase the number of commodity hardware easily.
  - It is a fault tolerant solution. When one node goes down other nodes can process the data.
  - Data can be stored as a Structured, Unstructured and semi-structured mode. So it to more flexible.

Need - Hadoop

Brief History on Hadoop

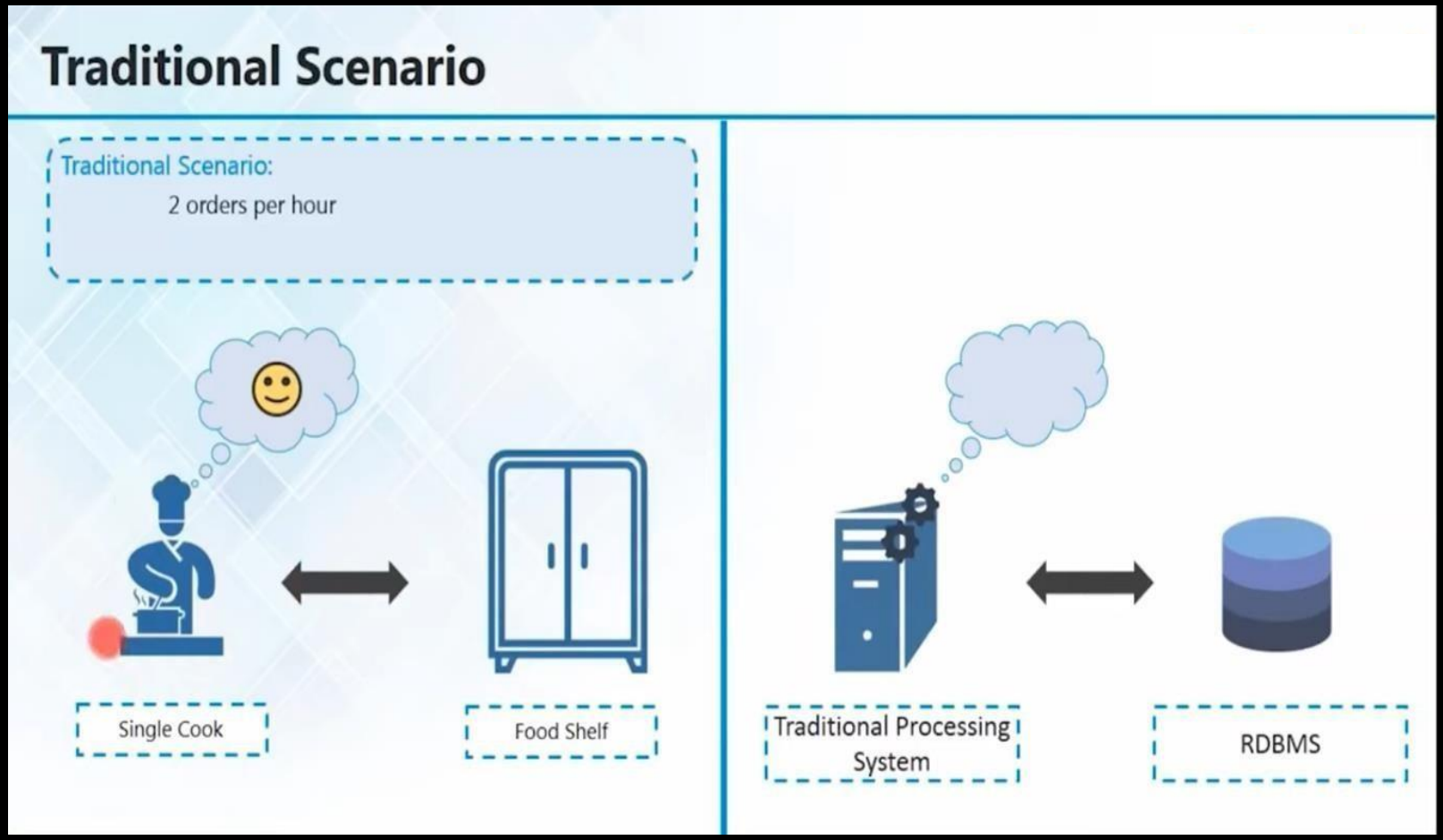


Hadoop - Need

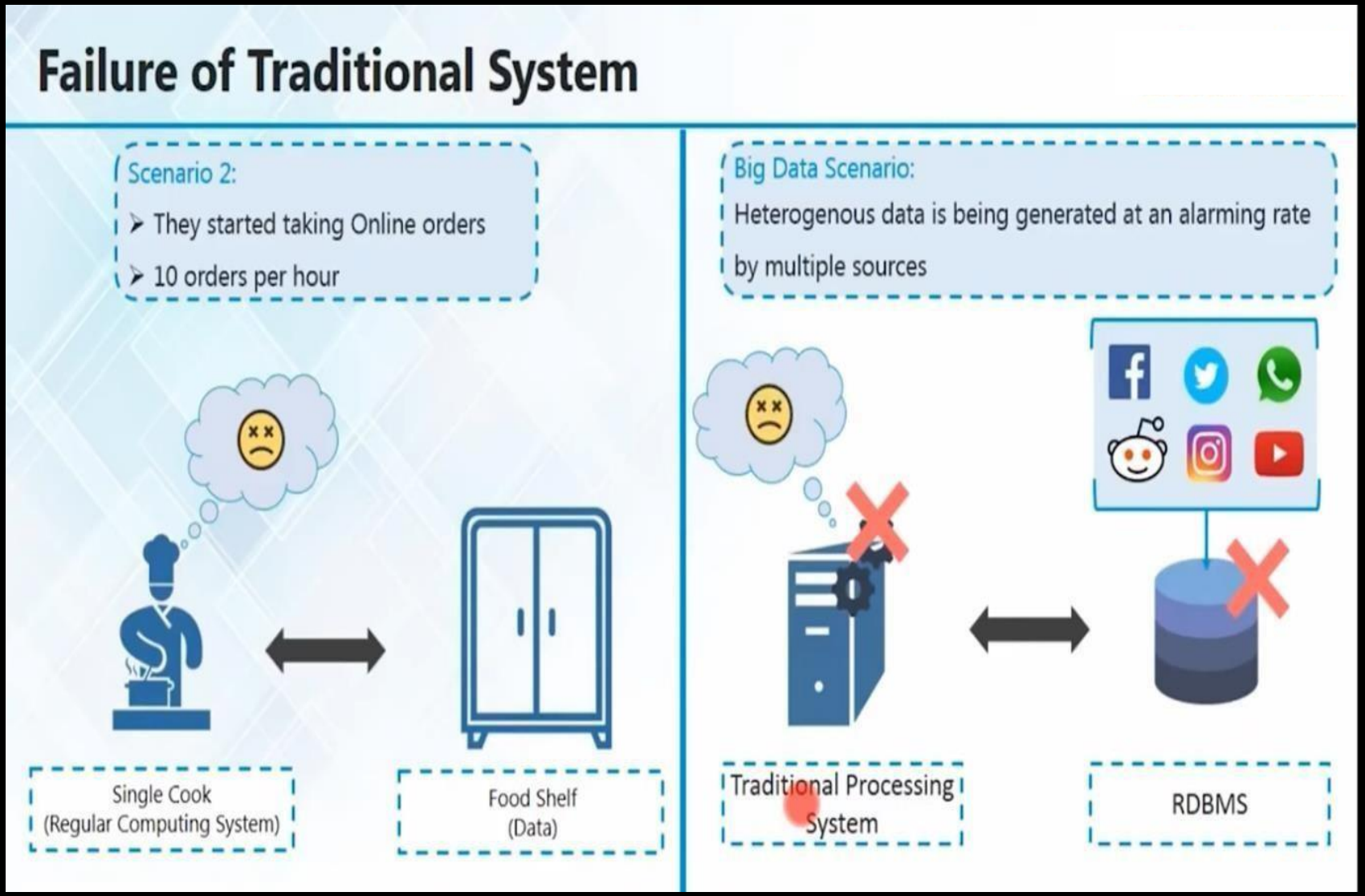
Story of Big Data & Traditional System

Scenario:
Bob has opened a small restaurant in his city

Hadoop - Need

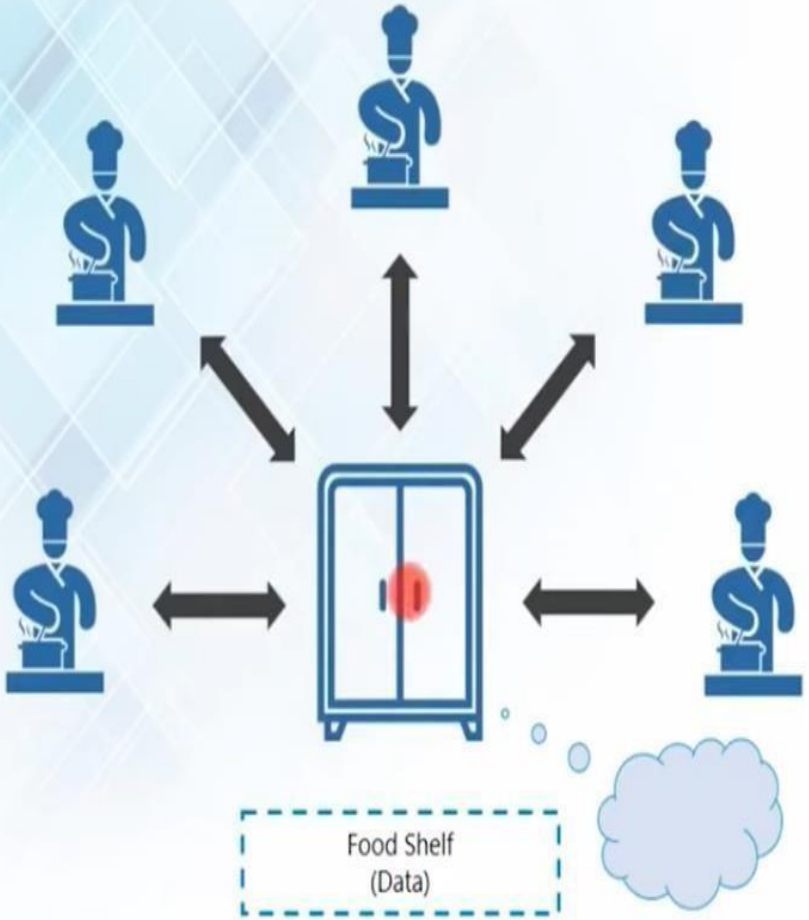

Traditional Scenario
Traditional Scenario:
2 orders per hour
Single Cook
Food Shelf
Traditional Processing System
RDBMS

Hadoop


Failure of Traditional System
Scenario 2:
They started taking Online orders
