Roll No: 33231

ASSIGNMENT 1

Title : Study of Hadoop Installation

Problem Statement :

1) Study of Hadoop installation on single Node
2) Study of Hadoop installation on multiple nodes

Objectives :

1) To understand installation and configuration of Hadoop on single node.
2) To understand installation and configuration of Hadoop on multiple nodes

Theory :

What is Hadoop ?

- Hadoop is an open source, Java-based programming framework which supports the processing and storage of extremely large sets of data in a distributed computing environment using simple programming models.
- Computer scientists Doug Cutting and Mike Cafferella created Hadoop in 2006 to support distribution for the Nutch (search engine). In 2008, Yahoo released Hadoop as an open-source project. Now, Apache Software Foundation (ASF) manages the Hadoop's framework and ecosystem of technologies
- Hadoop has very strong processing power and the

ability to handle virtually unlimited parallel tasks.

- Hadoop has quickly emerged as a foundation for big data processing tasks like scientific analytics of data, planning of business and sales and processing enormous volumes of data including social media data.

Hadoop Architecture :

<table>
<tr><td colspan="2">Hadoop 2.0</td></tr>
<tr><td>Mapreduce
(data processing)</td><td>other
(Data processing)</td></tr>
<tr><td colspan="2">YARN
(Cluster resource management)</td></tr>
<tr><td colspan="2">HDFS
(redundant, reliable storage)</td></tr>
</table>

1) HDFS : stands for "Hadoop distributed File System". It states that the file will be broken into blocks and stored in nodes over the distributed architecture. It provides high-throughput access to application data.

2) YARN : stands for "Yet Another Resource Negotiator". It is used for job schedulling and managing the cluster.

3) Map Reduce : This is YARN-based system for parallel

processing of large data set using key value pair. The Map task takes input data and converts it into a data set which can be computed in key-value pair.

4) Hadoop Common : These Java libraries and utilities are used to start Hadoop. They are used by other Hadoop modules. These ~~are used by other~~ libraries provide file system and OS level abstractions.

Hadoop Projects :

- With the help of Hadoop, application can run on systems with thousands of commodity hardware nodes. It can handle thousands of terabytes of data. Hadoop has distributed file system which facilitates rapid data tranfer rates among nodes. This allows the system to proceed even in case one or more nodes get failed. This approach avoids unexpected data loss.
- In 2006, scientists Dough Cutting and Mike Caferella created Hadoop to support dictribution for the Nutch. The main aim is to increase the speed of search results by the distribution of data and implement calculations on different computers by multitasking.

- Later on cutting joined Yahoo but he still worked on the Nutch project with ideas based on Google's early work with automating distributed storage and processing
- The Nutch project is divided into 2 parts -

1) Nutch - Web crawler portion
2) Hadoop - Distributed computing and processing portion.

- Yahoo released Hadoop ~~it~~ as an open-source project in 2008. Now, Apache software Foundation (ASF) manages Hadoop's framework and ecosystem of technologies.

Conclusion :

Thus we have successfully installed and tested single and multinode cluster.