Practical – 3

Aim: Installation of Hadoop Single node cluster on Ubuntu.

Theory:

Hadoop:

- Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications.
- Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

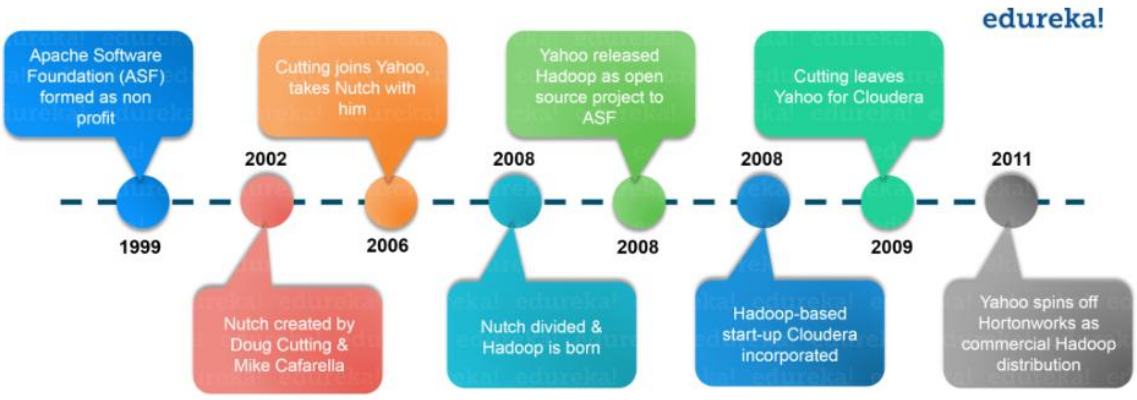


Figure 2.1 Evaluation of Hadoop

- The core components in the first iteration of Hadoop were MapReduce, the Hadoop Distributed File System (HDFS) and Hadoop Common, a set of shared utilities and libraries. It allows us to dump any kind of data across the clusters.

HDFS (Hadoop Distributed File System):

- HDFS creates an abstraction, let me simplify it for you. Similar as virtualization, you can see HDFS logically as a single unit for storing Big Data, but actually you are storing your data across multiple nodes in a distributed fashion. HDFS follows master-slave architecture.
- In HDFS, Namenode is the master node and Datanodes are the slaves. Namenode contains the metadata about the data stored in Data nodes, such as which data block is stored in which data node, where are the replications of the data block kept etc. The actual data is stored in Data Nodes.
- We actually replicate the data blocks present in Data Nodes, and the default replication factor is 3. Since we are using commodity hardware and we know the failure rate of these hardwares are pretty high, so if one of the DataNodes fails, HDFS will still have the copy of those lost data blocks.