10 orders per hour
Single Cook
(Regular Computing System)
Food Shelf
(Data)
Big Data Scenario:
Heterogenous data is being generated at an alarming rate by multiple sources
Traditional Processing System
RDBMS

Hadoop

Need of an Effective Solution



Scenario:

Multiple Cook cooking food

Issue:

Food Shelf becomes the BOTTLENECK

Hadoop

Need of an Effective Solution

Scenario:

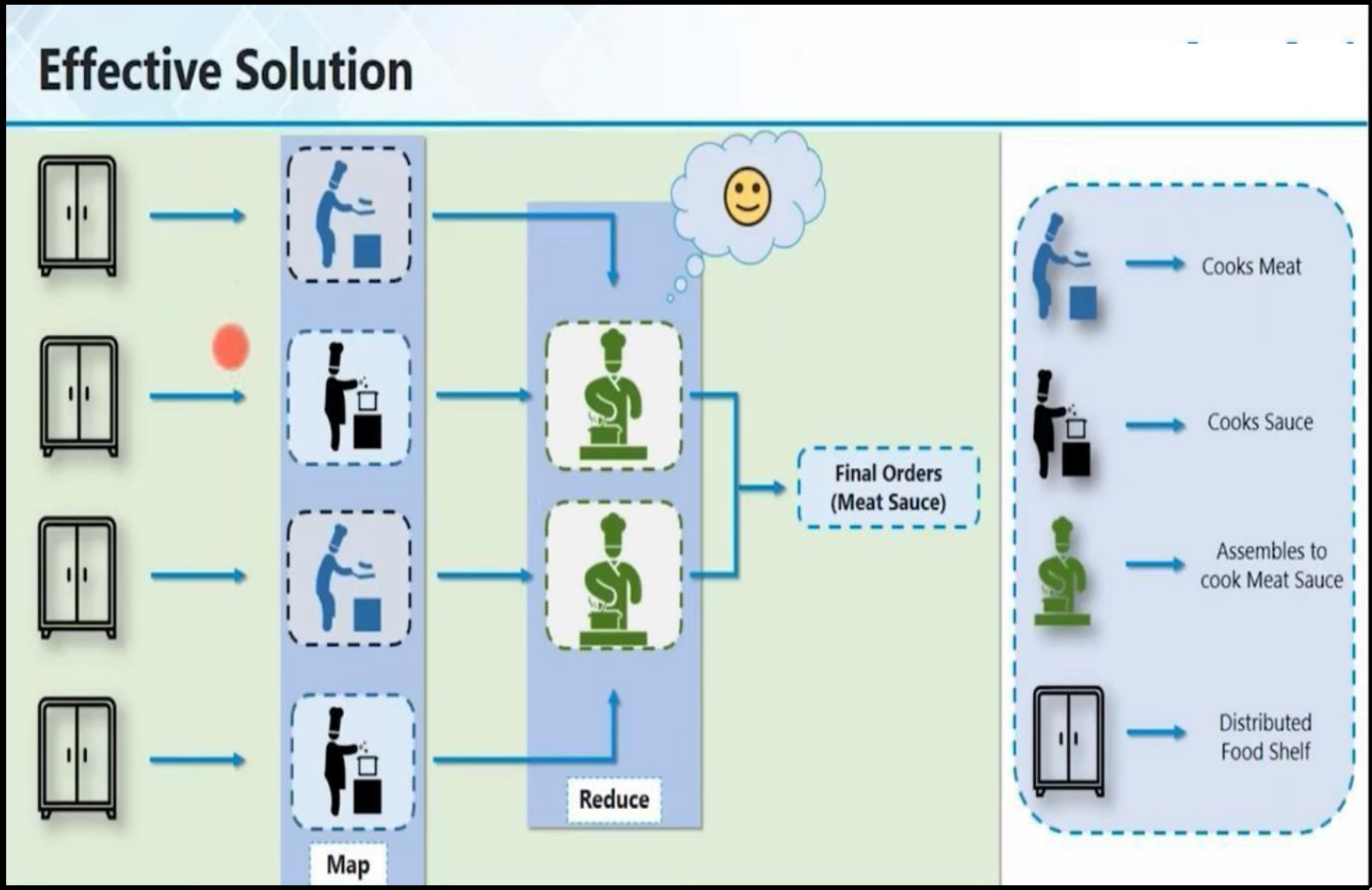
Multiple Processing Unit for data processing

Issue:

Bringing data to processing generated lots of Network overhead

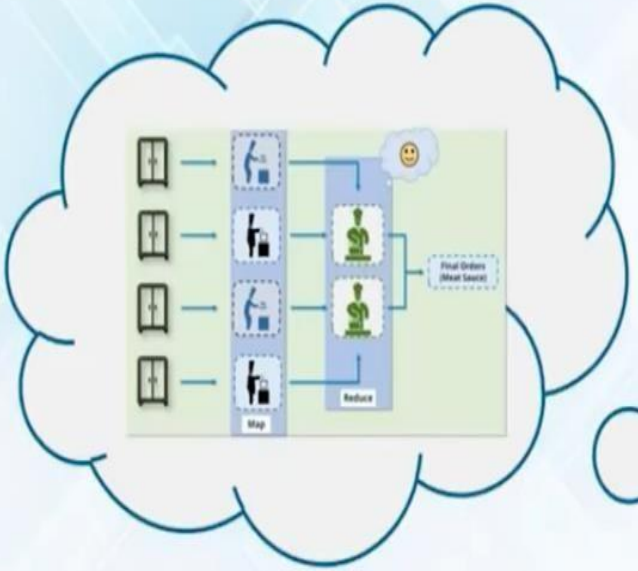
Data Warehouse

Hadoop


Effective Solution
Map
Reduce
Final Orders
(Meat Sauce)
Cooks Meat
Cooks Sauce
Assembles to
cook Meat Sauce
Distributed
Food Shelf

