Week-1	Hadoop Framework
1. a)	Introduction to Hadoop Framework

1.1 Hadoop

- **Hadoop** is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation.
- The core of Apache Hadoop software framework consists of a storage part, known as **Hadoop Distributed File System (HDFS),** and a processing part which is a **MapReduce** programming model.

1.2 Key Aspects of Hadoop

- Open-source software
- Framework
- Distributed
- Massive storage
- Faster processing

1.3 Hadoop Core Components

1) HDFS

a) Storage component
b) Distributes data across several nodes
c) Natively redundant

2) MapReduce

a) Computational framework
b) Splits a task across multiple nodes
c) Processes data in parallel

1.4 Hadoop Ecosystem: Hadoop Ecosystem are support projects to enhance the functionality of Hadoop Core Components.

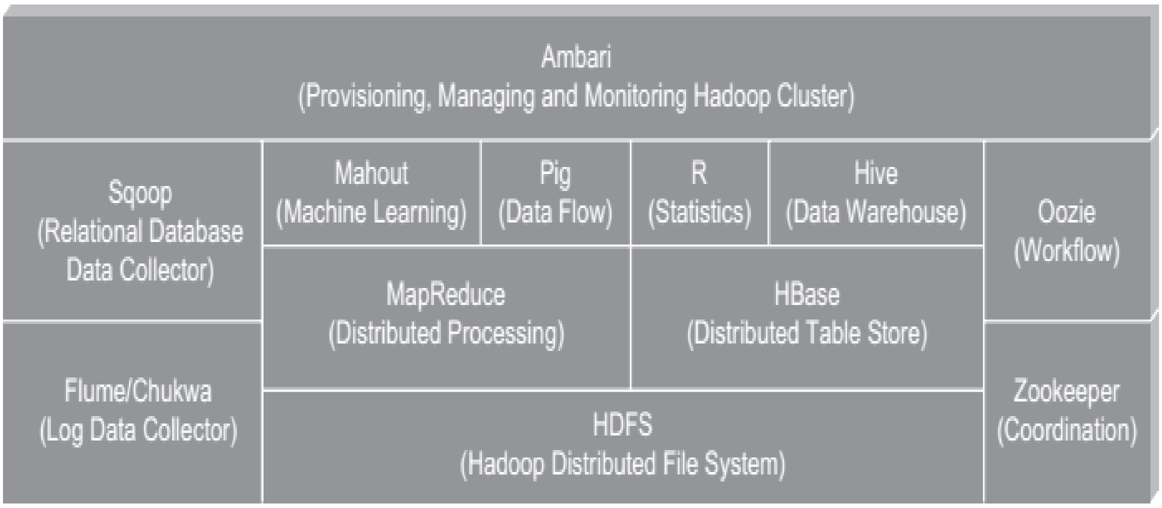


Fig 1a-: Hadoop Ecosystem

1.5 High-Level Architecture of Hadoop

- It is a distributed Master-Slave Architecture.
- Master node is known as NameNode and
- Slave nodes are known as DataNodes.

