**BENEDICT WAIHARO**

FLIGHT DATA ANALYSIS

CS 644

PROFESSOR BADER

Objectives

Design, implement, and run MapReduce jobs to find out

* the 3 airlines with the highest and lowest probability, respectively, for being on  
  schedule;
* the 3 airports with the longest and shortest average taxi time per flight (both in and  
  out), respectively; and
* the most common reason for flight cancellations.

Submission

A project report in PDF that includes:

* A detailed description of the algorithm you designed to solve each of the problems
* A performance measurement plot that compares the execution time in  
  response to an increasing data size (from 1 year to 22 years) and an in-depth  
  discussion on the observed performance comparison results

**Oozie Workflow**

A picture containing electronics, device

Description automatically generated

**Algorithm**

Flight Delay

Used 1 mapper and 1 reducer Mapper <k,v>:<UC,1 or0>

Read file, line by line. Considered arrival delay. If data of ArrDelaycol <= 5 mins == 1 else 0

Reducer: sum values of mapper. Same Key. Calculate total # of 0 and 1. Find probability of on schedule or not.

Sort and output 3 airlines.

Cleanup. If NULL, then no value == no output

Taxi Time

1 mapper and 1 reducer used

Mapper <k,v>: <IATA airport code, Taxi time>:<Origin,TaxiOut> or <Dest,Taxi>

Reducer <k,v>:<IATA airport code, Avg Taxi Time>

Reducer: Find avg taxi time. Sum total taxi time divided by Total count Value

Sort and output 3 airlines.

Cleanup. If NULL, then no value == no output

Cancellation

1 mapper and 1 reducer used

Read file and check for canceled value. If 1 and cancelation value is NA, commit entry value == 1

Mapper <k,v>: <Cancellation code, 1>

Reducer: <k,v>: <Cancellation code, sum of the 1>

Reducer. Sum value of from mapper, Same key.

This is the most common reason for cancellation.