Hadoop

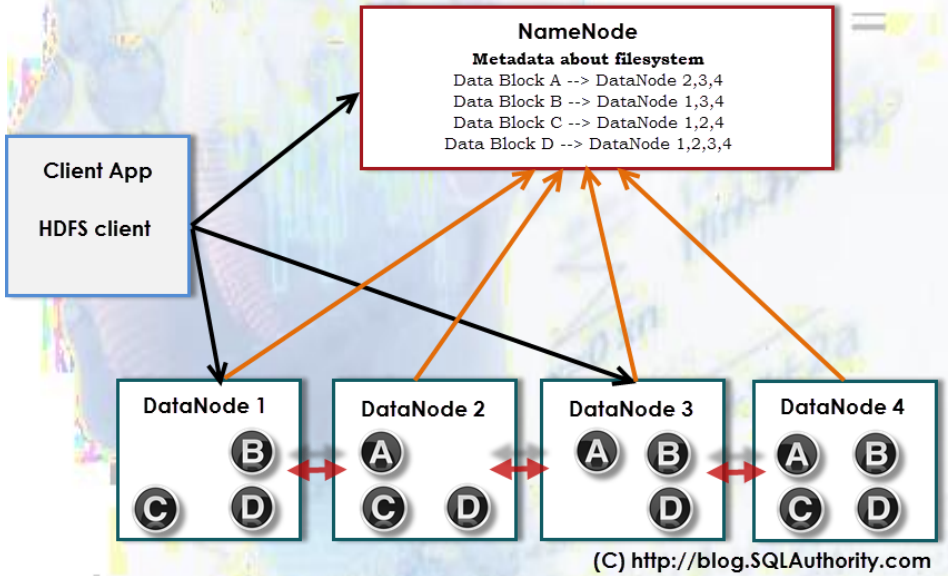
Need of a Framework



Do we have a framework that works like that ?

Hadoop Distributed File System

- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit.
- HDFS has two core components, i.e. NameNode and DataNode.
- The NameNode is the main node that contains metadata about the data stored.
- Data is stored on the DataNodes which are commodity hardware in the distributed environment.



~~Problems with Big Data~~ Apache Hadoop

Problem 1: Storing exponentially growing datasets

- **Hadoop Solution:** HDFS
  - Distributed File System
  - Divide input (data) into chunks and store across the cluster (with replicas)
  - Scale as required

~~Problems with Big Data~~ Apache Hadoop

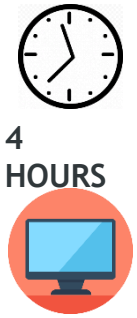
Problem 2: Storing different types of Data

- **Hadoop Solution: HDFS**
  - Allows to store any kind of data:
    - Structured
    - Semi-Structured
    - Unstructured
  - Write Once Read Many (WORM)
  - No schema validation is done while dumping data

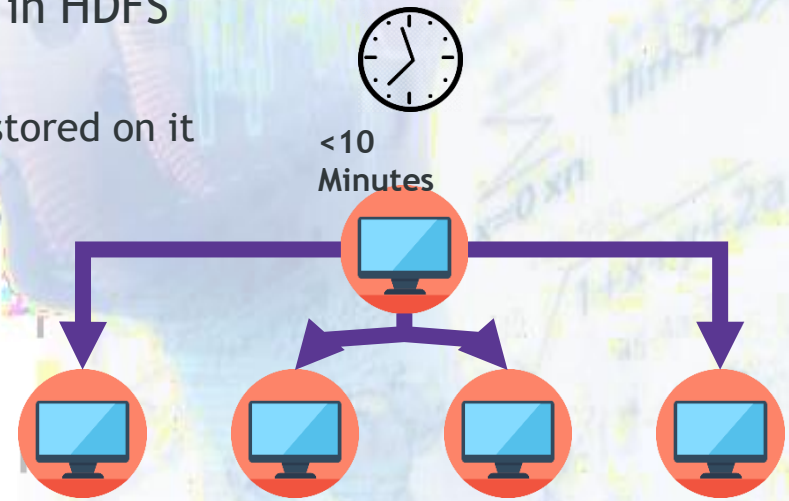
~~Problems with Big Data~~ Apache Hadoop

Problem 3: Process Data Faster

- **Hadoop Solution: MapReduce**
  - Parallel processing of data present in HDFS
  - Process data locally;
    - Each node works on part of the data stored on it

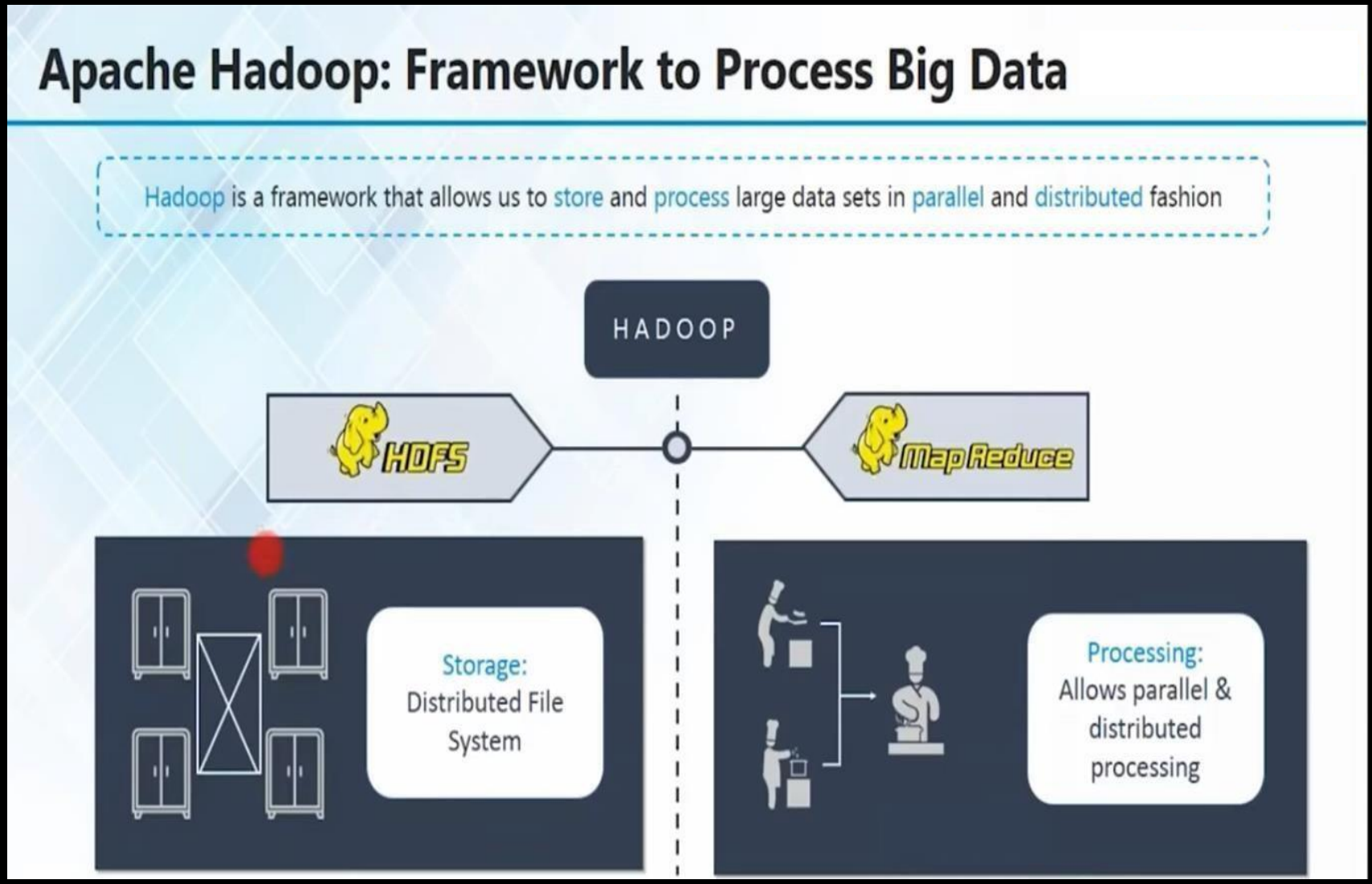


Single Node (commodity hardware)

