Lesson 2

Hadoop Distributed File System (HDFS)

HDFS

- HDFS─ a core component of Hadoop
- Designed to run on a cluster of computers and servers at cloud-based utility services
- HDFS stores Big Data which may range from GBs (1 GB = 2^{30} B) to PBs (1 PB = 10^{15} B, nearly the 2^{50} B)

HDFS– Hadoop Data Store Concept

- Implies storing the data at a number of clusters
- Each cluster has a number of data stores, called racks.
- Each rack stores a number of DataNodes
- Each DataNode has a large number of data blocks.

HDFS

- The racks distribute across a cluster
- The nodes have processing and storage capabilities
- The nodes have the data in data blocks to run the application tasks
- The data blocks replicate by default at least on three DataNodes at same or remote nodes.

HDFS File Storage

- A file, containing the data divides into data blocks. A data block default size is 64 MBs (HDFS division of files concept is similar to Linux or virtual memory page in Intel x86 and Pentium processors where the block size is fixed and is of 4 KB)

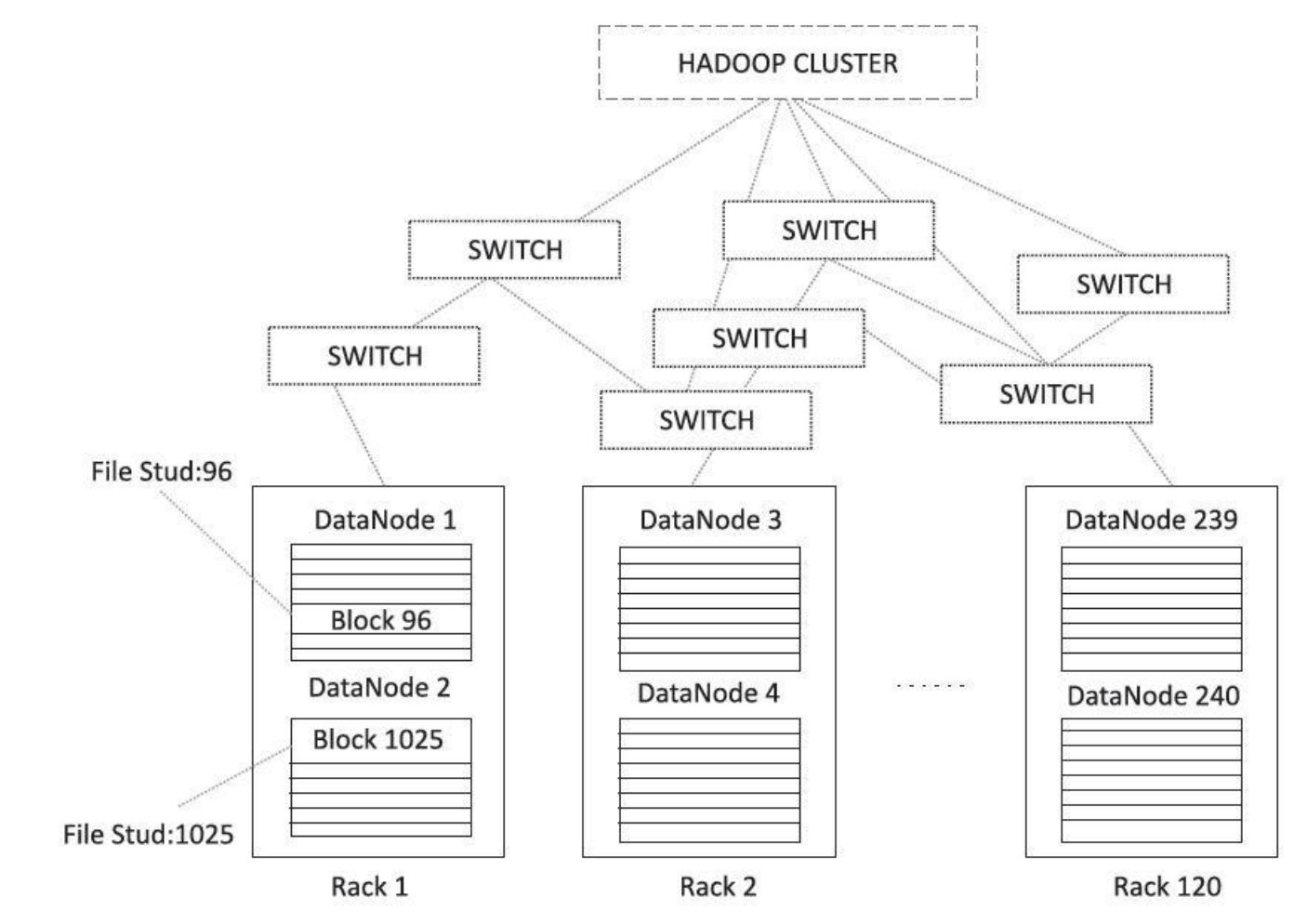
HDFS Running of Applications

- Data at the stores enable running the distributed applications including analytics, data mining, OLAP using the clusters

Example

- Example 2.2 for a data storage for University students in HDFS
- Each student data, stuData which is in a file of size less than 64 MB (1 MB = 2^{20}B). A data block stores the full file data for a student of stuData_idN, where N = 1 to 500

Example 2.2 Figure 2.3 A Hadoop cluster example, and the replication of data blocks in racks for two students of IDs 96 and 1025



Conventional file system

- Uses directories
- Each directory consists of folders
- Each folder consists of files
- When data processes, the data sources identify by pointers for the resources
- Resource Pointers at data-dictionary

…Conventional File System

- Master tables at the dictionary store at a central location. (Section 1.5.1 for
- the details)
- The centrally stored tables enable administration easier when the data sources change during processing

HDFS

- Similarly, the identification of data-blocks, DataNodes and Racks using MasterNodes (NameNodes) for processing the data at slave nodes
- A NameNode stores the file's meta data.

