Assignment: 1

AIM: Study of Hadoop Installation.

PROBLEM STATEMENT / DEFINITION:
1. Study of Hadoop Installation on Single Node.
2. Study of Hadoop Installation on Multiple Nodes.

OBJECTIVE:
I. To understand Installation & Configuration of Hadoop on single Node.
II. To understand Installation & Configuration of Hadoop on multiple nodes.

THEORY:

a) THE Single Node:

Introduction

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. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality— nodes manipulating the data they have access to— to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking. The base Apache Hadoop framework is composed of the following modules:

Hadoop Common – contains libraries and utilities needed by other Hadoop modules.

Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.

Hadoop YARN – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users & applications.

Hadoop MapReduce – an implementation of the MapReduce programming model for large scale data processing.

A.1) Apache Hadoop Installation

Steps to Install Apache Hadoop 2.7.0 Single Node Cluster on Ubuntu 16.04 are as follows:

1. Move hadoop.tar file in home directory and right click on hadoop.tar file → extarct here

2. Open terminal in Ubuntu and type following commands

sudo apt-get update

#Installation of OpenSSH Server

sudo apt-get install openssh-server

sudo cp /etc/ssh/sshd_config /etc/ssh/sshd_config.factory-defaults

#Open sshd_config in a text editor by running the command below for Ubuntu versions

sudo gedit /etc/ssh/sshd_config

Changes to be made in configuration file

1.Disable password authentication by changing this line in the configuration file #PasswordAuthentication yes to PasswordAuthentication no.

2.Add the following Lines to the End (Check username for eg. Slii here)

AllowUsers slii
PermitRootLogin no
PubkeyAuthentication yes

3.Change LogLevel INFO to LogLevel VERBOSE.

4.Save and Restart SSH

sudo systemctl restart ssh

ssh-keygen -t rsa -P ""

#Testing SSH

ssh localhost

sudo apt-get install eclipse eclipse-cdt

Step 2: Downloading and Installing Hadoop

sudo mv hadoop /usr/local/

sudo chown -R slii:slii /usr/local/hadoop

sudo apt-get install vim

vim .bashrc

Enter insert key on keyboard
Press ‘enter’ after ‘fi’ to shift to next line
Type the following at the end of file:

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop/hadoop-2.7.0
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib/native"

press ‘escape’ to shift to next line
press shift;wq and then press ‘enter’
Save the file and exit

Save the Changes to Bashrc File

source ~/.bashrc

sudo mkdir -p /app/hadoop/tmp
sudo chown slii:slii /app/hadoop/tmp

sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode
sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode
sudo chown -R slii:slii /usr/local/hadoop_store

hdfs namenode -format

#Start the single node cluster

start-dfs.sh
start-yarn.sh
jps