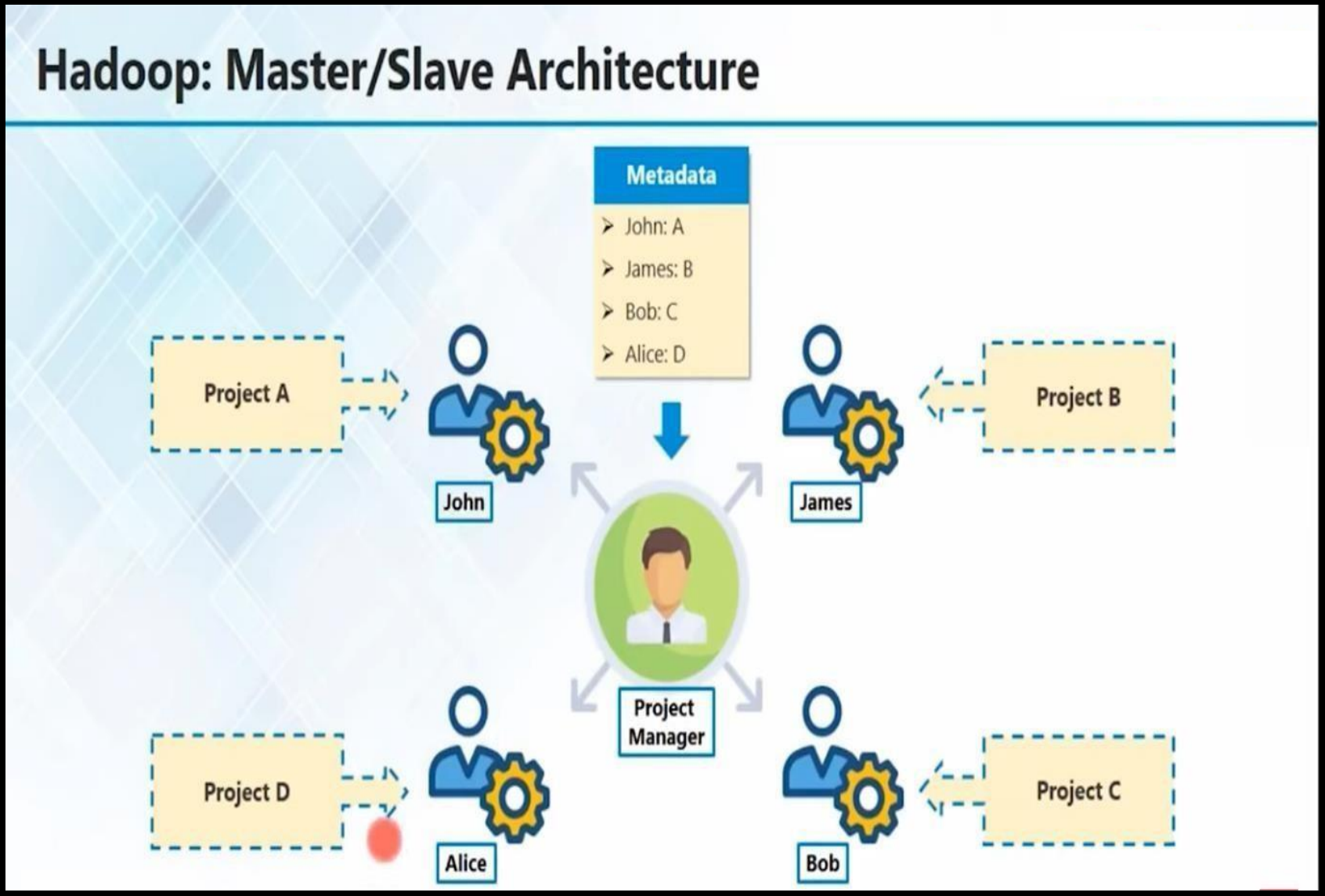


Cluster of Nodes (commodity hardware)

