Lab Assignment B2

Title:

Locate dataset (eg sample-weather txt) for working on weather data which reads the text input files and finds average for temperature, dew point, wind speed

Objective of the assignment:

Students should be able to install and Hadoop MapReduce framework on local-standalone set-up and they should able to write a code in Java for working on weather data which reads the text input files and finds average for temperature, dew points and wind speed

Prerequisite:

(1) Java - Java JDK (installed)

(2) Hadoop - Hadoop package

Theory:

1) Hadoop MapReduce framework

1) Install Java

2) Install Hadoop

3) Configure Hadoop

4) Test Hadoop Installation

5) Create MapReduce program

6) Input file to mapreduce

7) Display the output

A Set up Hadoop:

Install and configure Hadoop on your system or a Hadoop cluster.

B. Prepare the dataset:

Ensure your weather dataset (eg. "sample_weathe available and accessible to Hadoop. You may need to u the dataset to the HDFS or make it available thr other means.

C. Write a MapReduce program: Create a Java progr that implements the MapReduce paradigm to process weather data and calculate the average valu

D. Map function:

In the map function, you will parse each inp record from the dataset and extract the temperature, dew point and wind speed values Emit key-value pairs with the key set to constant value. and the values set to the ex temperature, dew point and wind speed.

E. Reduce function:

In the reduce function, you will receive the key pair emitted by the map function. Iterat over the values and calculate the sums of temperature dew points and wind speed.

F. Output:
Emit a single key-value pair with the key set to a constant value and the values set to a string representation of the calculated averages.

G. Submit the mapreduce job:
Use the Hadoop command line interface (CLI) or a job submission framework to submit the MapReduce job to the Hadoop cluster.

H. Retrieve the result:
Once the MapReduce job completes, you can retrieve the output files containing the calculated averages from the Hadoop cluster.

> How to Install single node cluster Hadoop on Windows?

Step1: Verify the Java installed
Step2: Installing Hadoop
Step3: Hadoop Configuration.
Step4: Testing Hadoop Installations
Step5: Create a program for working on weather data

3) <u>Hadoop - Running MapReduce Example:</u>

Step 1: Store the dataset file, such as "sample-w
·txt".

Store the dataset file in HDFS using hadoop fs
command.

eg:

```
hadoop fs -put /path/to/sample weather.txt /i
```

Step 2: Write a MapReduce program in Java.

Write a mapreduce program in Java to read and
the data in the 'sample-weathen txt' file.

eg: (i) WeatherDataMapper.java

(ii) WeatherDataReducen java

Step 3: Write a MapReduce program in Java for Hado
configuration

Create a Hadoop job configuration and specify the
and output paths as well as mapper and reducer
classes.

Step 4: Compile the java code and package into

Step 5: Run Hadoop job using the following comma

```
hadoop jar/ path/to /WeatherDataAnalyzer.jar.
```

Step 6: Output

View the output using the commands:

```
hadoop fs -cat /output/part-r-00000
```

This will display the average temperature, dew points, and wind speed values in the console.

Conclusion:

The java code for weather dataset using Hadoop MapReduce function was implemented.

Assignment B3

Title: Write a simple program in SCALA using Apache spark framework

Objectives of the assignment:

Students Should be able to write a Simple program in SCALA using Apache spark framework.

Prerequisites:
1) Basic knowledge of Scala
2) Basic knowledge of Java Syntax
3) Installation of Java.
4) Operating System recommended: 64-bit opensource Linux / Windows.

Theory:
1) Scala:
Scala is an acronym for 'scalable language'. It is a general purpose programming language designmed for the programmers who want to write program in concise, elegant way Scala is object oriented and functional programming language Scala enables programmers to be more productive Scala is a compiler based language

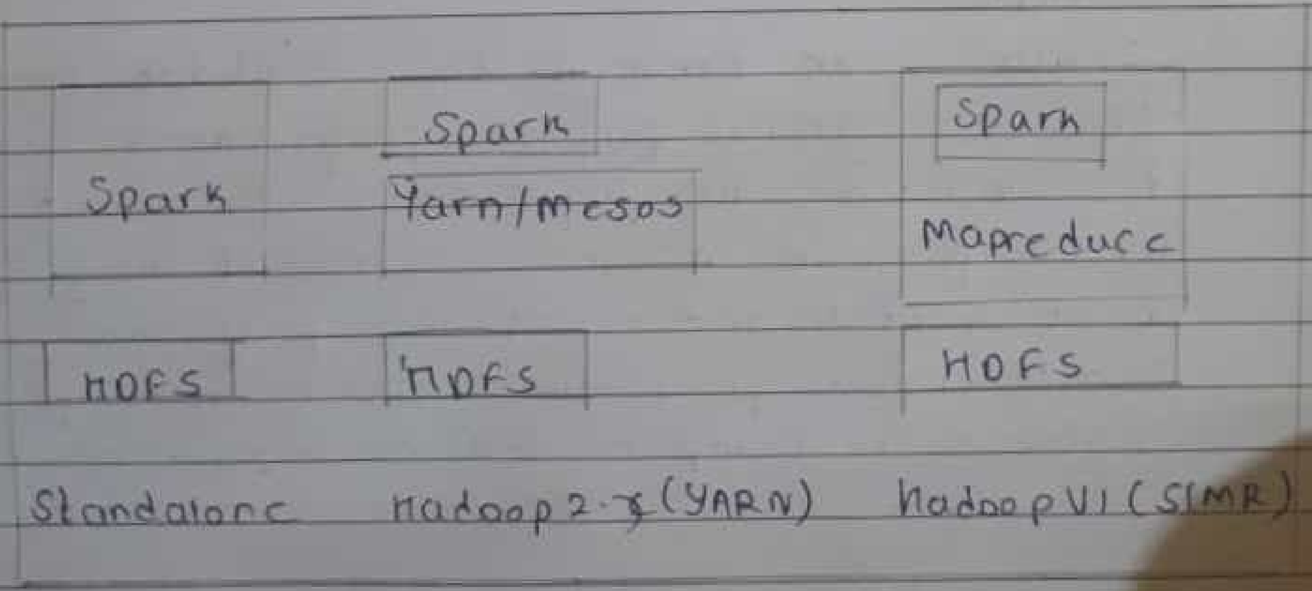
2) Apache spark
Apache spark is an open source data processing framework for performing big data analytics on

distributed computing cluster.

Spark was nitially Started by Matei Zaharia at UC Berkeley's. Apache Spark has other features:

1) Supports wide variety of operations compared to map and Reduce functions.
2) Provides concise and consistent API's in scala Java and Python.
3) Spark is written in Scala Programming Language and runs in JVM.
4) Features interactive shell for scala and Python
5) It leverages the distributed cluster memory for doing computations for increased speed and data processing.

Spark buit on Hadoop:



Steps to nstall Scala and Apache spark framework on Windows

Step 1: Java Installation
Use following command to verify the scala version
JAVA -version.

Step 2: Scala Installation
Use following command to verify the scala installation
Scala -version

Step 3: Apache spark download and install

Step 4: Configuring window environment for Apache Sparks.

Step 5: Download and Install Scala IDE.

Step6: Test the envionment.

Step 7: Choose a development environment

Step 8: Run your first Scala program in shell

Step 9: Write and run a program in scala using an edition.

Step 10: compile a Scala program

Conclusion:

In this way ,we have written and implemented a simple program in scala using Apache Spark Framework.