OS for Big Data – Hadoop Architecture

Unit II Syllabus

- Data Product, Building Data Products at Scale with Hadoop, Data Science Pipeline and Hadoop Ecosystem

  Ref: (Chapter 1 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- Operating System for Big Data: Concepts, Hadoop Architecture, Working with Distributed file system, Working with Distributed Computation

  Ref: (Chapter 2 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- Framework for Python and Hadoop Streaming, Hadoop Streaming, MapReduce with Python, Advanced MapReduce.

  Ref: (Chapter 3 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- In-Memory Computing with Spark, Spark Basics, Interactive Spark with PySpark, Writing Spark Applications

  Ref: (Chapter 4 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

Today’s Topics

Operating System for Big Data

- Hadoop Architecture
- Working with Distributed file system
- Working with Distributed Computation

Ref: (Chapter 2 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

Components of Hadoop

Hadoop HDFS

- Data is stored in a distributed manner in HDFS. There are two components of HDFS - name node and data node. While there is only one name node, there can be multiple data nodes.

- Hadoop enables you to use commodity machines as your data nodes. This way, you don’t have to spend millions of dollars just on your data nodes. However, the name node is always an enterprise server.

Hadoop HDFS…

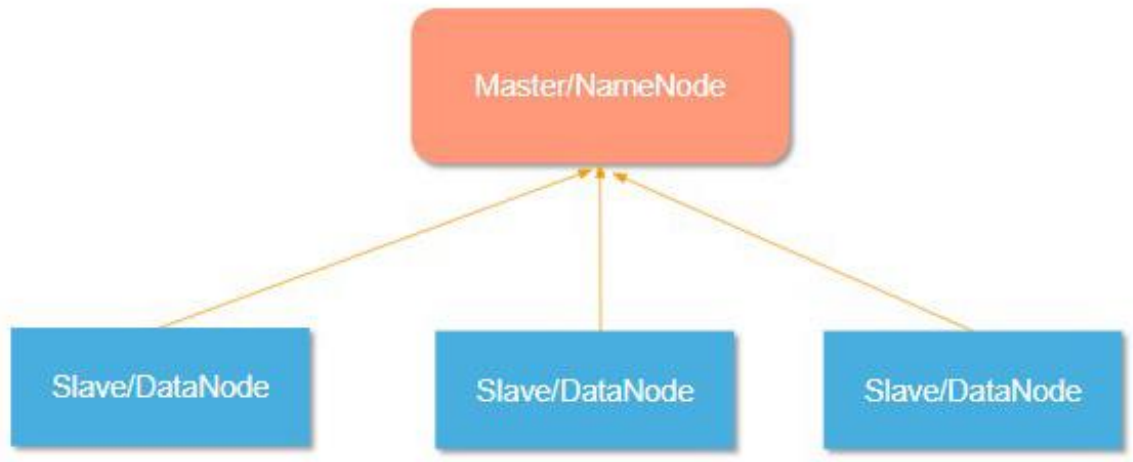
Features of HDFS

- Provides distributed storage
- Can be implemented on commodity hardware
- Provides data security
- Highly fault-tolerant - If one machine goes down, the data from that machine goes to the next machine

Master and Slave Nodes

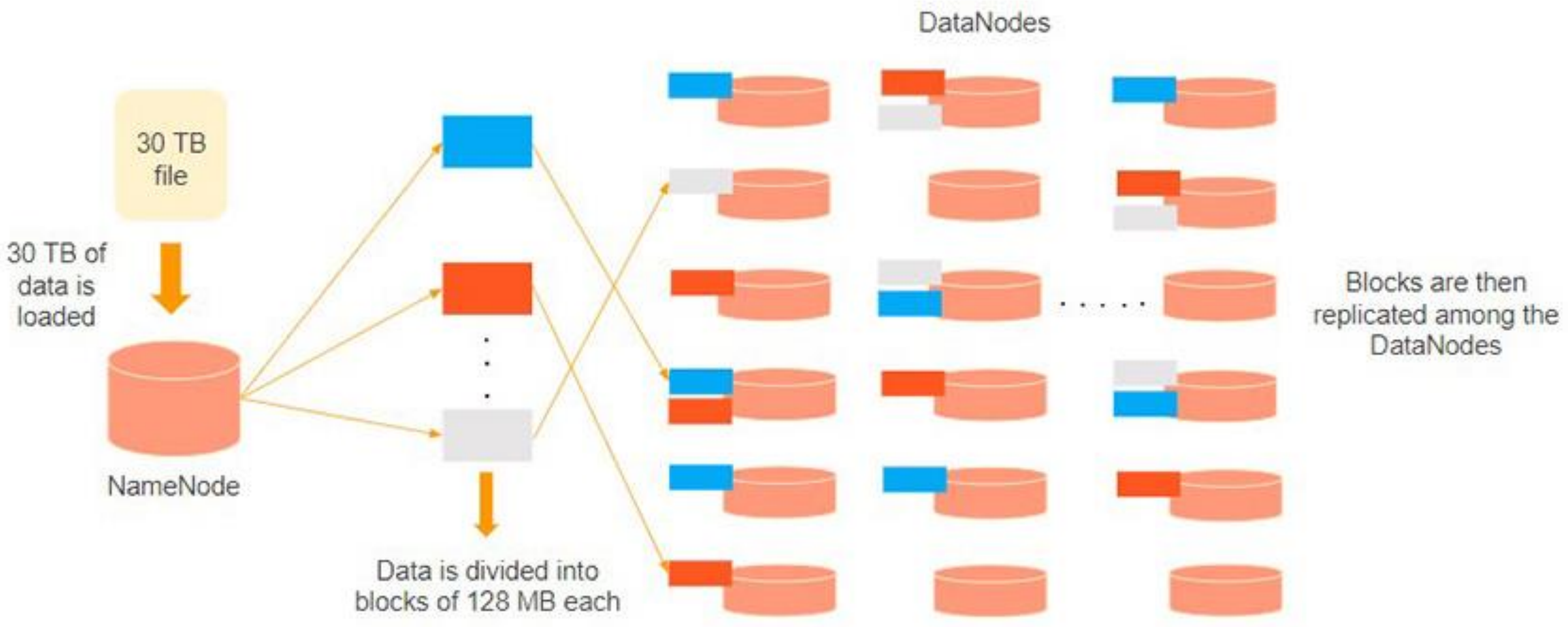
- Master and slave nodes form the HDFS cluster. The name node is called the master, and the data nodes are called the slaves.

- The name node is responsible for the workings of the data nodes. It also stores the metadata.



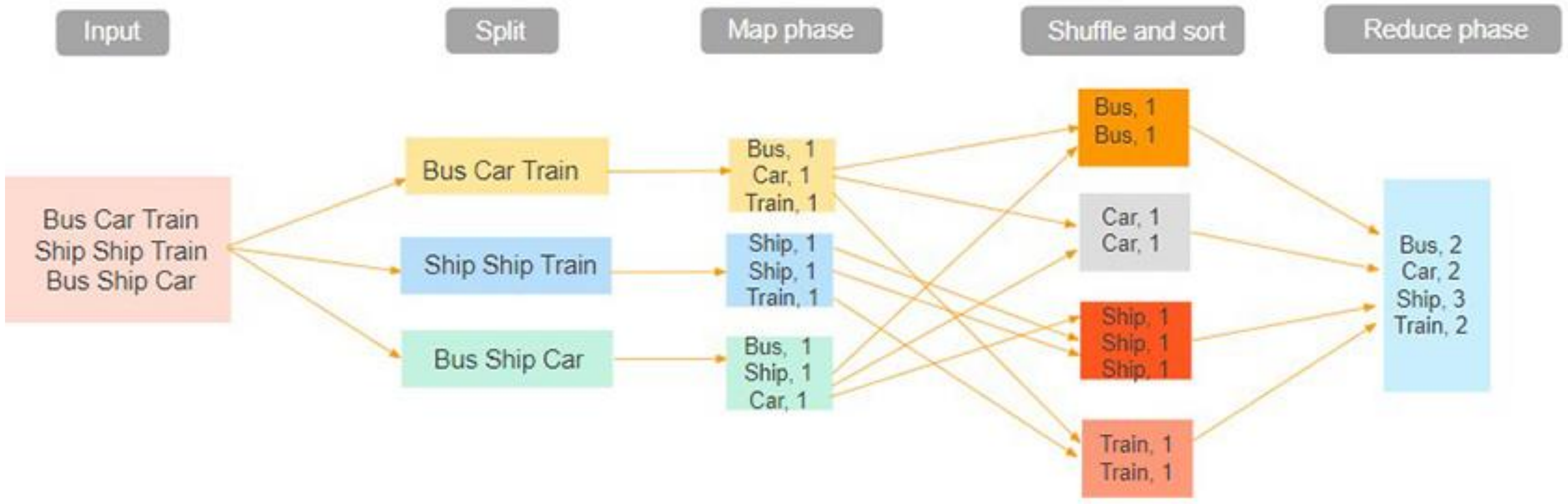
Master and Slave Nodes

- The data nodes read, write, process, and replicate the data. They also send signals, known as heartbeats, to the name node. These heartbeats show the status of the data node.


