Lesson 1
Hadoop

Big Data Programming Model

- Distributed pieces of codes as well as the data at the computing nodes
- Distributed data storage systems do not use the concept of joins
- Hadoop provides that model

Big Data Distributed Computing Model in Hadoop

- Distributed model which requires no sharing between data nodes
- Multiple tasks of an application also distribute, run using machines associated with multiple data nodes and execute at the same time in parallel.

Big Data Storage Model in Hadoop

- Data partitions into data blocks and written at one set of nodes
- The blocks replicate at multiple nodes to take care of possibilities of network faults; (When a network fault occurs, then replicated node makes the data available)

Big Data Computing Model

- Fault tolerant due to replication
- Follows CAP theorem– out of three properties (consistency, availability and partitions), two must at least be present

Hadoop

- Hadoop consisted of two components: data store in blocks in the clusters and the other is computations at each individual cluster in parallel with another.
- Hadoop system uses the Big Data programming and storage models

Hadoop

- Jobs or tasks assigned and scheduled on the same servers which hold the data
- The system provides faster results from Big Data and from unstructured data as well

Hadoop Infrastructure

- Execution of instructions in two interrelated entities, such as a query and the database
- Cloud for clusters
- A cluster consists of sets of computers or PCs

Hadoop Platform

- Provides a low cost Big Data platform, which is open source and uses cloud services

Hadoop

- Tera Bytes of data processing takes just few minutes
- Hadoop enables distributed processing of large datasets (above 10 million bytes) across clusters of computers using a programming model called MapReduce.

Hadoop System Characteristics

- Scalable
- Self-manageable
- Self-healing
- Distributed file system

Scalability

- Means can be scaled up (enhanced) by adding storage and processing units as per the requirements failure.

Self Manageability

- Means creation of storage and processing resources which are used, scheduled and reduced or increased with the help of the system itself

Self Healing

- Means taken care of by the system itself in case of faults
- Enables functioning and resources availability
- Software detect and handle failures at the task level and also Software enable the task execution on communication failure.

Hadoop Hardware Need

- The hardware scales up from a single server to thousands of machines that store the clusters
- Each cluster stores a large number of data blocks in racks. Default data block size is 64 MB.
- IBM BigInsights, built on Hadoop deploys default 128 MB block size. of data.

Big Data analytics applications

- Software applications that leverage large-scale data
- The applications analyze Big Data using massive parallel processing frameworks
- Hadoop provides that framework

