DS&BDL
Assignment 13

Title: Locate dataset (e.g., sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

Aim:

Write a code in JAVA for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

Objective:

By completing this task, students will learn the following

1. Hadoop Distributed File System.
2. MapReduce Framework.

Software/Hardware Requirements: 64-bit Open source OS-Linux, Java, Hadoop.

Theory:

Here, we will write a Map-Reduce program for analyzing weather datasets to understand its data processing programming model. Weather sensors are collecting weather information across the globe in a large volume of log data. This weather data is semi-structured and record-oriented. This data is stored in a line-oriented ASCII format, where each row represents a single record. Each row has lots of fields like longitude, latitude, daily max-min temperature, daily average temperature, etc. for easiness, we will focus on the main element, i.e. temperature. We will use the data from the National Centres for Environmental Information(NCEI). It has a massive amount of historical weather data that we can use for our data analysis.

1. Input Data

We can download the dataset for various cities in different years. choose the year of your choice and select any one of the data text-file for analyzing. We can get information about data from *README.txt* file available on the NCEI website.

2. Make a project in Eclipse with below steps:

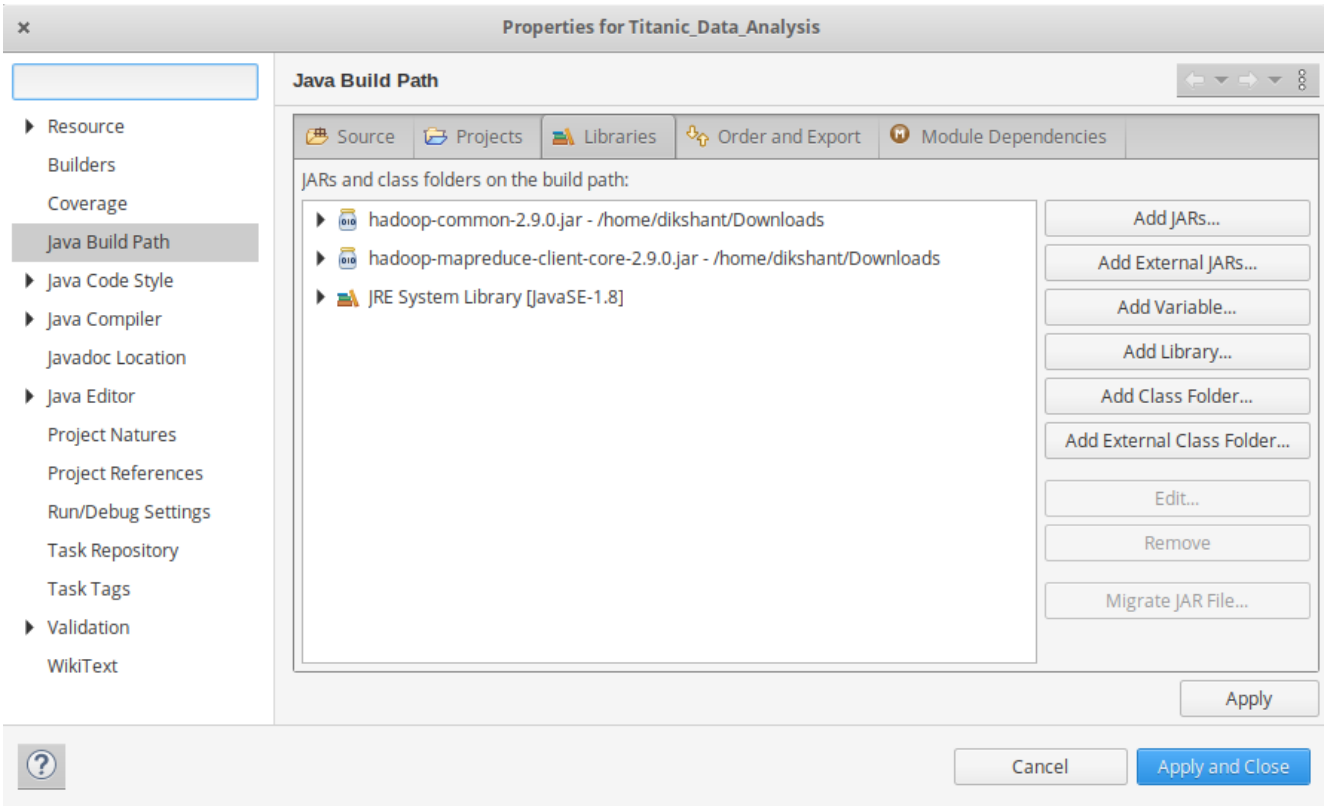
First Open Eclipse -> then select File -> New -> Java Project ->Name it MyProject -> then select use an execution environment -> choose JavaSE-1.8 then next -> Finish.

In this Project Create Java class with name **MyMaxMin** -> then click **Finish**

Write code for mapper and reducer in MyMaxMin class.

Now we need to add external jar for the packages that we have import. Download the jar package Hadoop Common and Hadoop MapReduce Core according to your Hadoop version.

3 . Now we add these external jars to our MyProject. Right Click on MyProject -> then select Build Path-> Click on Configure Build Path and select Add External jars…. and add jars from it's download location then click -> Apply and Close.



4 . Now export the project as jar file. Right-click on MyProject choose Export.. and go to Java -> JAR file click -> Next and choose your export destination then click -> Next.
choose Main Class as MyMaxMin by clicking -> Browse and then click -> Finish -> Ok.

5. Start our Hadoop Daemons

start-dfs.sh

start-yarn.sh

6. Move your dataset to the Hadoop HDFS.

```
hdfs dfs -put /file_path /destination
```

7. Now Run your Jar File with below command and produce the output in **MyOutput** File.

```
hadoop jar /jar_file_location /dataset_location_in_HDFS /output-file_name
```

8. Now Move to *localhost:50070/*, under utilities select *Browse the file system* and download **part-r-00000** in **/MyOutput** directory to see result.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	dikshant	supergroup	38.78 KB	Jul 04 09:39	1	122.07 MB	CRND0103-2020-AK_Fairbanks_11_NE.txt
drwxrwxr-x+	dikshant	supergroup	0 B	Jun 23 14:23	0	0 B	Hadoop_File
drwxr-xr-x	dikshant	supergroup	0 B	Jul 04 09:44	0	0 B	MyOutput
drwxrwxrwx	dikshant	supergroup	0 B	Jun 14 21:43	0	0 B	tmp
drwxr-xr-x	dikshant	supergroup	0 B	Jun 14 21:43	0	0 B	user

/MyOutput Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	dikshant	supergroup	0 B	Jul 04 09:44	1	122.07 MB	_SUCCESS
-rw-r--r--	dikshant	supergroup	3.85 KB	Jul 04 09:44	1	122.07 MB	part-r-00000

9. See the result in the Downloaded File.

For Example, **20200101** means year = 2020

month = 01

Date = 01

Conclusion:

In this assignment, we have learned how to locate dataset in Hadoop through jar files.