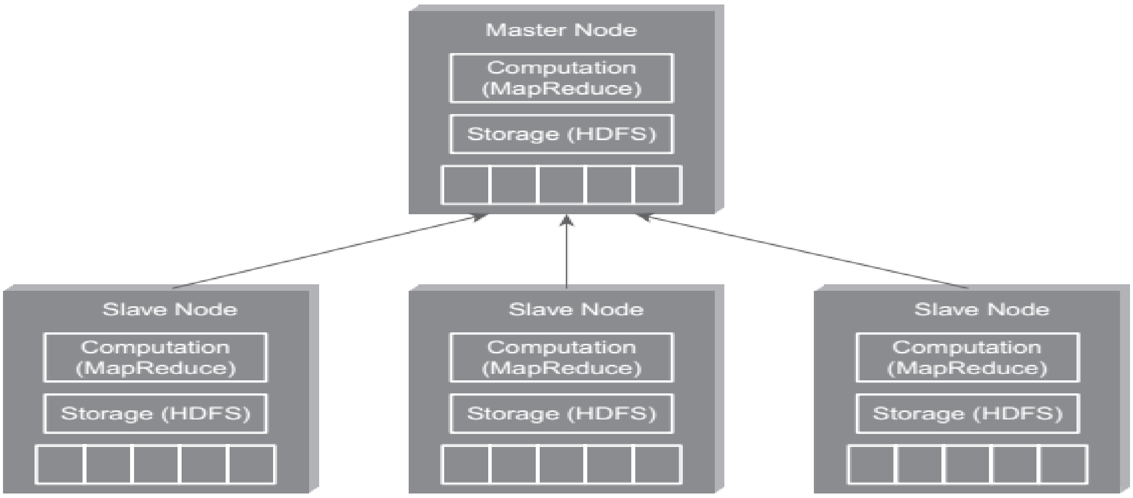
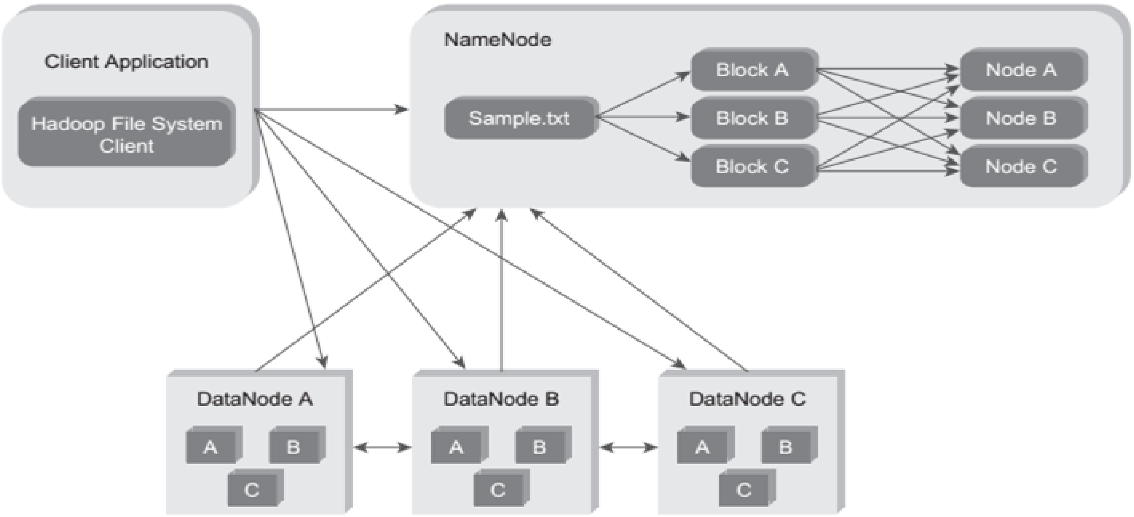


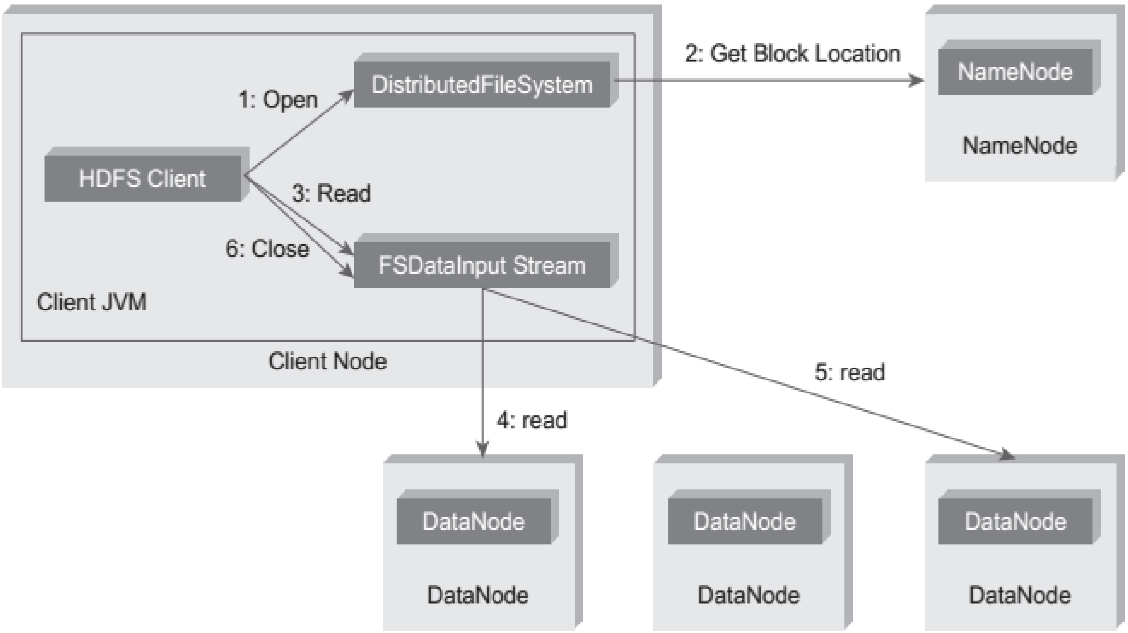
Fig 1b-: High-Level Architecture of Hadoop