DataNodes
30 TB file
30 TB of data is loaded
NameNode
Data is divided into blocks of 128 MB each
. . . . .
Blocks are then replicated among the DataNodes

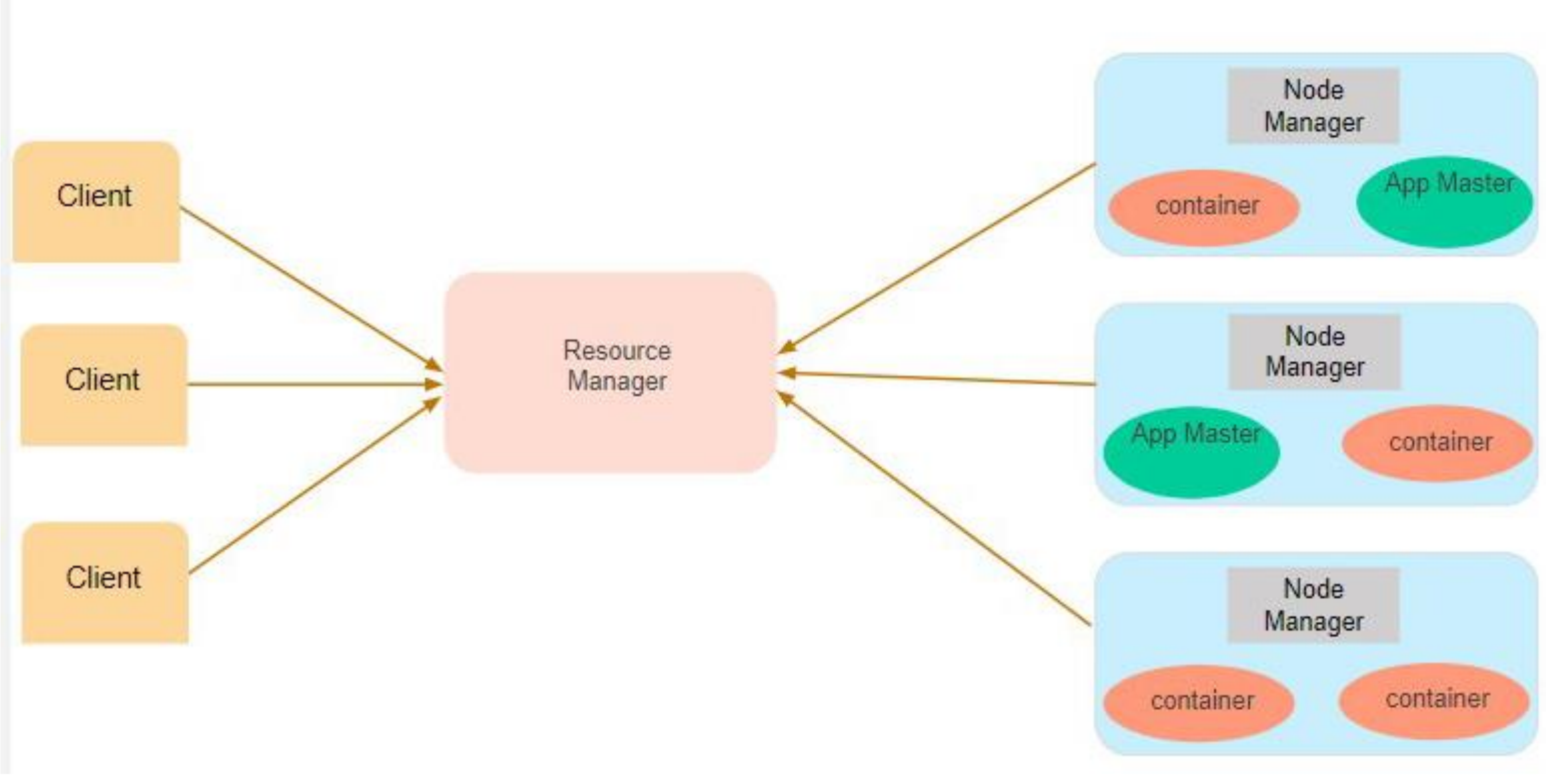
Hadoop MapReduce

Hadoop MapReduce is the processing unit of Hadoop. In the MapReduce approach, the processing is done at the slave nodes, and the final result is sent to the master node.



Hadoop YARN

- Hadoop YARN stands for Yet Another Resource Negotiator. It is the resource management unit of Hadoop and is available as a component of Hadoop version 2.



Architecture of Hadoop

Hadoop HDFS

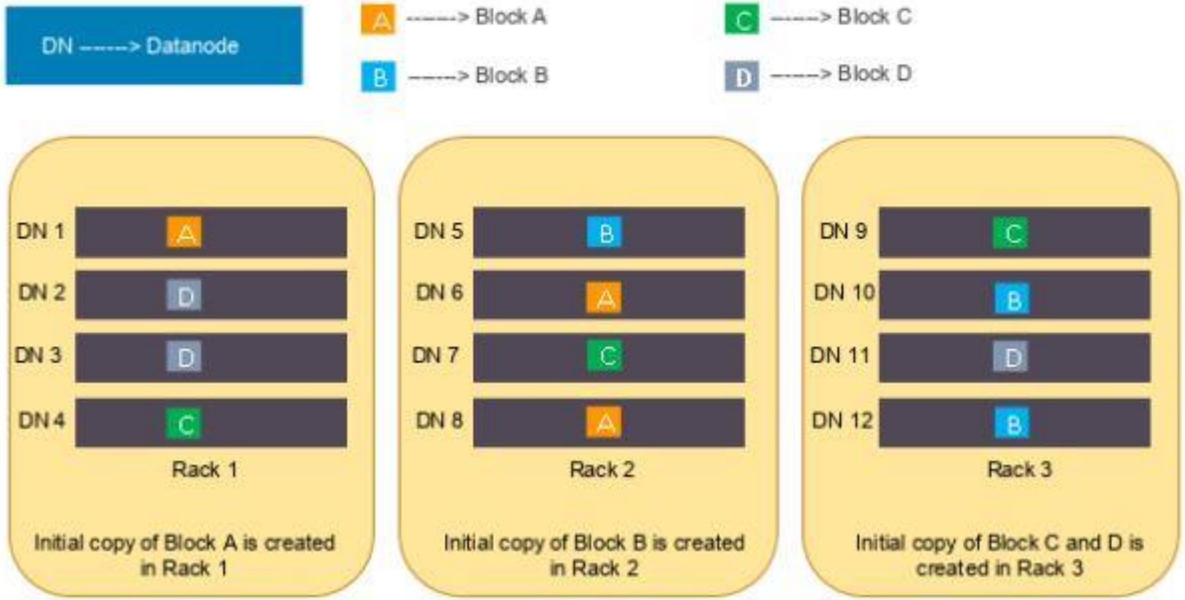
Hadoop is a framework permitting the storage of large volumes of data on node systems. The Hadoop architecture allows parallel processing of data using several components:

- Hadoop HDFS to store data across slave machines
- Hadoop YARN for resource management in the Hadoop cluster
- Hadoop MapReduce to process data in a distributed fashion
- Zookeeper to ensure synchronization across a cluster