HDFS Metadata

- Meta data gives information about the file of user application, but does not participate in the computations
- The DataNode stores the actual data files in the data blocks.

Name Nodes (Master Nodes)

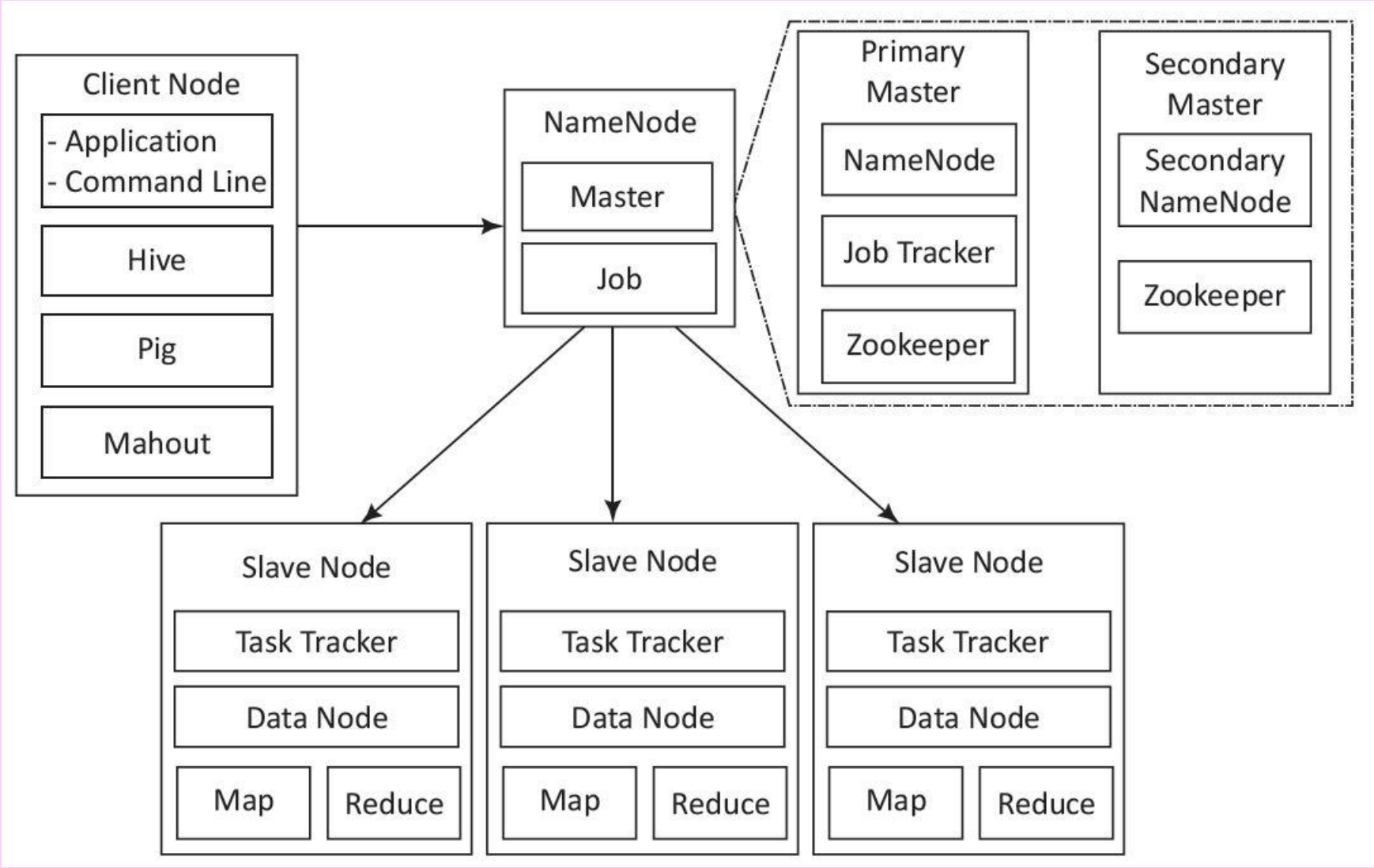
- Few nodes in a Hadoop cluster act as NameNodes, termed as MasterNodes
- Different configuration supporting high DRAM and processing power
- Masters use much less local storage
- Majority of the nodes in Hadoop cluster act as DataNodes and TaskTrackers.

File System Information

- NameNode stores all the file system related information such as:

1. The file section is stored in which part of the cluster
2. Last access time for the files
3. User permissions like which user has access to the file.

Figure 2.4 The Client, Master Namenode, Masternodes And Slave Nodes



ZooKeeper – Coordination service

- Zookeeper uses synchronization, serialization and coordination activities
- Enables functioning of a distributed system as a single function

Summary

We learnt :

- Hadoop Distributed File System
- Clusters, Racks, Data Nodes, Data blocks in HDFS
- Client Nodes
- Master (Name Nodes)
- Slave Nodes

End of Lesson 2 on

Hadoop Distributed File System (HDFS)