University of Sydney

DATA3404

Data Science Platforms

Group Assignment

Team members:
450411920

Semester 1, 2019

Contents

Job Design Documentation

Task 1: Top 3 Cessna Models

We imported the csv files containing the relevant fields. We then joined the ontimeperformance_flights with ontimeperformance_aircrafts and projected the relevant columns. We then applied a filter function to only return aircrafts with the model equaled to "CESSNA". Finally, we applied a flatmap to count all the instances of the Cessna model, rank them in descending order, and return the top 3 Cessna model aircrafts.

Task 2 Average Departure Delay

We import the csv files with the relevant fields. We first filter the dataset according to

1. Contain only US Airlines;
2. Select the specified year by the user;
3. Removing cancelled flights.

Then, we sort the data by airline name in ascending order. We then join the datasets and project relevant fields. We then apply a combination of flatMaps and group by's and then aggregate the result using airline names to compute the cumulative, minimum, and maximum delay time. We then apply a flatMap function to compute the average based off the previous step.

Task 3: Most Popular Aircraft Types

We read in the data from all 3 files to include relevant fields. We then join the flights data on the aircrafts data and project the revelant fields. We then take the result of the previous join and join that with the airlines dataset and again project the relevant fields. We then apply a filter function to retrieve US flights. Finally, we apply 2 GroupReduceFunction where the first GroupReduceFunction iterates through and construct a count of how frequent each model appears. We then sort the results of this in descending order and retrieve the first 5 results. We then apply another reduceGroup function to retrieve and format the string necessary for the output according to specifications.

Tuning Decisions and Evaluation

Task 1: Top-3 Cessna Models

Blah blah blah

Task 2: Average Departure Delay

We use a broadcash hash join if a file is significantly smaller than another. The rationale is that a hash join will be executed whereby the data will be split into buckets, that are later merged. This achieves a speed up in run time in comparison to traditional loop joins.

Task 3: Most Popular Aircraft Types

Join join join.

Performance Evaluation

Task 1: Top-3 Cessna Models

blah blah blah

Task 2: Average Departure Delay

Blah blah blah

Task 3: Most Popular Aircraft Types

Join join join.