Hadoop HDFS…

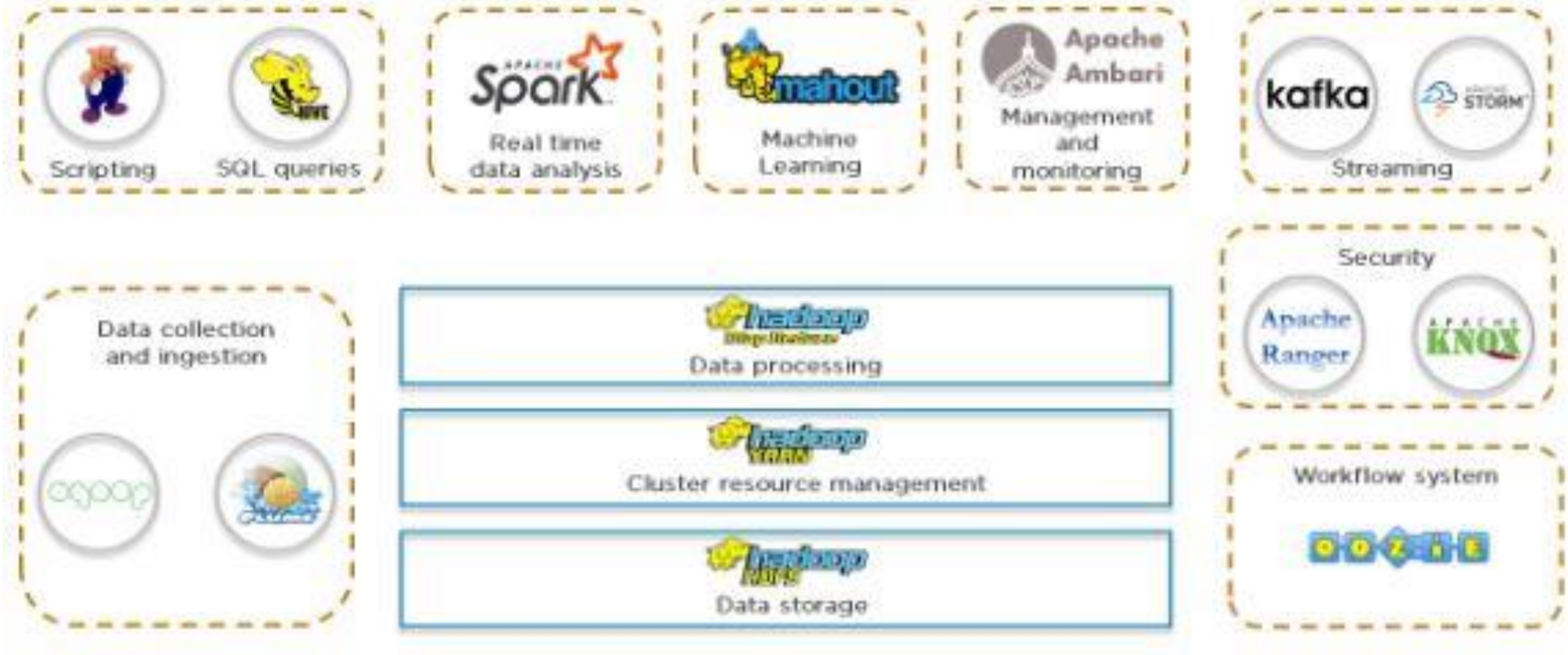
- DFS in Hadoop Architecture divides large data into different blocks. Replicated three times by default, each block contains 128 MB of data. Replications operate under two rules:

1. Two identical blocks cannot be placed on the same DataNode
2. When a cluster is rack aware, all the replicas of a block cannot be placed on the same rack