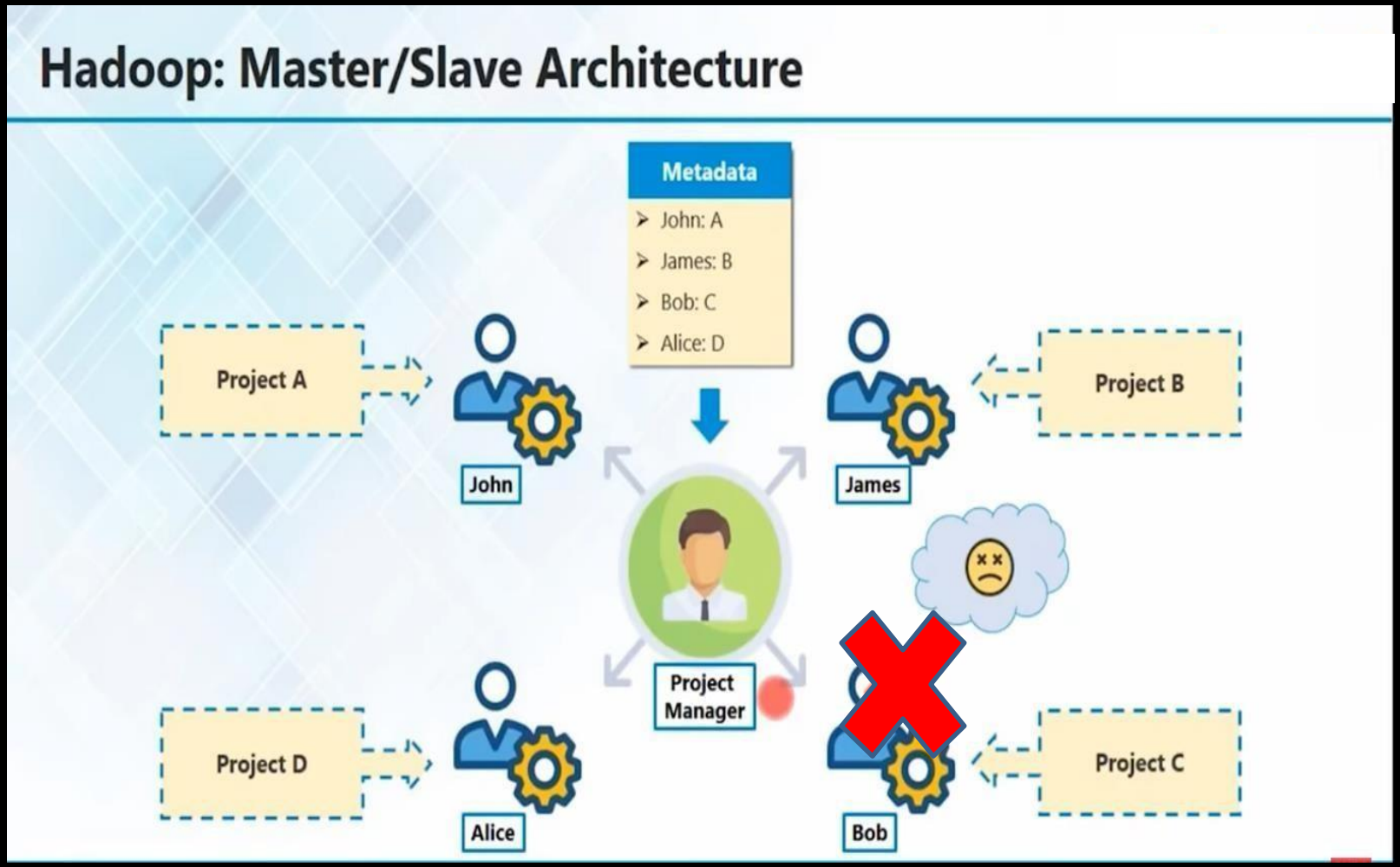
Hadoop - Framework



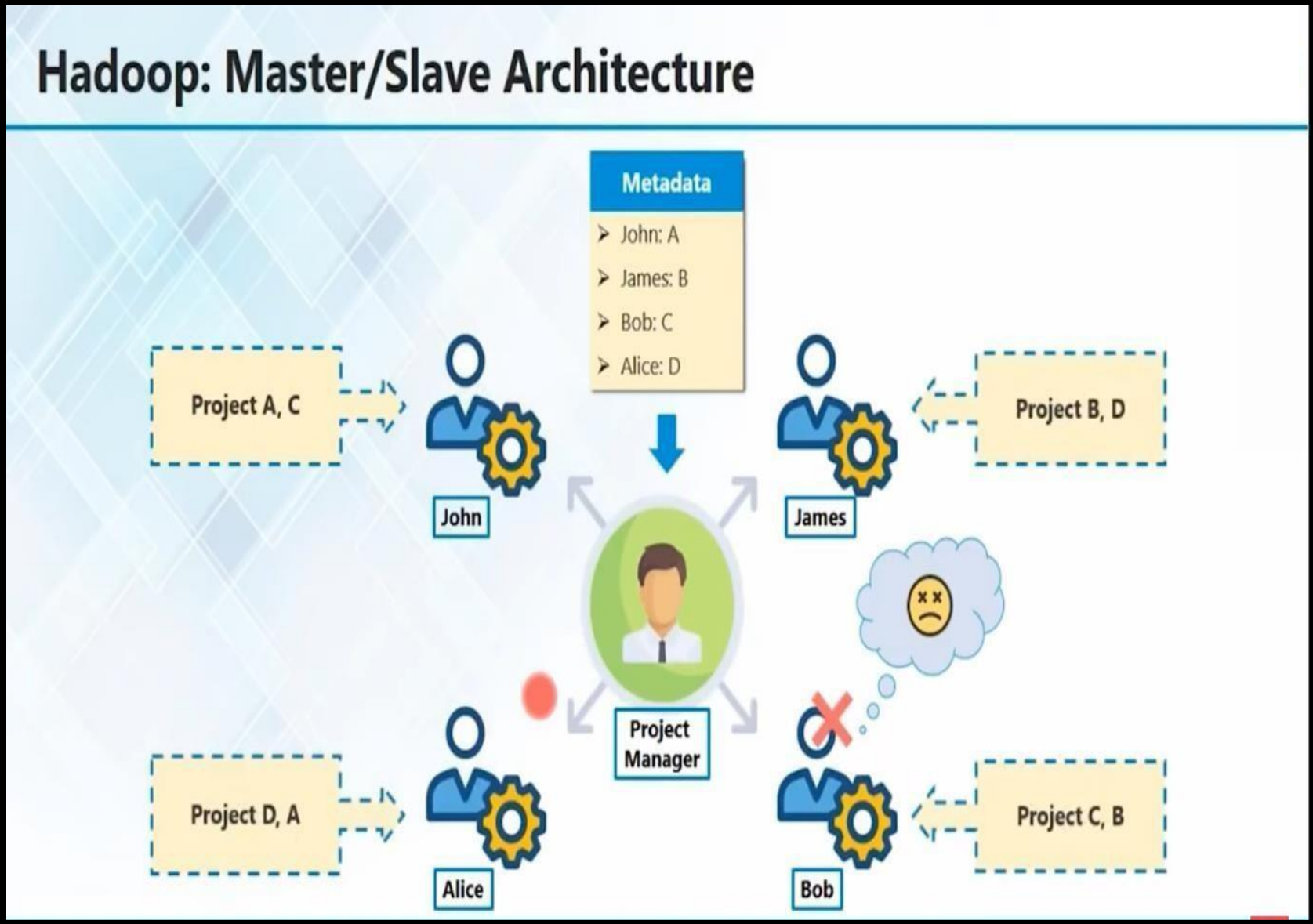
Hadoop - Architecture



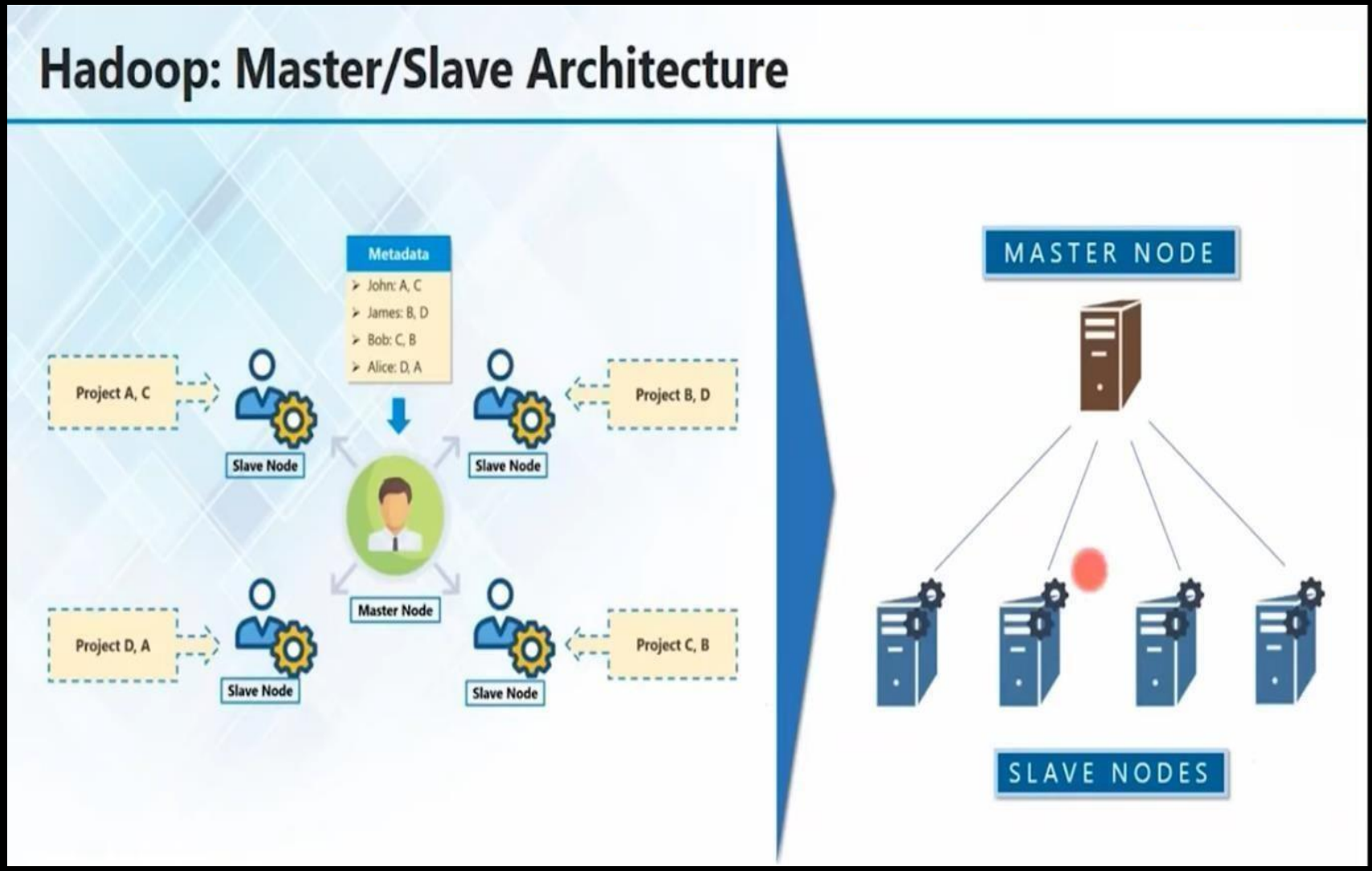
Hadoop


Hadoop: Master/Slave Architecture
Metadata
John: A
James: B
Bob: C
Alice: D
Project A
Project B
Project C
Project D
John
James
Alice
Bob
Project Manager

