Exercise 05: Weather data analytics using MapReduce

This exercise's MapReduce process is doing Weather analysis

Prerequisites

Ensure that Hadoop is installed, configured and is running. More details: Single Node Setup for first-time users.

Cluster Setup for large, distributed clusters.

Inputs and Outputs

i. **Input file should be in : /weather/in00/**

WAData.txt

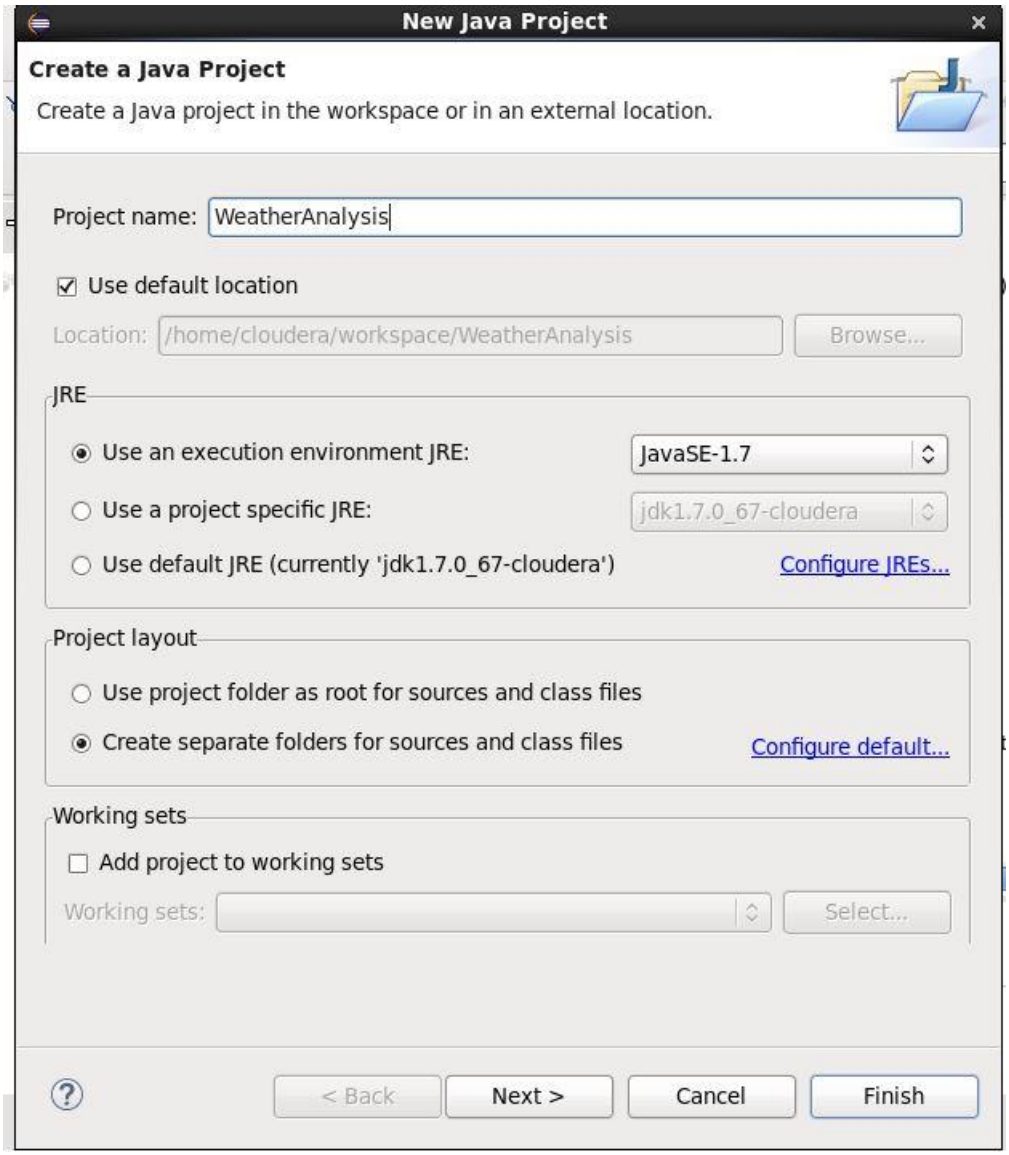
Copy the content text from sample_weather.txt, Which is attached in Google classroom.