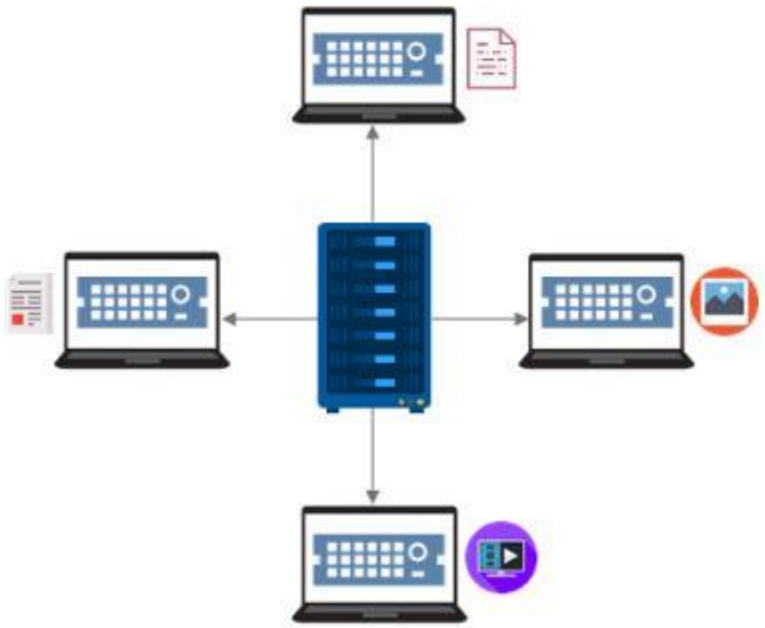


Hadoop Ecosystem

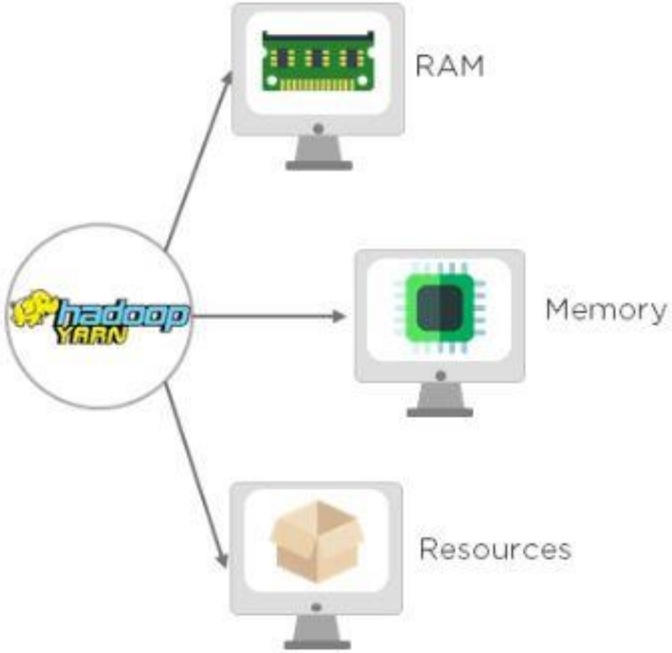
Hadoop Ecosystem



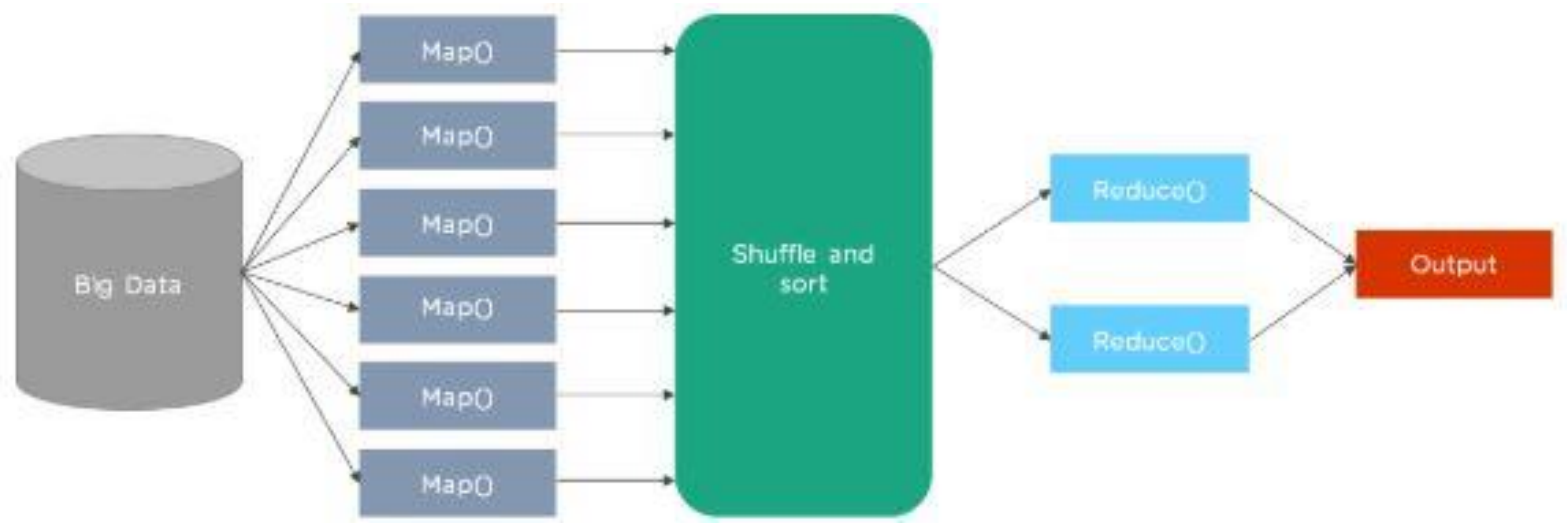
HDFS



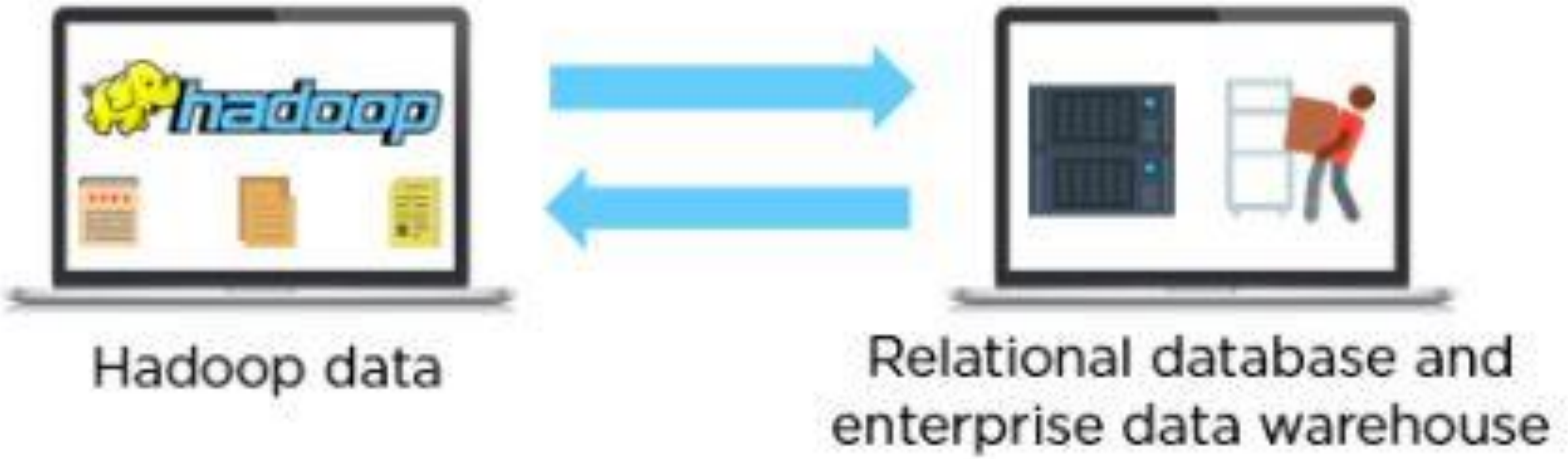
YARN



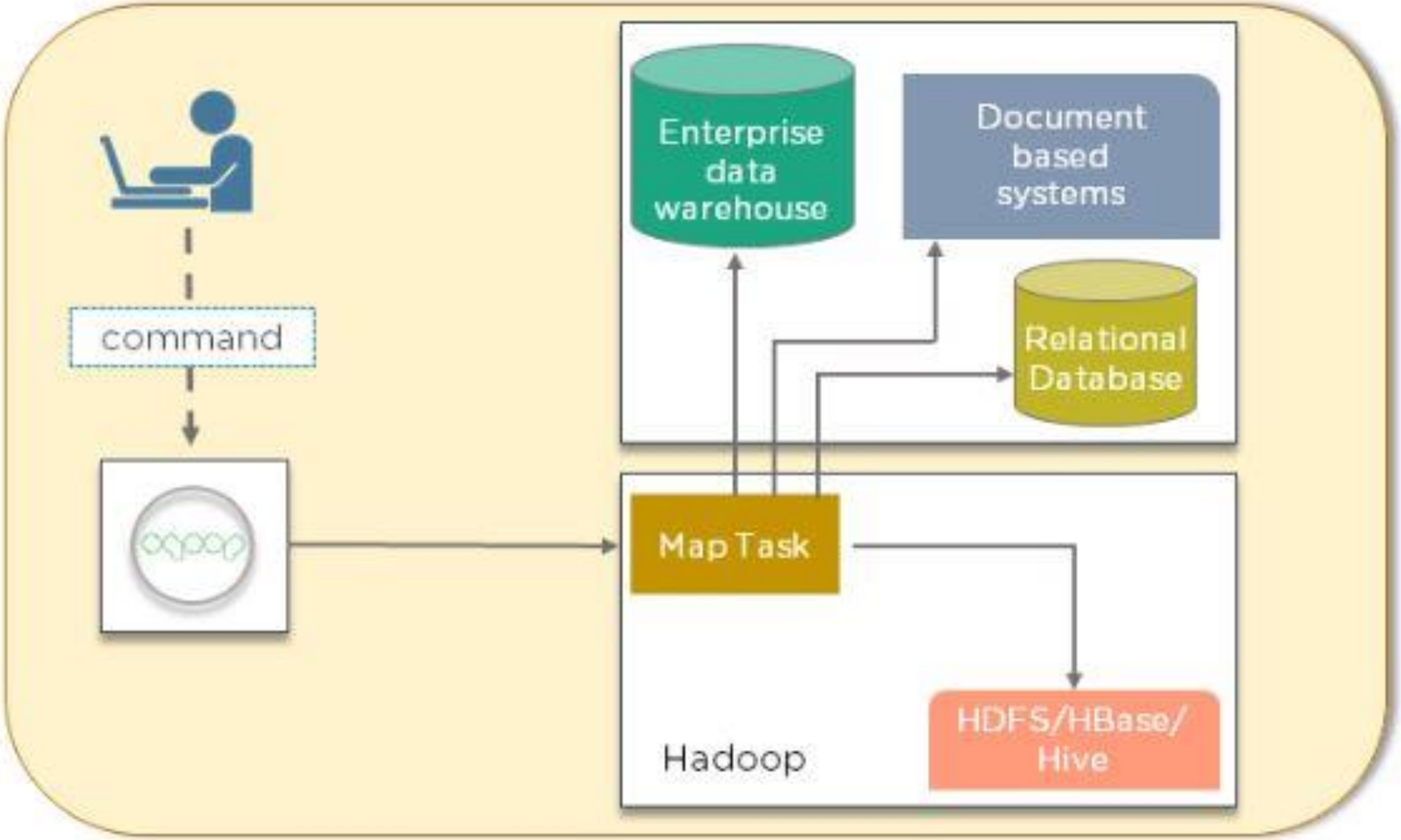
MapReduce


Big Data
Map()
Map()
Map()
Map()
Map()
Map()
Shuffle and sort
Reduce()
Reduce()
Output