1.6 Hadoop Distributed File System Architecture and HDFS Daemons

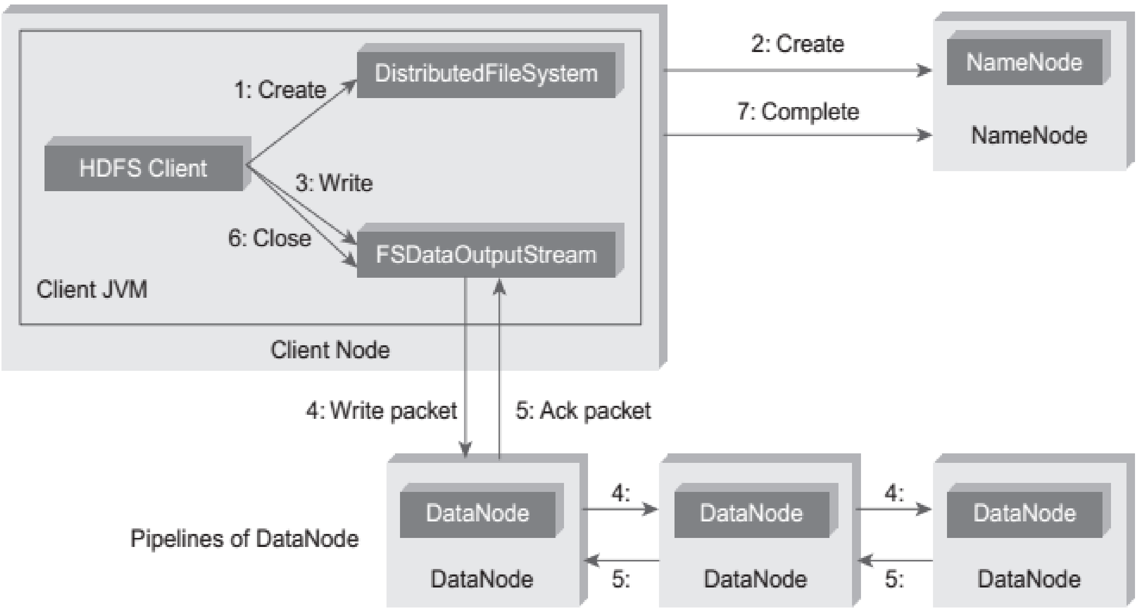
- Client Application interacts with NameNode for metadata related activities and communicates with DataNodes to read and write files.
- **Datanodes** are the slave nodes that divides the input files of varied formats into blocks and store the actual data. DataNodes converse with each other for pipeline reads and writes.
- **NameNode** is the master node that manages the File System Namespace, controlling the client's access to file-related operations such as read, write, create, delete and naming files and directories. NameNode starts up and reads FsImage and EditLog.
- If the NameNode has not restarted for months, the **Secondary NameNode** applies edits log on FSImage at regular intervals.
- **Hadoop 1.x** can configure to **64 MB** while **Hadoop 2.x** and **Hadoop 3.x** cluster can have 64MB/ **128MB (Default)** / 256MB/ 512 MB. Hadoop Administrator have control over block size to be configured for Cluster.



1.7 Anatomy of File Read



1.8 Anatomy of File Write


2: Create
NameNode
7: Complete
NameNode
1: Create
DistributedFileSystem
HDFS Client
3: Write
6: Close
FSDataOutputStream
Client JVM
Client Node
4: Write packet
5: Ack packet
Pipelines of DataNode
DataNode
DataNode
4:
5:
DataNode
DataNode
4:
5:
DataNode
DataNode