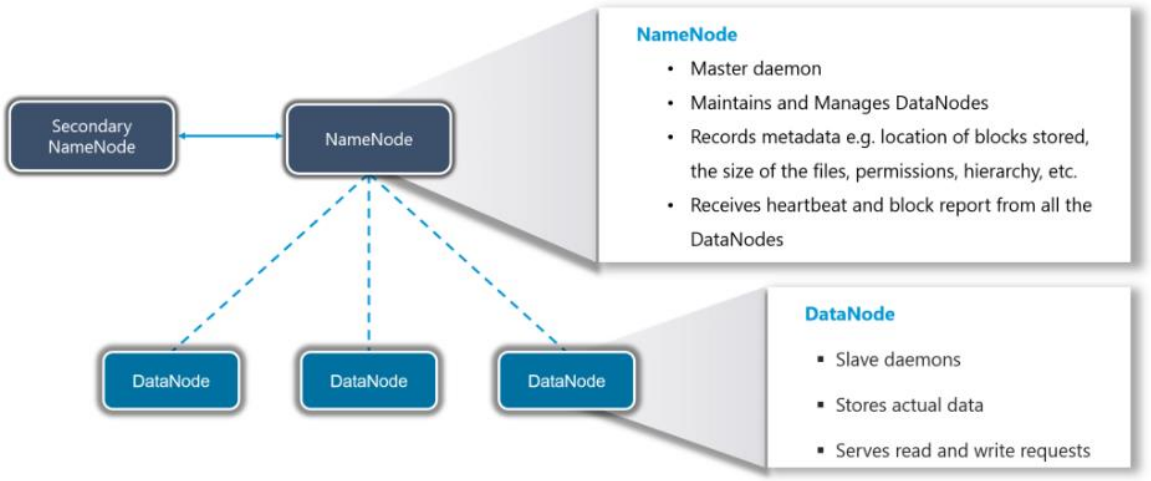
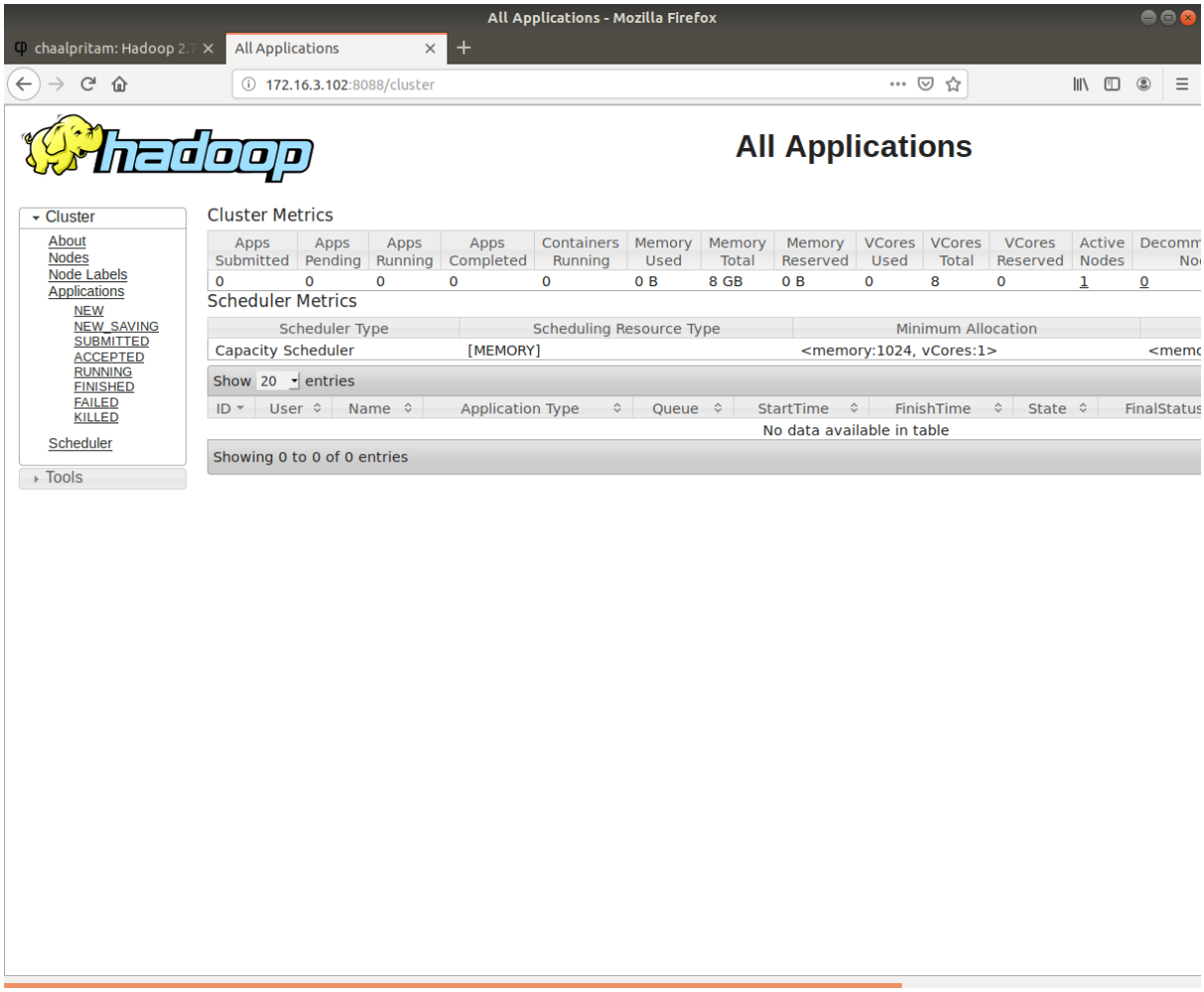


Figure 2.2 Hadoop – HDFS details

Output:

```
student@hadoopsn: ~
File  Edit  View  Search  Terminal  Help
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.
util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.
7.0.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.
security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflect
ive access operations
WARNING: All illegal access operations will be denied in a future release
19/07/25 16:18:52 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-student-resourc
emanager-hadoopsn.out
student@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-student-
nodemanager-hadoopsn.out
student@hadoopsn:~$ jps
3440 NameNode
3699 SecondaryNameNode
3876 ResourceManager
4233 NodeManager
4348 Jps
student@hadoopsn:~$
```



Conclusion:

We successfully performed and installation of Hadoop single node cluster on Ubuntu as well as we studied basics of HDFS and nodes like Data Node and Name Node and functionalities of those components.