Sqoop

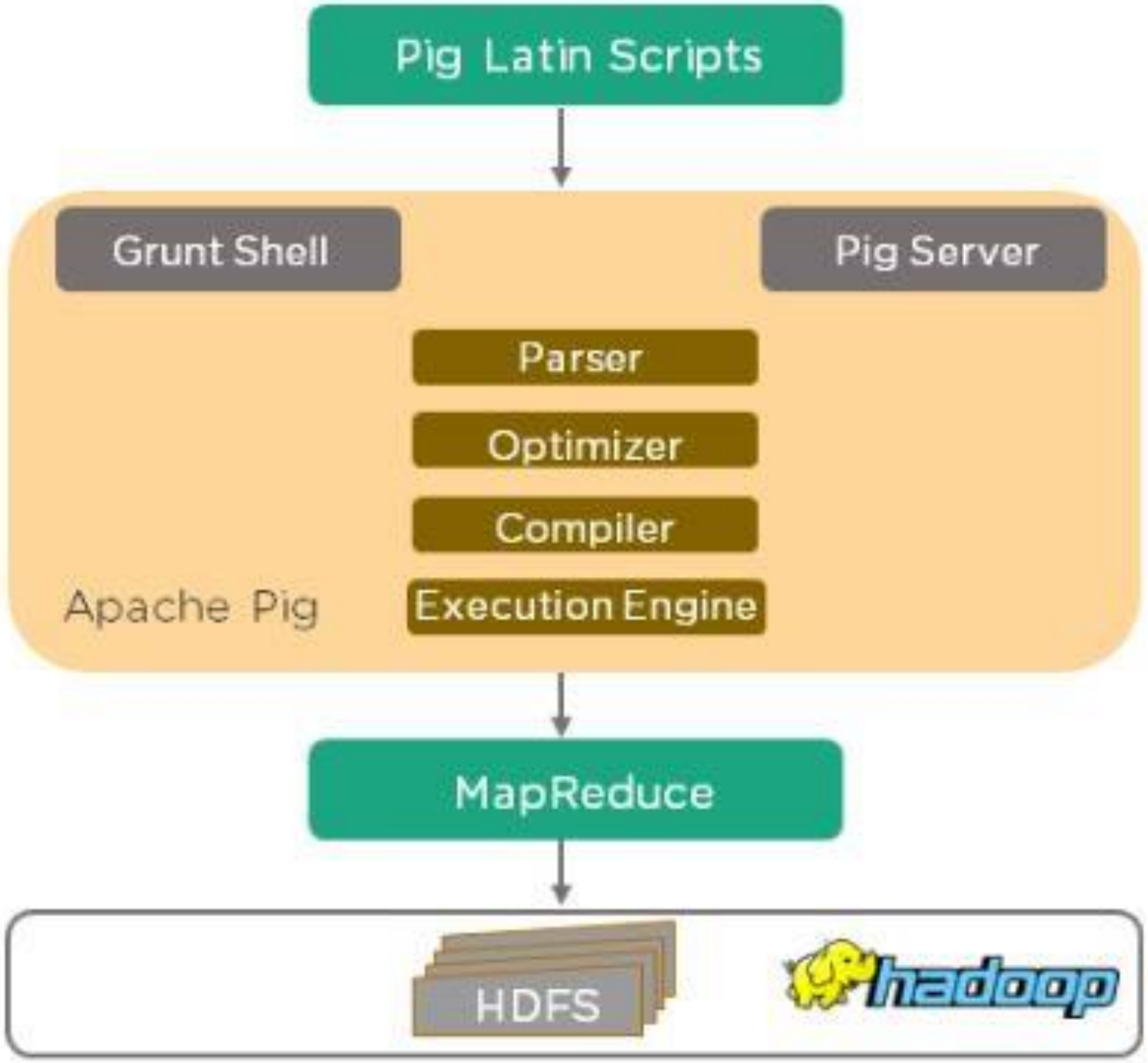

hadoop
Hadoop data
Relational database and enterprise data warehouse

Flume

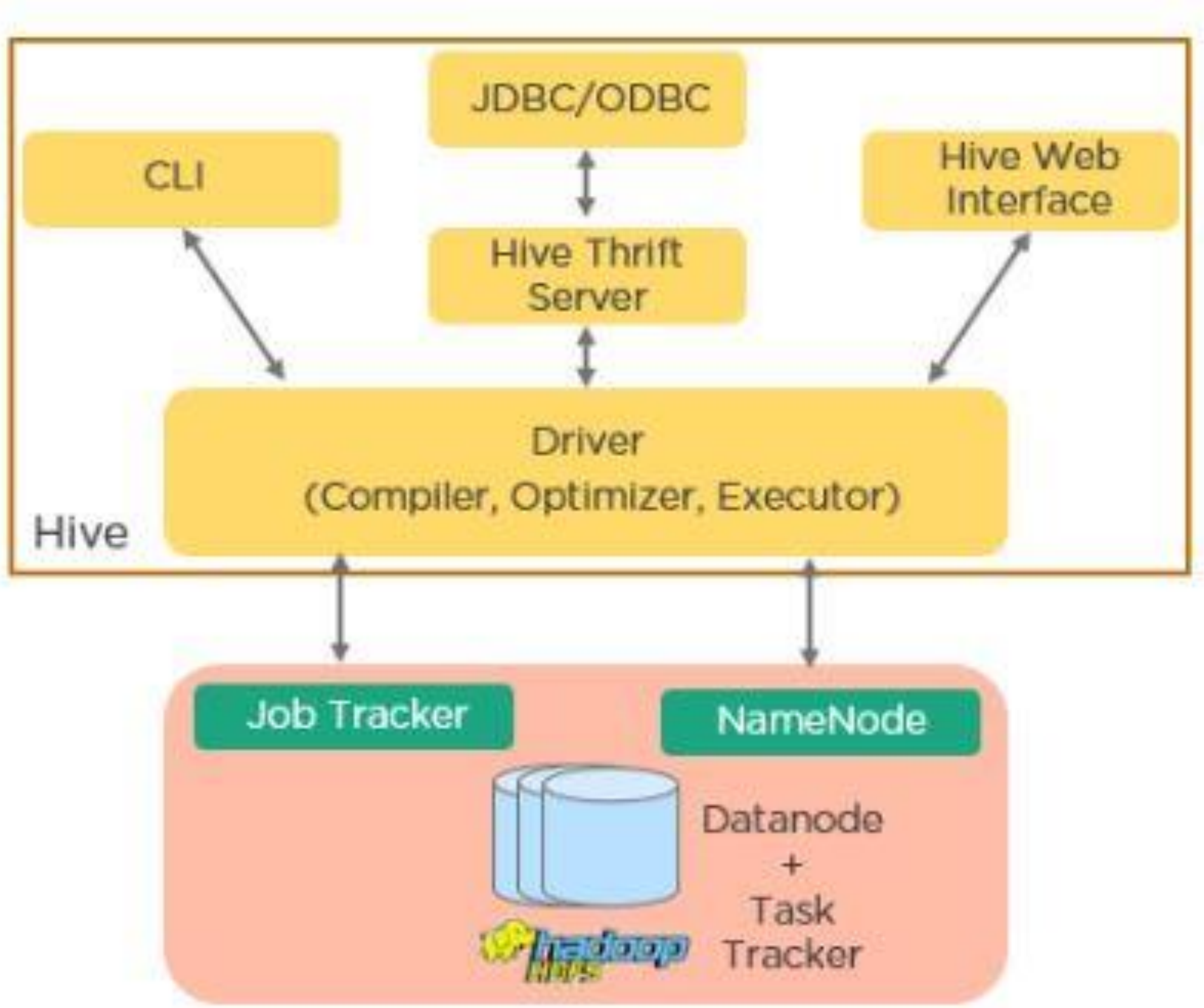

command
Enterprise data warehouse
Document based systems
Relational Database
Map Task
HDFS/HBase/ Hive
Hadoop

Pig

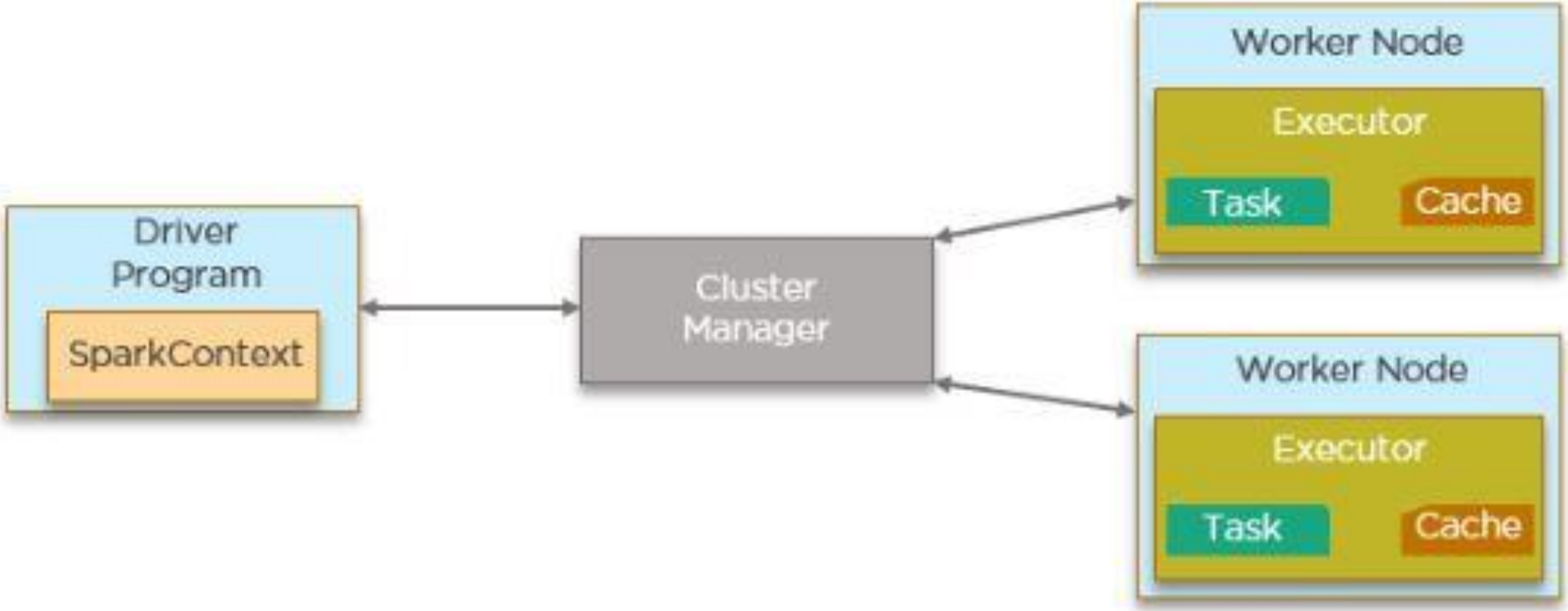
1.Pig Latin - This is the language for scripting
2.Pig Latin Compiler - This converts Pig Latin code into executable code



