THE UNIVERSITY OF SYDNEY

University of Sydney

DATA3404

Data Science Platforms

Group Assignment

Team members:
450411920

Semester 1, 2019

Contents

Job Design Documentation **2**
Task 1 2
Task 2 2
Task 3 2

Tuning Decisions and Evaluation **3**
Task 1 3
Task 2 3
Task 3 3

Performance Evaluation **4**
Task 1: Top-3 Cessna Models 4
Task 2: Average Departure Delay 4
Task 3 4

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

Blah blah.

Task 3

Join join join.

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

blah blah blah

Task 2

Blah blah blah

Task 3: Most Popular Aircraft Types

Join join join.