DISTRIBUTED COMPUTING: HADOOP PROGRAMMING

Lab Work
Hadoop as a MapReduce Platform and Java Classes as a Programming Framework

February 3rd, 2018

Jeremy Williams

Problem Statement

The goal of this lab work is to find out the average temperature in cloudera environment using python MapReduce based on the weather dataset provided. First python mapper is executed to collect the reduced dataset of year and temperate of that year. Once we have mapper results in hand, we invoke the reducer python to collect the temperature set against a year and then to sum of the temperature and then divided by count to get the average temperature of the year from the various temperature values of the year.

Approach to Solution

The source code is written in python and pure python map-reduce techniques is used for the results. I used a Linux-shell to initially run the mapper class then feed the mapper output to the reducer. The code has business logics that uses pure python map-reduce techniques.

Solution Description

As the code is being run from Linux-shell so, the pipeline feeding of output is available.
The Execution logic is to run the mapper first to get the reduced dataset to work on.
Then find out and filter out only the valid data which are passing the validation criteria.
Once the valid and filter dataset is available, the dataset will consist of **year** and **temperature** tuples.

Now the reduction mechanism is applied on the output of the above mapper output through 'reduceByKey' mechanism. This basically calculates the sum of the temperatures of a year first then divide by number of entries.

Data File

weather.txt

Sample data

0029029070999991901010106004+64333+023450FM-
12+000599999V0202701N015919999999N0000001N9-
00781+99999102001ADDGF108991999999999999999999

0029029070999991901010113004+64333+023450FM-
12+000599999V0202901N008219999999N0000001N9-
00721+99999102001ADDGF104991999999999999999999

0029029070999991901010120004+64333+023450FM-
12+000599999V0209991C000019999999N0000001N9-
00941+99999102001ADDGF108991999999999999999999

0029029070999991901010206004+64333+023450FM-
12+000599999V0201801N008219999999N0000001N9-
00611+99999101831ADDGF108991999999999999999999

0029029070999991901010213004+64333+023450FM-
12+000599999V0201801N009819999999N0000001N9-
00561+99999101761ADDGF108991999999999999999999

Results/Output

-using Python

```
[cloudera@quickstart AverageTemp1]$ hdfs dfs -ls /hduser/output18
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2018-01-21 00:46 /hduser/output18/_SUCCESS
-rw-r--r--   1 cloudera supergroup         19 2018-01-21 00:46 /hduser/output18/part-00000

[cloudera@quickstart AverageTemp1]$ hdfs dfs -cat /hduser/output18/part-00000

1901    46.6985070079
```

-using JAVA

```
[cloudera@quickstart AverageTemp1]$ hdfs dfs -ls /hduser/output17
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2018-01-21 00:32 /hduser/output17/_SUCCESS
-rw-r--r--   1 cloudera supergroup         18 2018-01-21 00:32 /hduser/output17/part-r-00000

[cloudera@quickstart AverageTemp1]$ hdfs dfs -cat /hduser/output17/part-r-00000

 average         46.69850
```