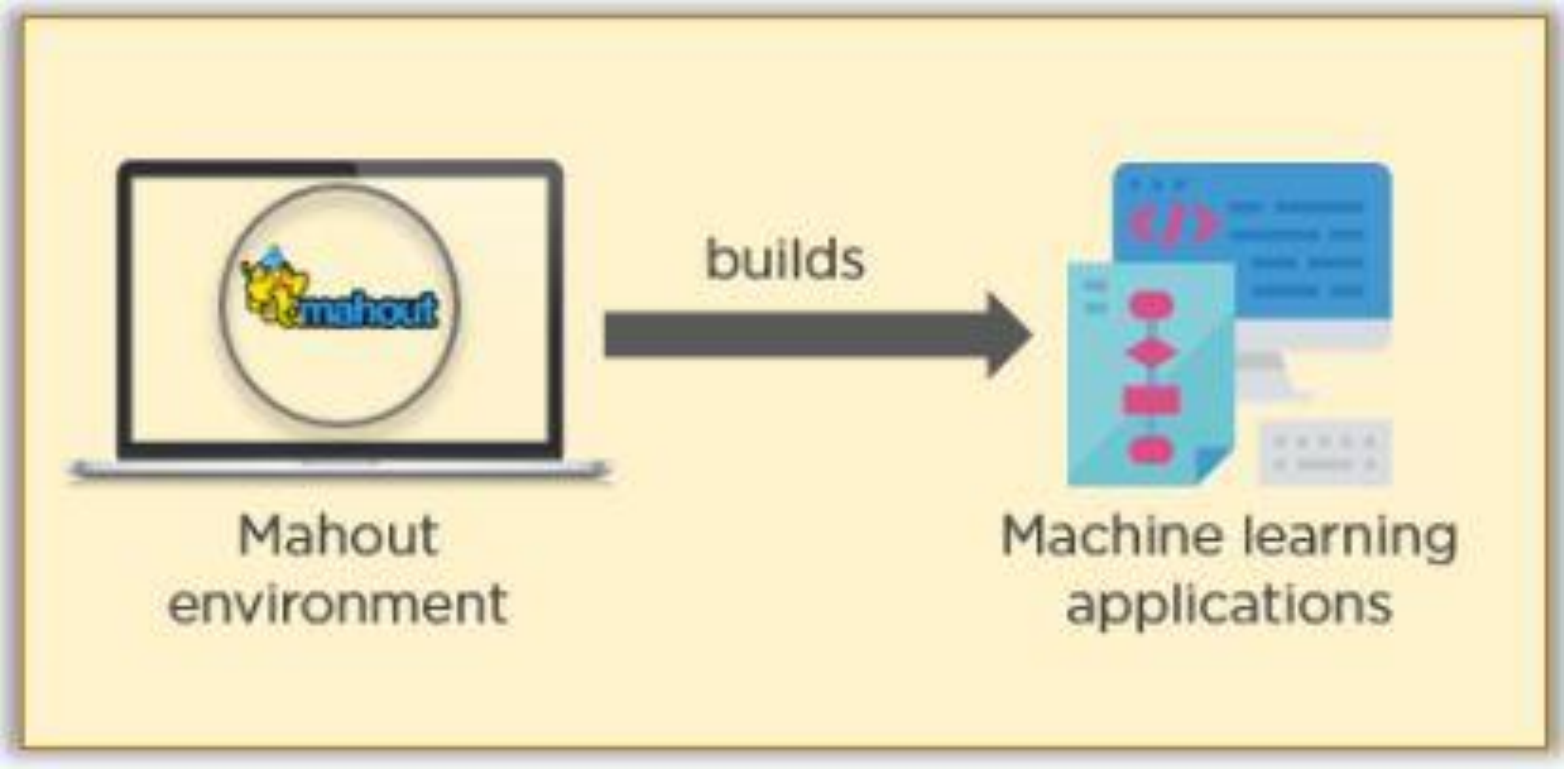
Hive


JDBC/ODBC
CLI
Hive Web Interface
Hive Thrift Server
Driver
(Compiler, Optimizer, Executor)
Hive
Job Tracker
NameNode
Datanode
+
Task
Tracker
hadoop
HDFS

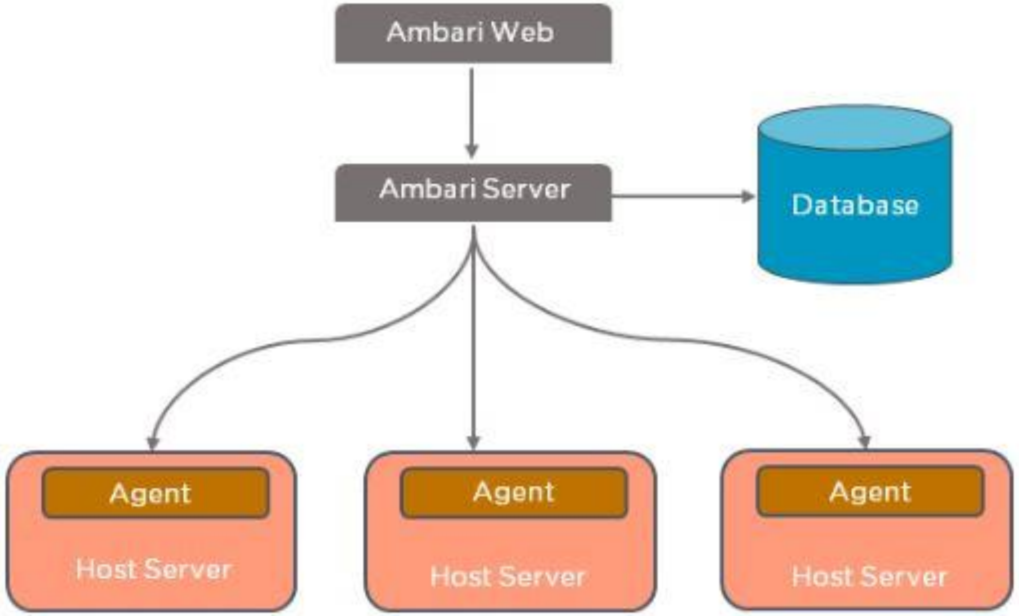
Spark


Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Task
Cache
Worker Node
Executor
Task
Cache

