Homework 4: Hadoop and Spark

Large-Scale Data Analysis 2020

Stefano Pellegrini (mlq211)

June 13, 2020

1 Analyzing Airline Data (50 pts.)

1.1 Data Analysis with Hadoop:

(a) ***Total Departure Delay (5 pts):*** Create a directory airline data on the Hadoop cluster and copy all csv files from the local file system to this HDFS directory. Write a mapper and reducer that computes the total departure delay per airport in minutes (ignore all negative values for the delay, which indicate a departure before the scheduled time). What is the total departure delay for airport LAX?

I run the following bash command in my Virtual Machine:

```
# Set the environment
./.start_dfs_yarn.sh
source .venvs/lsda/bin/activate
source .activate_jupyter_pyspark
pyspark

# Combine the 2016 csv files in a single file.
cat 2016*.csv > air_all

# Upload the file to the airline_data directory in Hadoop
hadoop fs -mkdir airline_data
hadoop fs -put air_all airline_data

# Run the mapper and reducer to get the total departure delay
hrun mapper_1a.py reducer_1a.py airline_data/2016_all.csv air_all

# Copy the output to my local machine
hadoop fs -cat air_all/part-00000 > air_all

# Extract the total departure delay for LAX airport
grep LAX air_all
```

The total departure delay for the LAX airport is 2369088 minutes. Please find the code in `mapper_1a.py` and `reducer_1a.py` and the output in `air_all`.

(b) ***Maximal Departure Delay (5 pts):*** Write a mapper and reducer that computes the maximal departure delay per airport in minutes. What is the maximal departure delay for airport DTW?

I run the following bash command in my Virtual Machine:

```
# Run the mapper and reducer to get the max departure delay
hrun mapper_1b.py reducer_1b.py airline_data/2016_all.csv air_max

# Copy the output to my local machine
hadoop fs -cat air_max/part-00000 > air_cat

# Extract the max departure delay for DTW airport
grep DTW air_max
```

The maximum departure delay for DTW airport is 1216 minutes. Please find the code in `mapper_1b.py` and `reducer_1b.py` and the output in `air_max`.

(c) ***Mean and Standard Deviation for Departure Delays (10 pts):*** Write a mapper and reducer that compute the mean and the standard deviation for the departure delays for each airline ID. Consider all delay values including the negative ones (treat missing values as 0). A reducer should output, for each airline ID, the associated mean and standard deviation. What is the mean and standard deviation for airline ID 20366.

I run the following bash command in my Virtual Machine:

```
# Run the mapper and reducer to get the mean and standard deviation
hrun mapper_1c.py reducer_1c.py airline_data/2016_all.csv air_mean_sd

# Copy the output to my local machine
hadoop fs -cat air_mean_std/part-00000 > air_mean_sd

# Extract the mean and sd for delay for 20366 airline ID
grep 20366 air_mean_std
```

The mean and standard deviation for airline ID 20366 are respectively 7.021903 and 42.79789037023834. Please find the code in `mapper_1c.py` and `reducer_1c.py` and the output in `air_mean_sd`.

(d) ***Top-10 of Departure Delays (10 pts):*** Write a mapper and reducer that compute the 10 most delayed flights for each airline ID. The reducer should output, for each airline ID, a list containing the delays for the 10 most delayed flights w.r.t. the departure time. What are the 10 most delayed flights for airline ID 20366?

```
# Run the mapper and reducer to get the top10 delay per airport
hrun mapper_1d.py reducer_1d.py airline_data/2016_all.csv air_top10_delay

# Copy the output to my local machine
hadoop fs -cat top10_delay/part-00000 > air_top10_delay
```

The top 10 delays (in minutes) for each airport are:

```
19393   [587, 587, 628, 628, 651, 651, 669, 669, 779, 779]
19690   [940, 940, 959, 959, 1013, 1013, 1066, 1066, 1202, 1202]
19790   [1165, 1165, 1170, 1170, 1181, 1181, 1189, 1189, 1201, 1201]
19805   [1554, 1554, 1558, 1558, 1597, 1597, 1651, 1651, 1663, 1663]
19930   [649, 649, 653, 653, 905, 905, 906, 906, 994, 994]
19977   [1093, 1093, 1110, 1110, 1113, 1113, 1122, 1122, 1221, 1221]
20304   [1273, 1273, 1284, 1284, 1311, 1311, 1313, 1313, 1332, 1332]
20366   [1189, 1189, 1190, 1190, 1205, 1205, 1222, 1222, 1236, 1236]
20409   [651, 651, 655, 655, 735, 735, 790, 790, 814, 814]
20416   [584, 584, 590, 590, 600, 600, 681, 681, 887, 887]
20436   [641, 641, 646, 646, 666, 666, 704, 704, 725, 725]
21171   [368, 368, 384, 384, 392, 392, 401, 401, 404, 404]
```

Please find the code in `mapper_1d.py` and `reducer_1d.py` and the output in `air_top10_delay`.

1.2 Data Analysis with Spark (20 pts, 5 pts for each of the subtasks):

Implement the four jobs for the Airline dataset in Apache Spark!