ii. **Output file should be in /weather/out00/**

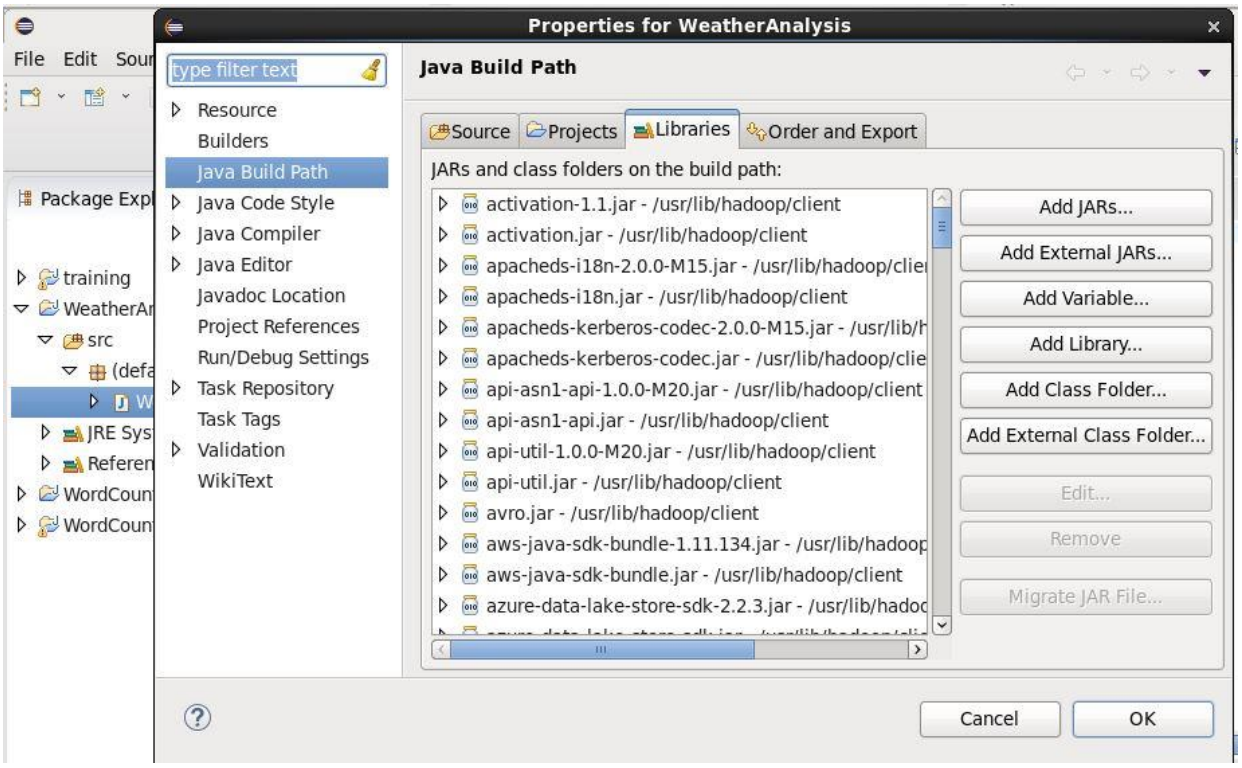
Step 1

Compile `WeatherAnalysis.java` and create a WeatherAnalysis.jar:

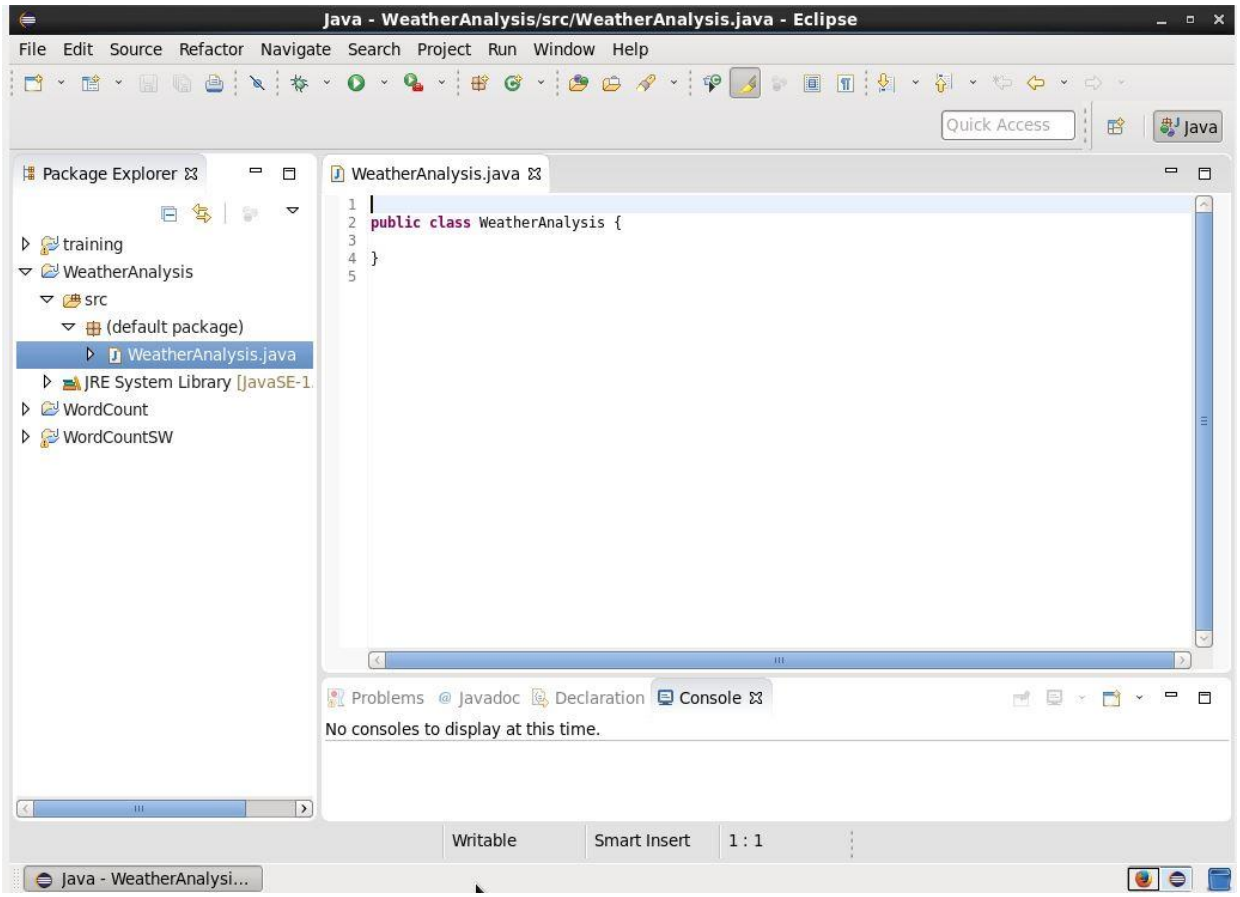
(i) Create WeatherAnalysis.java project.



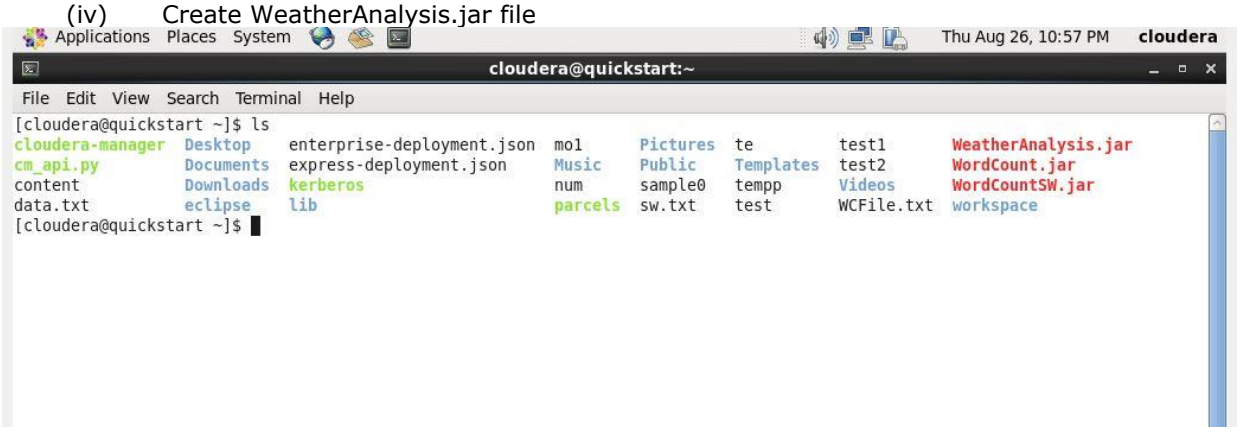
(ii) Import external .jar files



(iii) Create Weathernalysis class file using Google classroom attached WeatherAnalysis.java file.