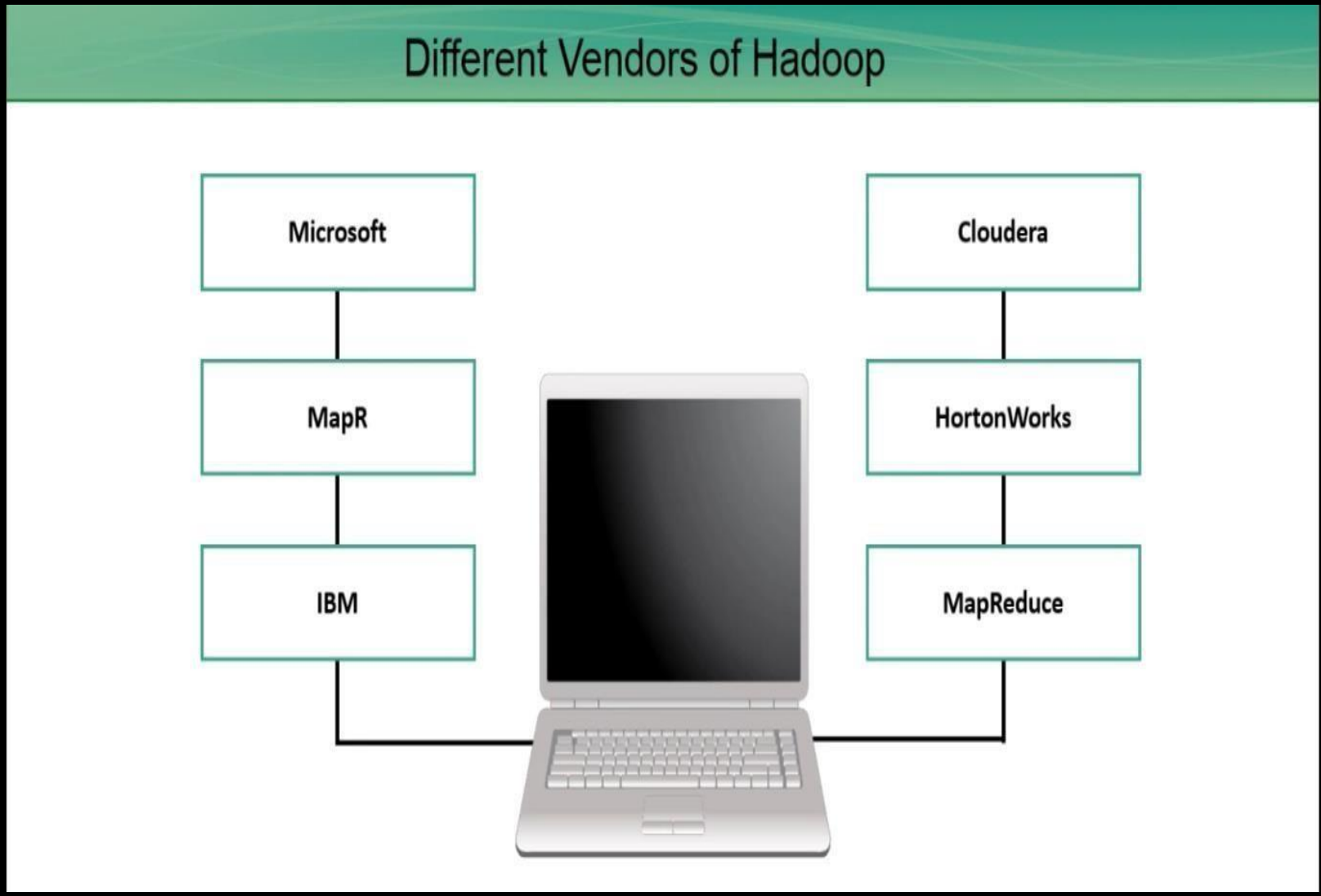
Hadoop


Hadoop: Master/Slave Architecture
Metadata
John: A
James: B
Bob: C
Alice: D
Project A, C
John
James
Project B, D
Project Manager
Project D, A
Alice
Bob
Project C, B

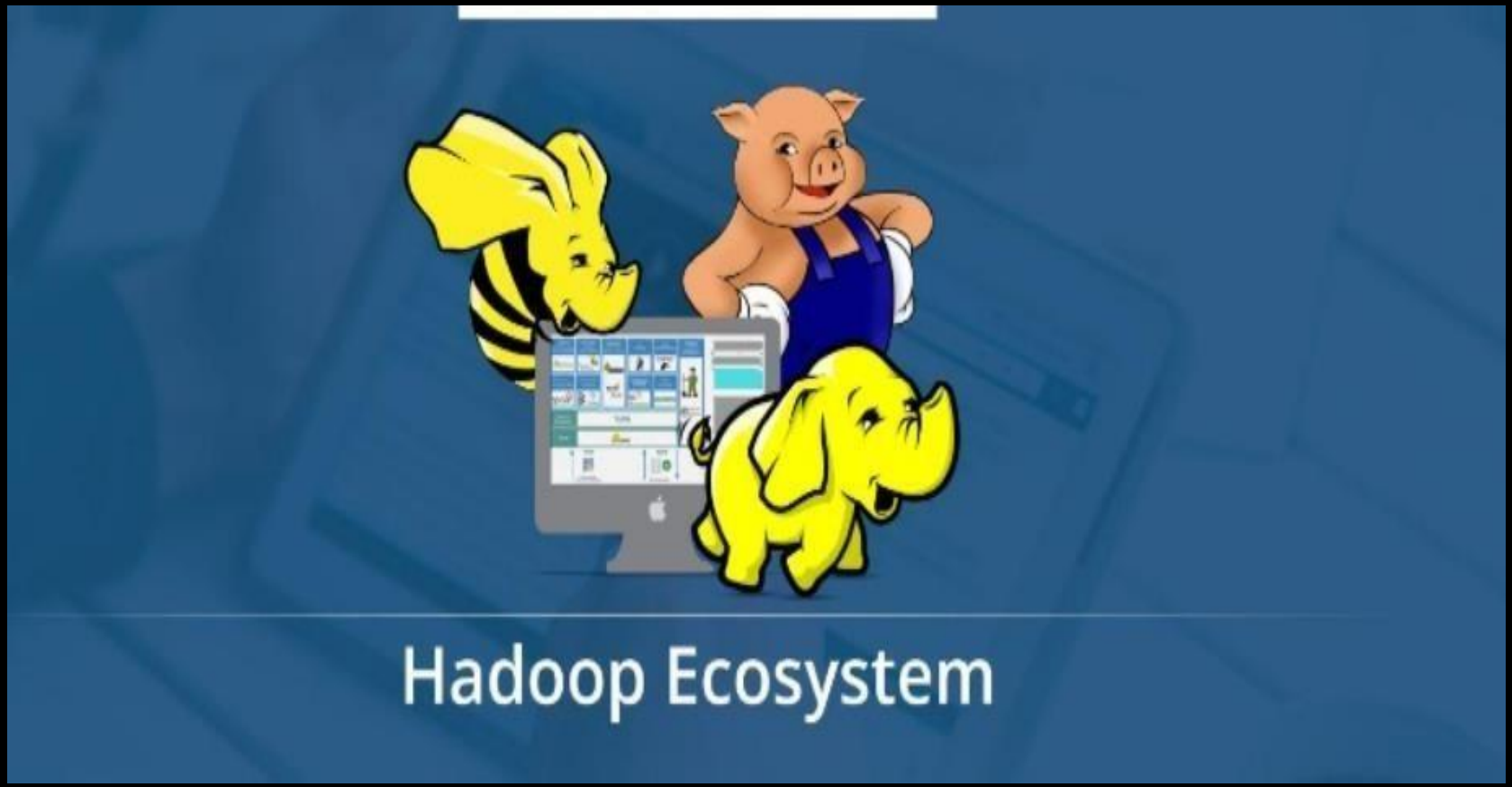
Hadoop


Hadoop: Master/Slave Architecture
Metadata
John: A, C
James: B, D
Bob: C, B
Alice: D, A
Project A, C
Project B, D
Project D, A
Project C, B
Slave Node
Slave Node
Slave Node
Slave Node
Master Node
MASTER NODE
SLAVE NODES