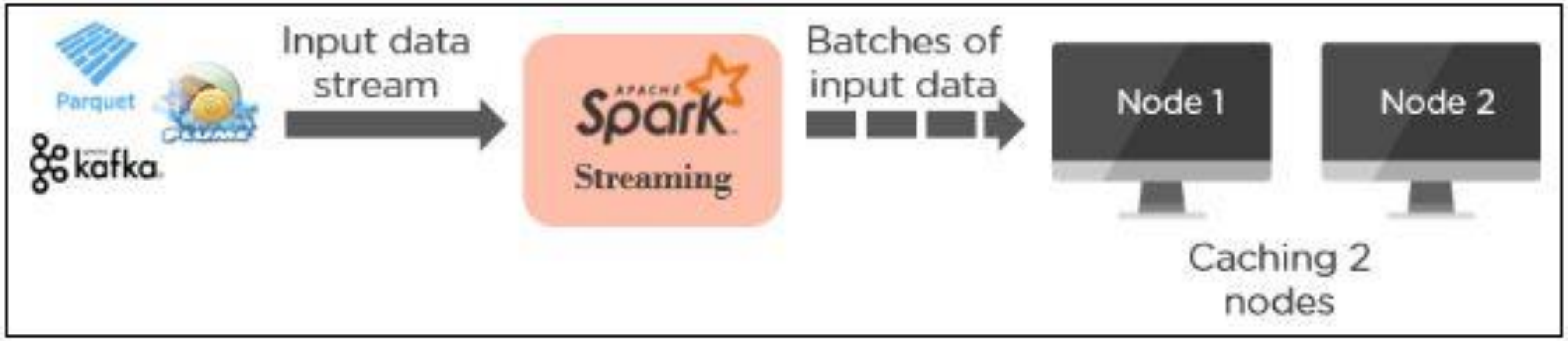
Mahout

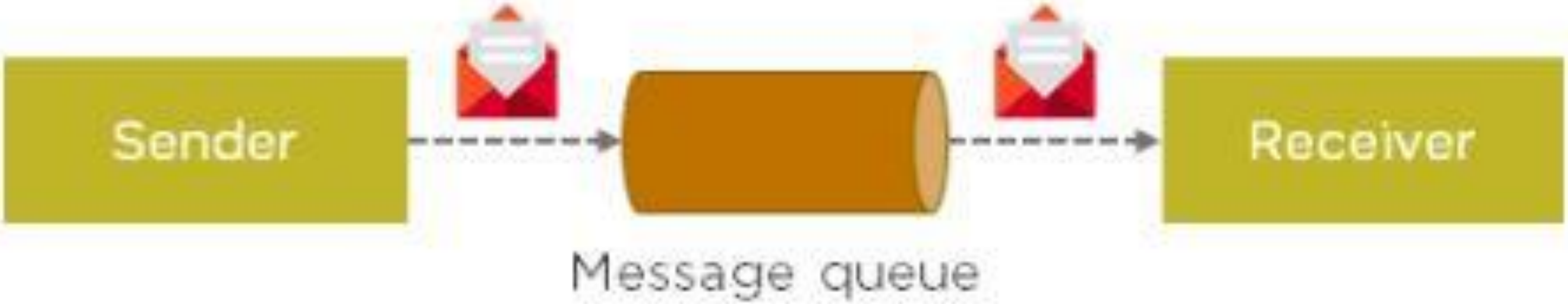

mahout
builds
Mahout environment
Machine learning applications

Ambari

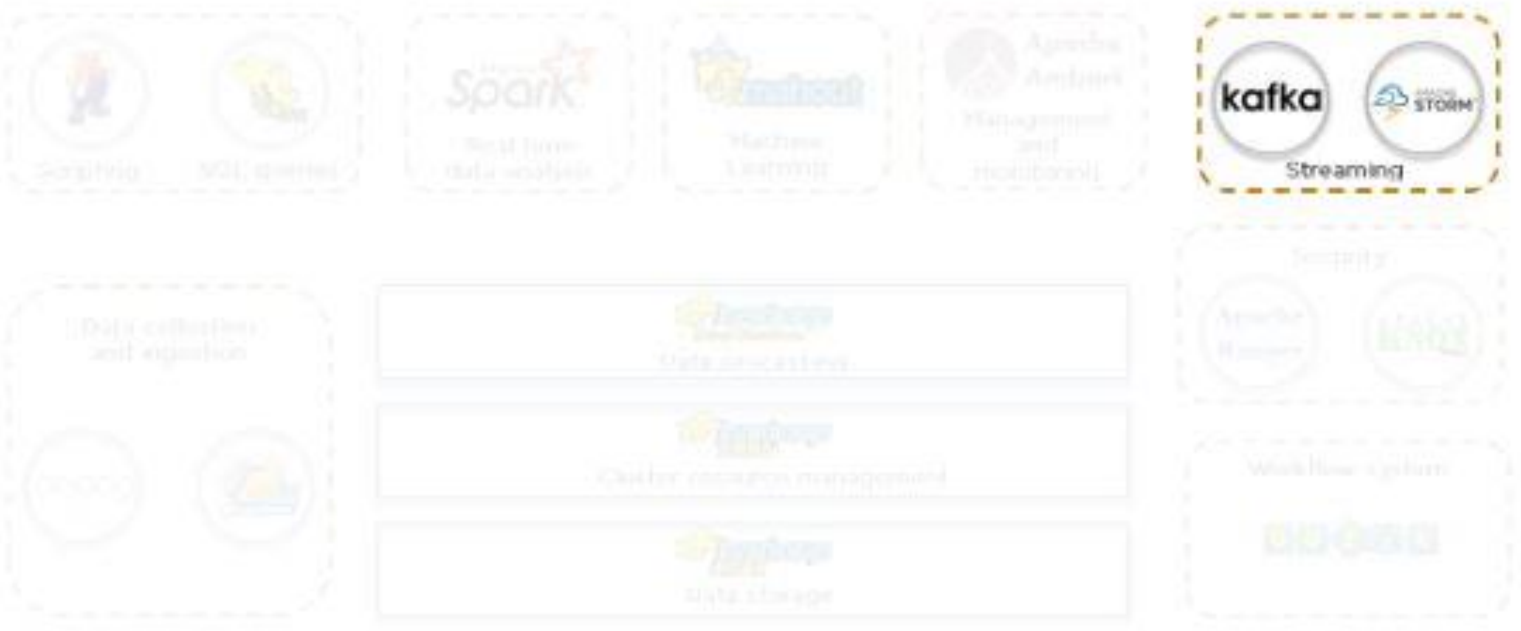

Ambari Web
Ambari Server
Database
Agent
Host Server
Agent
Host Server
Agent
Host Server

Kafka

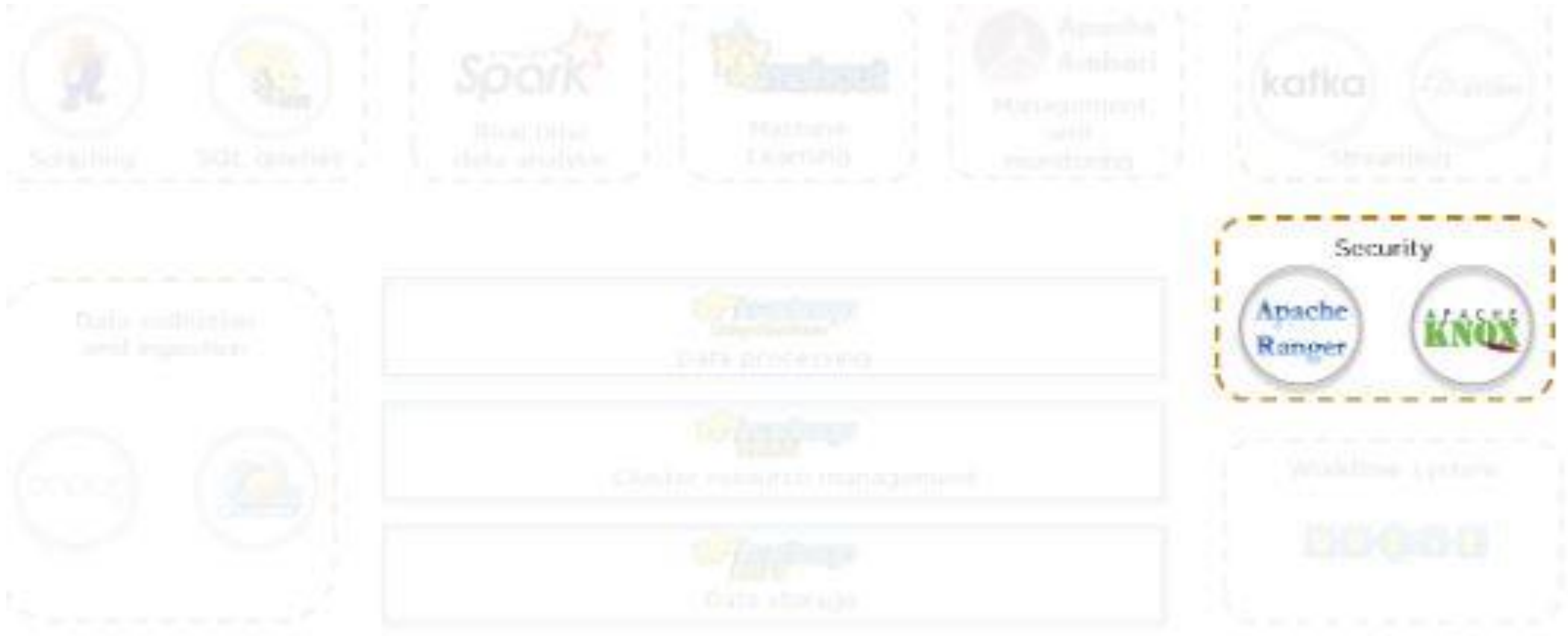

Parquet
kafka
Input data stream
Apache Spark Streaming
Batches of input data
Node 1
Node 2
Caching 2 nodes


