Assignment 1

→ **Problem Statement**

1. Study of Hadoop Installation on Single Node
2. Study of Hadoop Installation on Multiple Nodes.

→ **Theory** :

A] **Hadoop** :

1. Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.
2. All modules in Hadoop are designed in such a way that hardware failures are common & should be automatically handled by the framework.
3. It consists of a storage part called Hadoop Distributed File System (HDFS) & a processing part called MapReduce. It splits files into large blocks & distributes them across nodes in a cluster.
4. To process data, it transfers packaged code for nodes to process in parallel based on the data that needs to be processed.
5. The framework has following modules:

a. Hadoop Common : contains all libraries & utilities

b. HDFS : stores data on commodity system

c. Hadoop Yarn : Manages computing resources along with scheduling user's application.

d. Hadoop MapReduce : implementation of MapReduce function for large scale data processing

B] Installation on Single Node :

1. Apache Hadoop Installation

a. Open Terminal
b. Install JDK
c. Add a separate hadoop user
d. Sign in through new user.
e. Install SSH
f. Create & Setup SSH Certificates.
g. Download Hadoop
h. Set the Configuration Files.

2. Cloudera Hadoop Installion

a. Install VirtualBox.
b. Download Cloudera VM
c. Start VirtualBox
d. Import Cloudera & follow on screen instructions.
e. Make neccessary changes for the VM.
f. Lauch Cloudera VM
g. It will take several minutes to boot

→ Installation on Multiple Nodes

a. Install Multi Node Hadoop Cluster

b. Install Java on Master & Slaves.

c. Disable IPv6

d. Set up a Hadoop User

e. Login as Hadoop User

f. Download & Install Hadoop binaries on Master & Slave Nodes.

g. Setup Hadoop Environment on Master & Slave Node.

h. Update Configuration files.

→ Conclusion :

Topics Covered :

1. Hadoop & its modules.
2. Installation of Hadoop on Single and Multipl
Nodes.