Hadoop


Different Vendors of Hadoop
Microsoft
MapR
IBM
Cloudera
HortonWorks
MapReduce

Hadoop

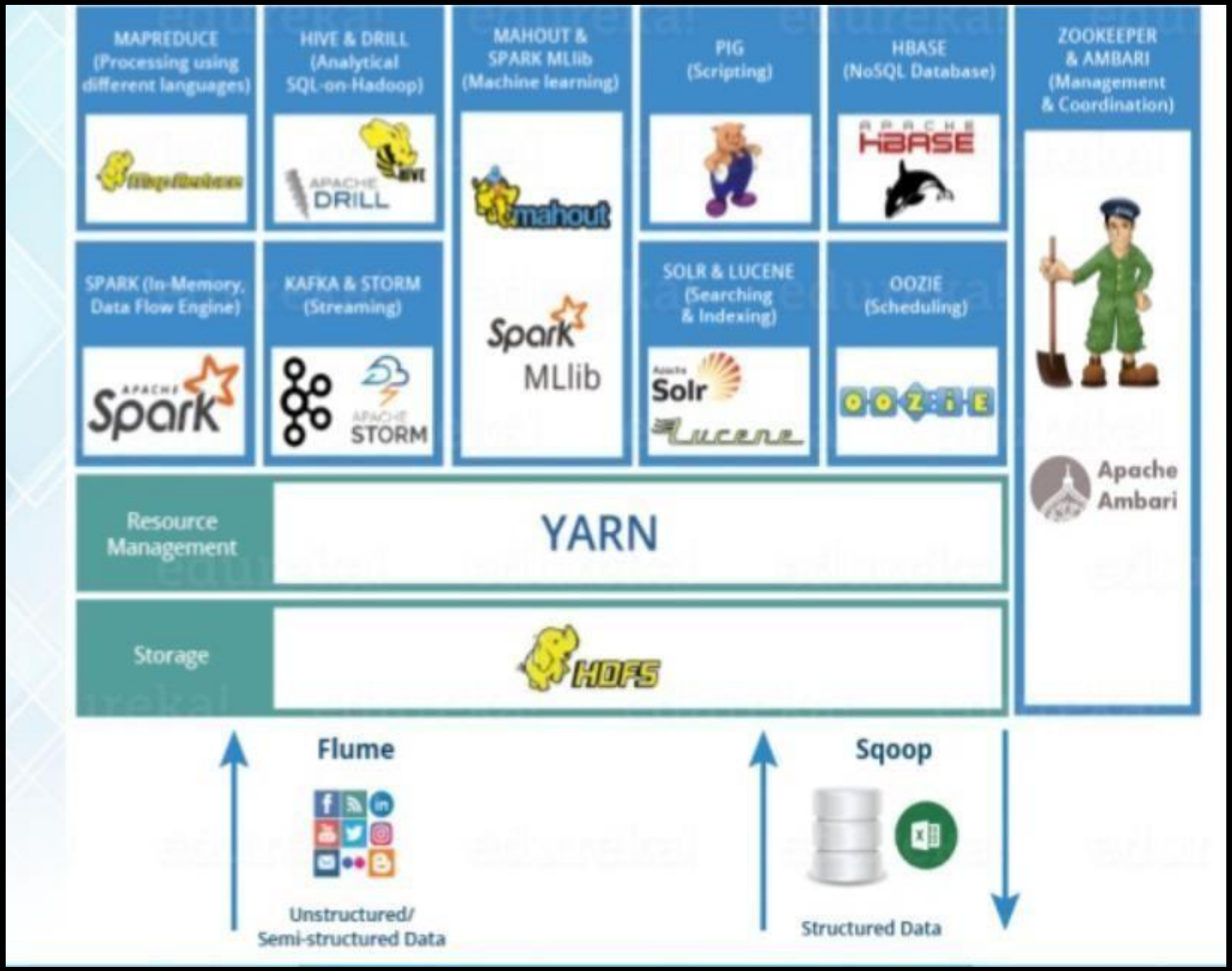


Hadoop Ecosystem

Hadoop Ecosystem

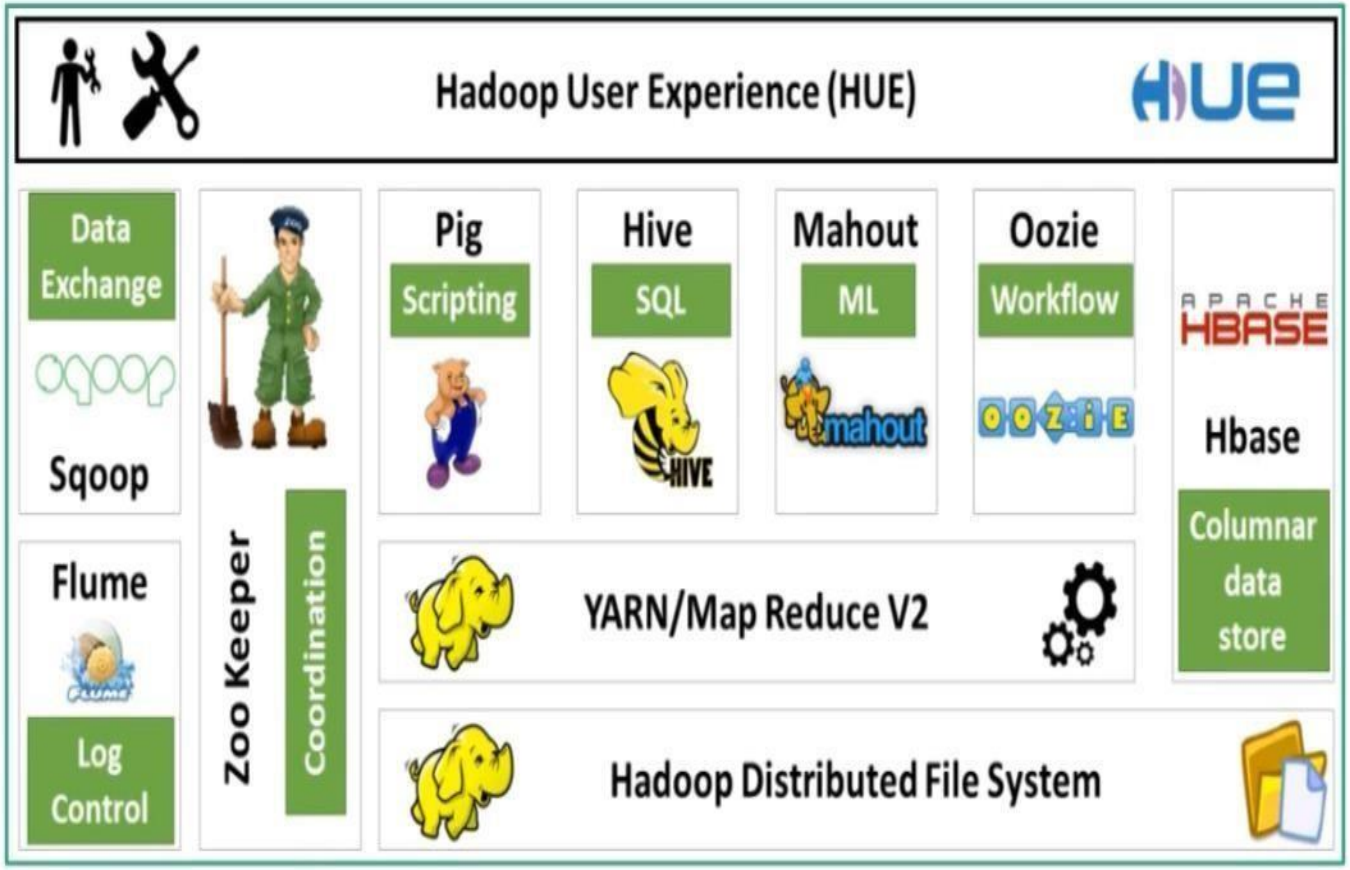
- HDFS -> Hadoop Distributed File System
- YARN -> Yet Another Resource Negotiator
- MapReduce -> Data processing using programming
- Spark -> In-memory Data Processing
- PIG, HIVE-> Data Processing Services using Query (SQL-like)
- HBase -> NoSQL Database
- Mahout, Spark MLlib -> Machine Learning
- Apache Drill -> SQL on Hadoop
- Zookeeper -> Managing Cluster
- Oozie -> Job Scheduling
- Flume, Sqoop -> Data Ingesting Services
- Solr & Lucene -> Searching & Indexing
- Ambari -> Provision, Monitor and Maintain cluster

Hadoop Ecosystem



Hadoop Ecosystem

Different Hadoop Modules



Hadoop Ecosystem - HDFS

- Stores different types of large data sets (i.e. structured, unstructured and semi structured data)
- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit
- Stores data across various nodes and maintains the log file about the stored data (metadata)
- HDFS has two core components, i.e. NameNode and DataNode

Hadoop Ecosystem – YARN

- Performs all your processing activities by allocating resources and scheduling tasks
- Two services: ResourceManager and NodeManager
- ResourceManager: Manages resources and schedule applications running on top of YARN
- NodeManager: Manages containers and monitors resource utilization in each container

Hadoop Ecosystem – Map Reduce

Data Processing using Programming

- Core component in a Hadoop Ecosystem for processing
- Helps in writing applications that processes large data sets using distributed and parallel algorithms
- In a MapReduce program, Map() and Reduce() are two functions
- Map function performs actions like filtering, grouping and sorting
- Reduce function aggregates and summarizes the result produced by map function

Hadoop Ecosystem – PIG

Data Processing using Query

- PIG has two parts: Pig Latin, the language and the pig runtime, for the execution environment
- **1 line of pig latin = approx. 100 lines of Map-Reduce job**
- The compiler internally converts pig latin to MapReduce
- It gives you a platform for building data flow for ETL (Extract, Transform and Load)
- PIG first loads the data, then performs various functions like grouping, filtering, joining, sorting, etc. and finally dumps the data on the screen or stores in HDFS.