Sender
Receiver
Message queue

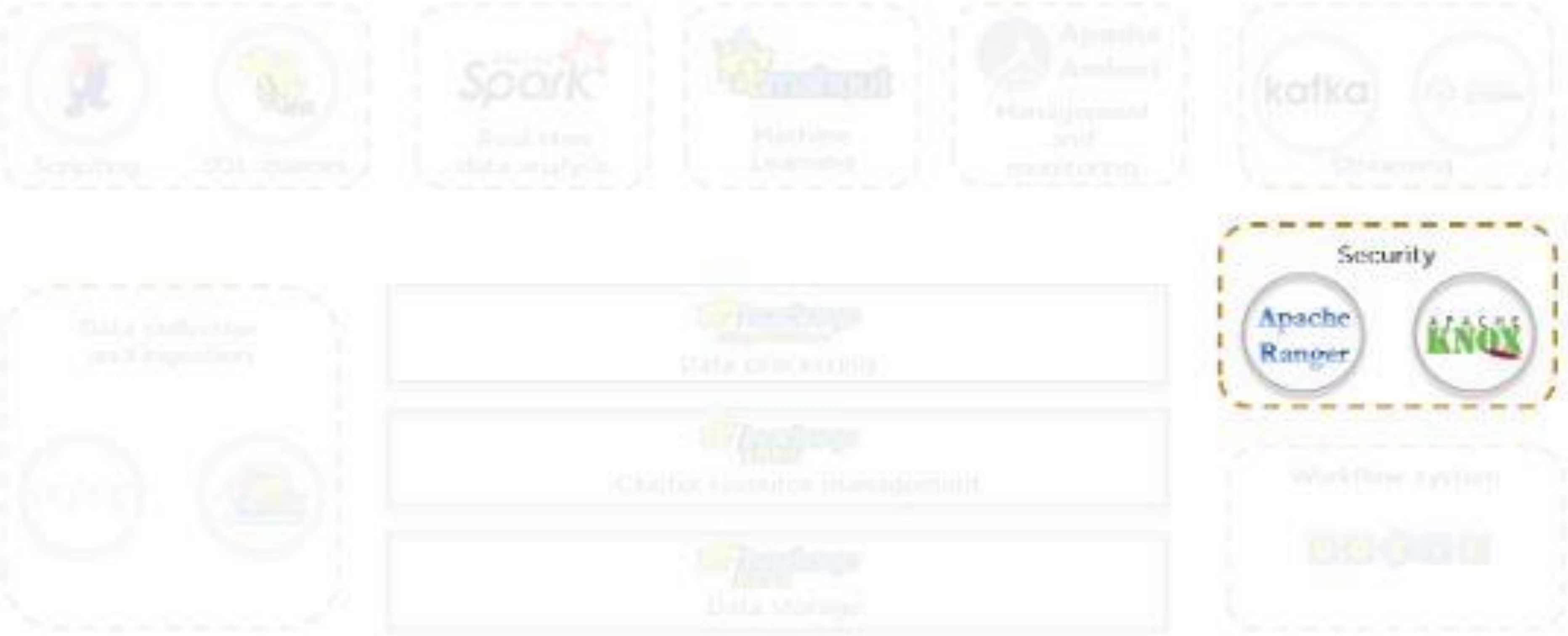
Storm


Scripting
SQL queries
Spark
Real time data analysis
Machine Learning
Apache Ambari
Management and Monitoring
kafka
STORM
Streaming
Security
Apache Ranger
Data collection and ingestion
Data processing
Cluster resource management
Data storage
Workflow system

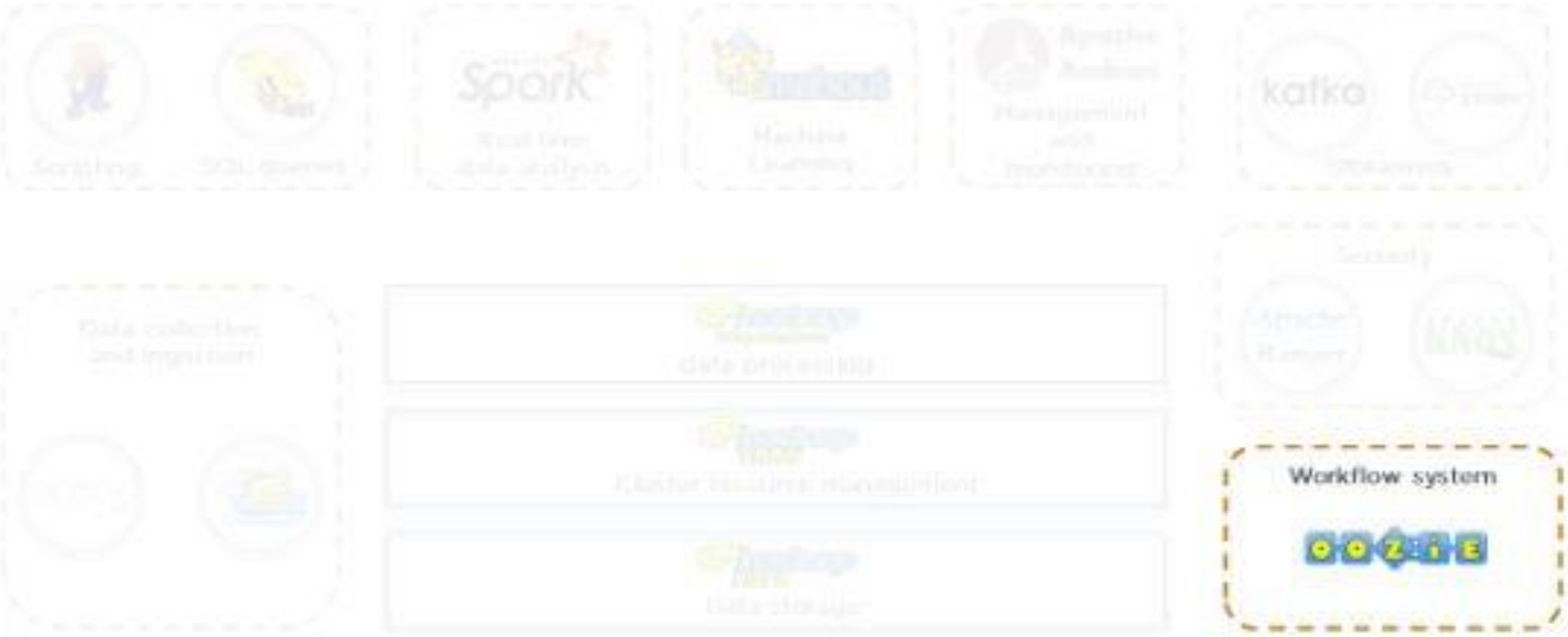
Ranger


Scripting
SQL queries
Spark
Real time data analysis
Machine Learning
Apache Ambari
Management and monitoring
kafka
Streaming
Security
Apache Ranger
APACHE KNOX
Data collection and ingestion
Hadoop
Data processing
Hadoop
Cluster resource management
Hadoop
Data storage
Workflow system

Knox


Scripting
SQL queries
Spark
Real-time data analysis
Machine Learning
Apache Ambari
Management and monitoring
kafka
Streaming
Security
Apache Ranger
APACHE KNOX
Data collection and ingestion
Data processing
Cluster resource management
Data storage
Workflow system

Oozie


Scripting
SQL queries
Spark
Real time data analysis
Machine Learning
Apache Ambari
Management and monitoring
kafka
Streaming
Security
Apache Ranger
Data collection and ingestion
data processing
Cluster resource management
YARN
HDFS
Data storage
Workflow system
OOZIE

Thank You….

Revise the topics from Syllabus References…

Fill Your Attendance Form….!

Syllabus References

1. Big Data and Analytics, Subhashini Chellappan Seema Acharya, Wiley
2. Data Analytics with Hadoop *An Introduction for Data Scientists*, Benjamin Bengfort and Jenny Kim, O'Reilly
3. Big Data and Hadoop, V.K Jain, Khanna Publishing

https://books.google.co.in/books?id=i6NODQAAQBAJ&pg=PA122&source=gbs_toc_r&cad=4#v=onepage&q&f=true