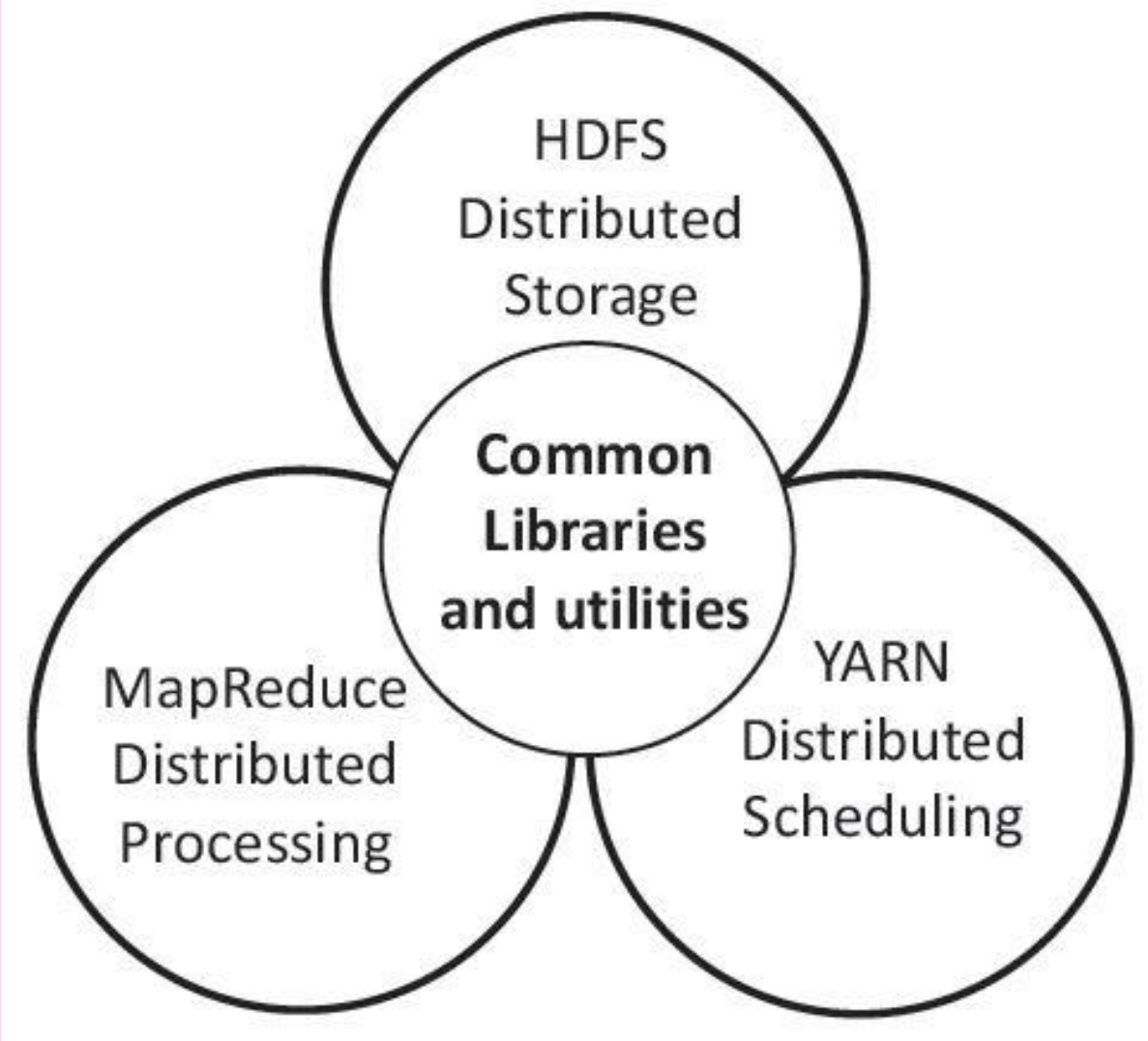
Hadoop Framework

- Provides the computing features of a system of distributed, flexible, scalable, fault tolerant computing with high computing power
- Provides an efficient platform for the distributed storage and processing of a large amount

Hadoop Big Data storage and cluster computing

- Manages both, large-sized structured and unstructured data in different formats, such as XML, JSON and text with efficiency and effectiveness
- Performs better with clusters of many servers when the focus is on horizontal scalability

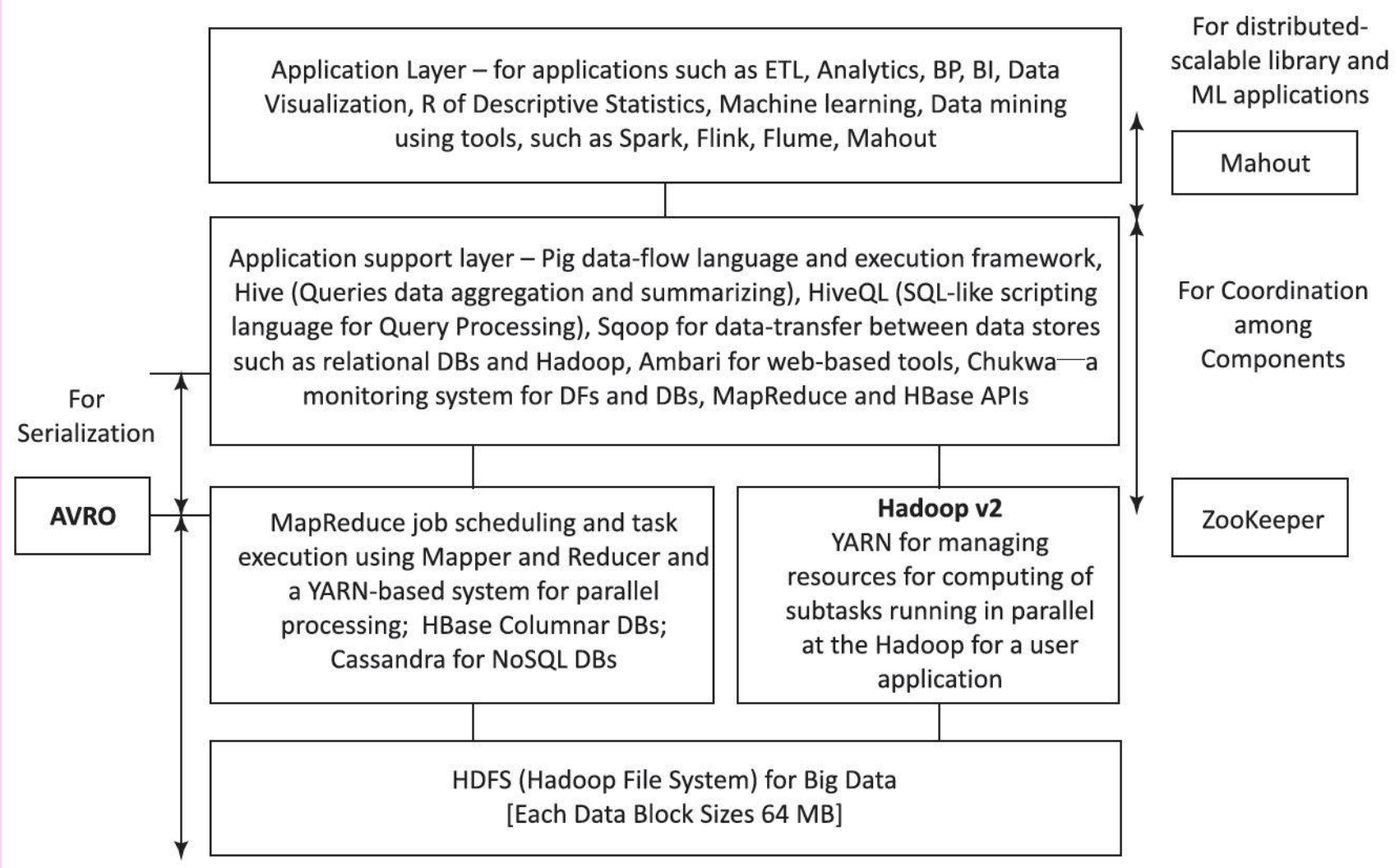
Figure 2.1 Core components of Hadoop



Hadoop

- Open Source Framework
- Java and Linux based: Hadoop uses Java interfaces
- Base is Linux but has its own set of shell commands support

Figure 2.2 Hadoop main components and ecosystem components



Summary

We learnt

- Hadoop Distributed model with pieces of codes as well as the data at the computing nodes which requires no sharing between data nodes
- Hadoop multiple tasks distribution, running using machines associated, execute at the same time in parallel

Summary

We learnt

- Partitionability
- Replication of Data
- Java, Linux based, Hadoop Shell Command Codes
- Hadoop Core Components and Ecosystem Tools

End of Lesson 1 on Hadoop