Hadoop Ecosystem – HIVE

- A data warehousing component which analyses data sets in a distributed environment using SQL-like interface
- The query language of Hive is called Hive Query Language(HQL)
- 2 basic components: Hive Command Line and JDBC/ODBC driver
- Supports user defined functions (UDF) to accomplish specific needs

Hadoop Ecosystem – MAHOUT

Machine Learning

- Provides an environment for creating machine learning applications
- It performs collaborative filtering, clustering and classification
- Provides a command line to invoke various algorithms.
- It has a predefined set of library which already contains different inbuilt algorithms for different use cases.

Hadoop Ecosystem – Spark

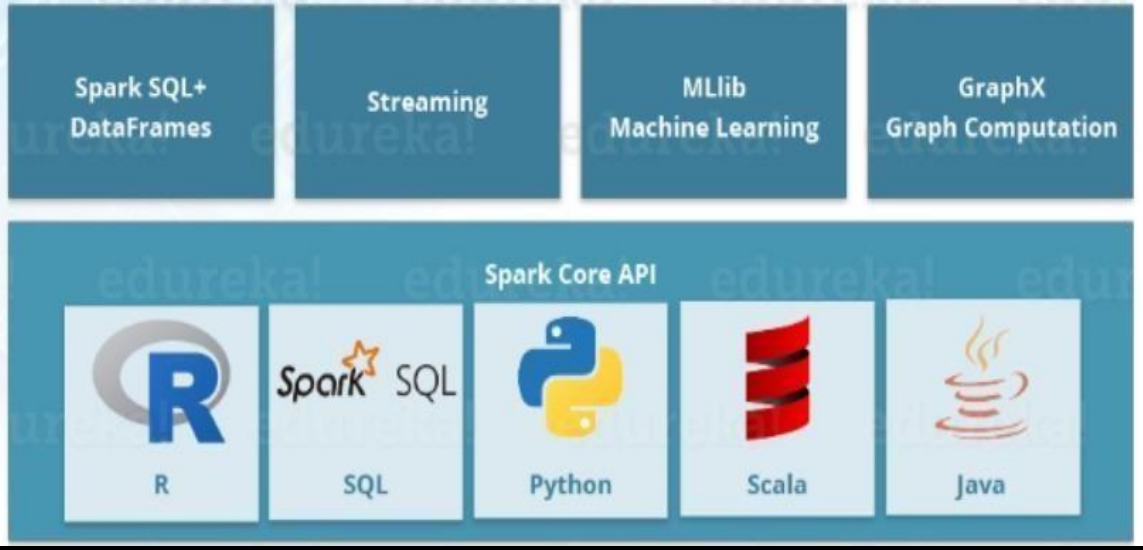
In Memory Data Processing

- A framework for real time data analytics in a distributed computing environment.
- Written in Scala and was originally developed at the University of California, Berkeley.
- It executes in-memory computations to increase speed of data processing over Map-Reduce.
- 100x faster than Hadoop for large scale data processing by exploiting in-memory computations

Hadoop Ecosystem – Spark

In Memory Data Processing

- Spark comes packed with high-level libraries
- Provides various services like MLlib, GraphX, SQL + Data Frames, Streaming services
- Supports various languages like R, SQL, Python, Scala, Java
- Seamlessly integrates in complex workflow



Hadoop Ecosystem – HBASE

NO Sql DataBase

- An open source, non-relational distributed database - a **NoSQL** database
- Supports all types of data and that is why, it's capable of handling anything and everything
- It is modelled after Google's BigTable
- It gives us a fault tolerant way of storing sparse data
- It is written in Java, and HBase applications can be written in REST, Avro and Thrift APIs

Hadoop Ecosystem – Drill

Sql on Hadoop

- An open source application which works with distributed environment to analyze large data sets
- Follows the ANSI SQL
- Supports different kinds NoSQL databases and file systems
- For example: Azure Blob Storage, Google Cloud Storage, HBase, MongoDB, MapR-DB HDFS, MapR-FS, Amazon S3, Swift, NAS and local files
- Combines a variety of data stores just by using a single query

Hadoop Ecosystem – OOZIE

Job Schedular

- Oozie is a job scheduler in Hadoop ecosystem
- Two kinds of Oozie jobs: Oozie workflow and Oozie Coordinator
- **Oozie workflow**: Sequential set of actions to be executed
- **Oozie Coordinator**: Oozie jobs which are triggered when the data is made available to it or even triggered based on time

Hadoop Ecosystem – Flume

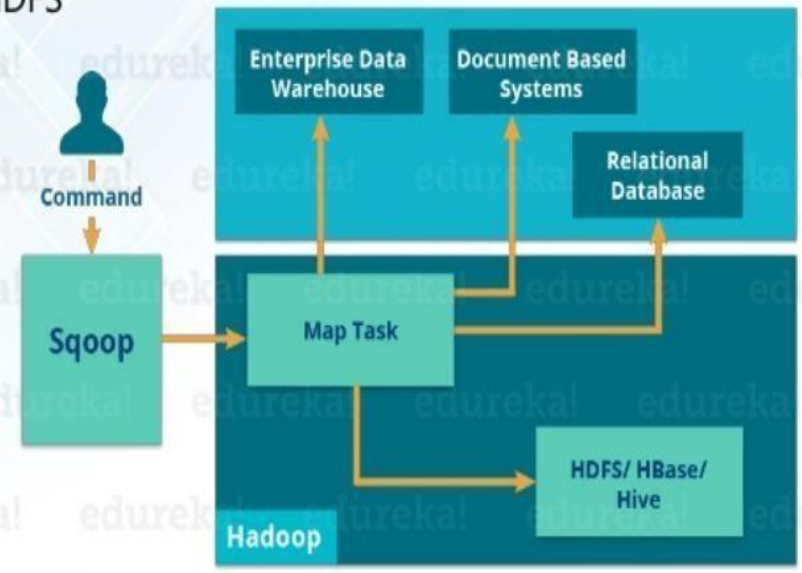
Data Ingesting Servive

- Ingests unstructured and semi-structured data into HDFS.
- It helps us in collecting, aggregating and moving large amount of data sets.
- It helps us to ingest online streaming data from various sources like network traffic, social media, email messages, log files etc. in HDFS.

Hadoop Ecosystem – Sqoop

Data Ingesting Servive

- Another data ingesting service
- Sqoop can import as well as export structured data from RDBMS
- Flume only ingests unstructured data or semi-structured data into HDFS



Hadoop Ecosystem – Ambari

Cluster Manager

- Software for provisioning, managing and monitoring Apache Hadoop clusters
- Gives us step by step process for installing Hadoop services
- Handles configuration of Hadoop services
- Provides a central management service for starting, stopping and re-configuring Hadoop services
- Monitors health and status of the Hadoop cluster

Apache Ambari

Hadoop Ecosystem – ZooKeeper

Coordinator

- An open-source server which enables highly reliable distributed coordination
- Apache Zookeeper coordinates with various Hadoop services in a distributed environment
- Performs synchronization, configuration maintenance, grouping and naming

THANK YOU