Assignment NO - 11

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Title - Big data Analytics 2.

Design a distributed application using map-reduce which processes a long File of a System.

objective - Students should be able to processes a log file of a system using the Hadoop map-reduce framework.

Prerequisite - 1. Basic of Java programming

Theory - Steps to install Hadoop for distributed enviroment.

Java code for processes a log file a system

Steps to install hadoop for distributed enviroment

Intially create one folder store input file (access-log-short-csv), Sales Country Reducer java, Sales Country Driver.java files)

Step 1) Go to hadoop home directory and format the NameNode.

cd hadoop-2.7.3.

bin/ hadoop namenode -format.

Step 2) once the NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and Tracks all the files stored across cluster.

/hadoop-daemon.sh start namenode.

Q) Start Data Node. Resource manager

Resource manager is the master that attributes all the avilable cluster resource and Thus. helps in managing the distributed applns. running on the YANN System. Its work is to manage each Node Managers and the each applications ~~running~~ each Node Managers and the each application's Appln Master.

3) Start Data Node -

on startup a DataNode connects to the namenode and it responds to the requests from the Namenode for different operations.

4) Start Node Manager.

The Node manager in each machine framework is the agent which is responsible for managing container, monitering the resource usage and reporting the same to the resource manager.

5) Start JobHistory Server -

JobHistory Server is responsible for servicing all job history related request from client | ~~mr~~ yarn - daemon.

Sh start nodemanager.

step 3. To cheak that all the Hadoop services are up and running, run the below command.

JPS.

step 4 - cd

step 5 - sudo mkdir mapreduce - Vijay

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Step 6 - sudo chmod 777 - R mapreduce_vijay/

Step 7 - sudo chown - R vijay mapreduce_vijay/

Step 8 - sudo cp/home/vijay/Desktop/logfiles/*.* /
mapreduce_vijay/

Step 9 - cd mapreduce_vijay/

Step 10 - ls

Step 11 - sudo chmod +r **.

Step 12 - export.

Step 13 - Javac -d - SalesMapper.java SalesCountry
Reducer.java SalesCountryDriver.java

Step 14 - ls.

Step 15 - cd SalesCountry/

Step 16 - ls (cheak is class files are created)

Step 17 - cd - -

Step 18 - gedit Manifest.txt

Step 19 - Jar - cfm mapreduce_vijay.jar Manifest.txt
SalesCountry/*.class

Step 20 - ls

Step 21 - cd

Step 22 - cd mapreduce_vijay/

Step 23 - Sudo mkdir /input 200.

Step 24 - Sudo cp access_log_Short.csv /input 200

Step 25. SHADOOP_HOME/bin/hdfs dfs -put /input
200/

Step 26. SHADOOP_HOME/bin/hadoop jar mapreduce_
vijay.jar /input 200/output 200

Step 27 - hadoop fs - ls /output 200.

Step 28 - hadoop fs - cat /out 231 / part 0000.

Step 29 - Now open the Mozilla browser and go to localhost : 50070/dfshealth.html to cheak the Namenode interface.

conclusion - Hence we study the Design a distributed appln using Map-reduce which processes a log file of a System.

TS	PR	UC	VA	RN	Total marks	sign
02	02	01	02	02	09	

Assignment No - 12.

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Title - A simple program in SCALA using Apache Spark Framework.

objective - write a simple program in SCALA using Apache spark framework.

Theory

- steps to install scala
- Apache Spark Framework Installation
- Source code

1. Install scala

Step 1) Java - version

Step 2) Install Scala from the apt repository by running the following commands to search for scala and install it.

Sudo apt search scala => Search for the package Sudo apt install scala => install the package.

Step 3) - To verify the installation of scala, run the following command, scala -version.

2. Apache Spark Framework Installation.

Apache Spark is an open-source, distributed processing system used for big data workloads It utilizes in memory caching, and optimized query execution for fast analytics queries

against data of any size.

Step 1: Now go to the official Apache spark download page and grab the latest version (i.e 3.2.1) at the time of writing this article. Alternatively, you can use the wget command to download the file directly in the terminal.

wget
https://apachemirror.wuchna.com/spark/spark-3.2.1/spark-3.2.1-bin-hadoop2.7.tgz.

Step 2) Extract the Apache Spark tar file. tar -xvzf spark-3.1.1-bin-hadoop2.7.tgz

Step 3) Move the extracted spark directory to/opt directory. Sudo mv.
Spark-3.1.1-bin-hadoop2.7/opt/spark

Configure Enviromental variables for Spark

Step 4: Now you have to set a few enviromental variables in .profile file before starting up the Spark.
echo "export SPARK_HOME=/opt/spark" >> ~/.profile.
echo "export
PATH=$PATH:/opt/spark/bin:/opt/spark/sbin" >> ~/.profile. echo "export

against data of any size.

Step 1: Now go to the official Apache spark download page and grab the latest version (i.e. 3.2.1) at the time of writing this article. Alternatively, you can use the wget command to download the file directly in the terminal.

wget

https://apachemirror.wuchna.com/spark/spark-3.2.1/spark-3.2.1-bin-hadoop2.7.tgz.

Step 2) Extract the Apache Spark tar file. tar -xvzf spark-3.1.1-bin-hadoop2.7.tgz

Step 3) Move the extracted spark directory to /opt directory. sudo mv.

Spark-3.1.1-bin-hadoop2.7 /opt/spark

Configure Enviromental variables for Spark

Step 4: Now you have to set a few enviromental variables in .profile file before starting up the Spark.

echo "export SPARK_HOME=/opt/spark" >> ~/.profile.

echo "export

PATH=$PATH:/opt/spark/bin:/opt/spark/sbin" >> ~/.profile echo "export

Page No.
Date

~/ profile.

Step 5 :- To make sure that these new enviroment variables are reachable within the shell and avilable to apache spark, it is also madatory to run the following command to take recent changes into effect.

Source ~/ profile.

Step 6 - ls -l /opt/spark.

Start Apache ~~browser~~ spark in Ubantu

Step 7) Run the following command to start the Spark master service and Slave service

Start - master.sh.

Start - workers.sh spark://localhost:7077.

(if workers not starting then remove and install openssh:

Sudo apt-get remove openssh-client openssh-Server

Sudo apt-get install openssh-client openssh-Server.

Step 8) once the service is started go to the browser and type the following URL Access spark page. from the page, you can see my master and slave service is started.

http://localhost:8080/

... analytics queries

step 9 - you can also cheak if spark shell works fine by launching the spark - shell command. Spark shell.

Sudo apt install snapd.

Snap find "intellij"

Sudo Snap install intellij-idea-community --classic

start intellij IDE community Edition.

conclusion - Hence we study have simple program using SCALA Apache ~~spark~~ framework.

~~PR~~ TS (2)	PR (2)	UC (2)	VA (2)	~~RN~~ (2)	Total marks (10)	Sign
02	02	01	02	01	09	