(iv) Create WeatherAnalysis.jar file



Step 2

Create following folders in HDFS:

- `/weather/in00` - input directory in HDFS
- `/weather/out00` - output directory in HDFS

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /weather
[cloudera@quickstart ~]$ hdfs dfs -mkdir /weather/in00
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 15 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - cloudera supergroup          0 2021-08-19 07:07 /data
drwxr-xr-x   - hbase    supergroup          0 2021-08-15 07:21 /hbase
drwxr-xr-x   - cloudera supergroup          0 2021-08-25 00:37 /in00
drwxr-xr-x   - cloudera supergroup          0 2021-08-25 00:34 /in000
drwxr-xr-x   - cloudera supergroup          0 2021-08-18 04:55 /out00
-rw-r--r--   1 cloudera supergroup          0 2021-08-16 23:17 /sample00
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxr-xr-x   - cloudera supergroup          0 2021-08-19 06:17 /temp
drwx------   - cloudera supergroup          0 2021-08-25 00:56 /tmp
drwxr-xr-x   - cloudera supergroup          0 2021-08-19 06:07 /tweet
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
drwxr-xr-x   - cloudera supergroup          0 2021-08-25 05:34 /wcsw
drwxr-xr-x   - cloudera supergroup          0 2021-08-26 23:08 /weather
[cloudera@quickstart ~]$ hdfs dfs -ls /weather
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2021-08-26 23:08 /weather/in00
[cloudera@quickstart ~]$
```

Step 3

Create and copy WAData.txt-files into input folder:

```
[cloudera@quickstart ~]$ cat >WAData.txt
^?690190 13910  20060201_0  51.75     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_1  54.74     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_2  50.59     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_3  51.67     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_4  65.67     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_5  55.37     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_6  49.26     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_7  55.44     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_8  64.05     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_9  68.77     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_10  48.93     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_11  65.37     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_12  69.45     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_13  52.91     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_14  53.69     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060201_15  53.30     33.0 24  1006.3 24   943.9 24   15.0 24   10.7 24   22.0   28.9    0.00I 999.9  000000
690190 13910  20060204_22  50.37     15.6 24  1020.0 24   955.9 24   15.0 24    5.1 24   11.1
690190 13910  20060204_23  51.75     15.6 24  1020.0 24   955.9 24   15.0 24    5.1 24   11.1
^C
[cloudera@quickstart ~]$ hdfs dfs -put WAData.txt /weather/in00
[cloudera@quickstart ~]$ hdfs dfs -ls /weather/in00
Found 1 items
-rw-r--r--   1 cloudera supergroup      12055 2021-08-26 23:11 /weather/in00/WAData.txt
[cloudera@quickstart ~]$
```

Java - WeatherAnalysi... | cloudera@quickstart:~

[cloudera@quickstart ~]$ hdfs dfs -ls

/weather/in00/ Found 1 items

-rw-r--r--   1 cloudera supergroup 12054 2021-08-26 15:48 /weather/in00/WAData.txt

Step 4

Run the MapReduce application :

[cloudera@quickstart ~]$ hadoop jar WeatherAnalysis.jar WeatherAnalysis /weather/in00/WAData.txt /weather/out00/

Show MapReduce Framework

```
Map-Reduce Framework
        Map input records=96
        Map output records=95
        Map output bytes=2962
        Map output materialized bytes=3158
        Input split bytes=120
        Combine input records=0
        Combine output records=0
        Reduce input groups=4
        Reduce shuffle bytes=3158
        Reduce input records=95
        Reduce output records=4
        Spilled Records=190
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=264
        CPU time spent (ms)=2410
        Physical memory (bytes) snapshot=362201088
        Virtual memory (bytes) snapshot=3016896512
        Total committed heap usage (bytes)=226365440
```

Step 5

Output:

[cloudera@quickstart ~]$ hdfs dfs -ls

/weather/out00/ Found 2 items

-rw-r--r-- 1 cloudera supergroup 0 2021-08-26 15:50 /weather/out00/_SUCCESS

-rw-r--r-- 1 cloudera supergroup 228 2021-08-26 15:50 /weather/out00/part-r-

00000 [cloudera@quickstart ~]$ hdfs dfs -cat /weather/out00/part-r-00000