FLIGHT DATA ANALYSIS

INTRODUCTION TO BIGDATA - CS 644

FALL 2019

PROFESSOR: Chase Wu

TA: Wuji Liu

BY

AKILESH R KANNAMUKKALA (ARK72)

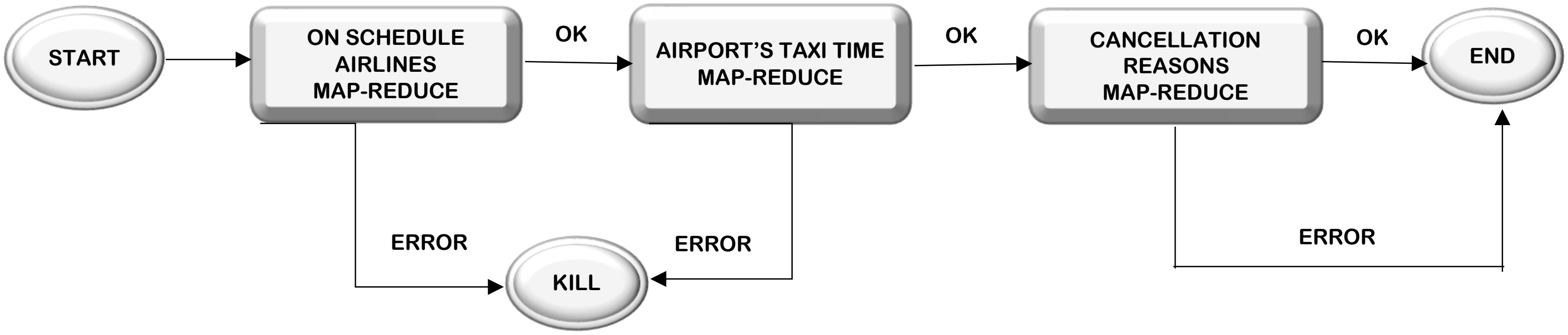
DARSHAN H SHAH (DS962)

INTRODUCTION

In this project, we have used Oozie workflow in conjunction with HDFS – MapReduce framework to analyze the Airline On-time Performance data set (flight data set) from the period of October 1987 to April 2008 obtained from the website http://stat-computing.org/dataexpo/2009/the-data.html. We have configured Hadoop in fully distributed mode. Furthermore, we have developed Oozie workflow for that to solve following 3 problems:

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

STRUCTURE OF OOZIE WORKFLOW


START
ON SCHEDULE AIRLINES MAP-REDUCE
OK
AIRPORT'S TAXI TIME MAP-REDUCE
OK
CANCELLATION REASONS MAP-REDUCE
OK
END
ERROR
KILL
ERROR
ERROR

As shown above, the OOZIE Workflow diagram depicts the different stages at which Map-Reduce works to deduce into desired results. As shown whenever the error occurs the process gets Killed. More detailed explanation about the **MAP-REDUCE STAGES** is given below.

ALGORITHM

First Map-Reduce: Airports Taxi Time

- Mapper <key,value>: <IATA airport code, TaxiTime>: <Origin,TaxiOut> or <Dest,TaxiIn>
- The Mapper reads the data line by line, ignores the first line. If the data of the TaxiIn or the TaxiOut column is not NA, output: <IATA airport code, TaxiTime>
- Reducer <key,value>: <IATA airport code, Average TaxiTime>
- Reducer sums the value from the mapper of the same key (normal) and calculates the total time this key is found (all). Then do the equation: normal/all to calculate the average TaxiTime of each key.
- Reducer then uses the Comparator function do the sorting. After sorting, output the 3 airports with the longest and shortest average taxi time.
- If the data is NULL, then output: There is no value can be used, so no output.

Second Map-Reduce: Cancellation Reasons

- Mapper <key,value>: < CancellationCode, 1>
- The Mapper reads the data line by line, ignores the first line. If the value of the Cancelled is 1 and the CancellationCode is not NA, output: < CancellationCode, 1>
- Reducer <key,value>: < CancellationCode, sum of the 1s>
- Reducer sums the value from the mapper of the same key.
- Reducer then uses the Comparator function do the sorting. After sorting, output the most common reason for flight cancellations.
- If the data is NULL, then output: There is no most common reason for flight cancellations.

Third Map-Reduce: On Schedule Airlines

- Mapper <key, value>:<UniqueCarrier,1 or 0>

- The Mapper reads the data line by line, ignores the first line and the NA data. If the data of the ArrDelay column which is less than or equal to 10 minutes, output: <UniqueCarrier,1>, otherwise output: <UniqueCarrier,0>
- Reducer <key, value>:<UniqueCarrier, probability>

- Probability = (# of 1) / (# of 1 and 0)

- Reducer sums the values from the mapper of the same key, the sum will be the number of this airline when it is on schedule. And calculates the total number of 0 and 1, then calculate the on schedule probability of this airline.

- Reducer then use the Comparator function do the sorting. After sorting, output the 3 airlines with the highest and lowest probability.

- If the data is NULL, then output: There is no value can be used, so no output.

PERFORMANCE MEASUREMENTS

Increasing number of VMs' with entire dataset

A performance measurement plot that compares the workflow execution time in response to an increasing number of VMs used for processing the entire data set (22 years) and an in-depth discussion on the observed performance comparison results

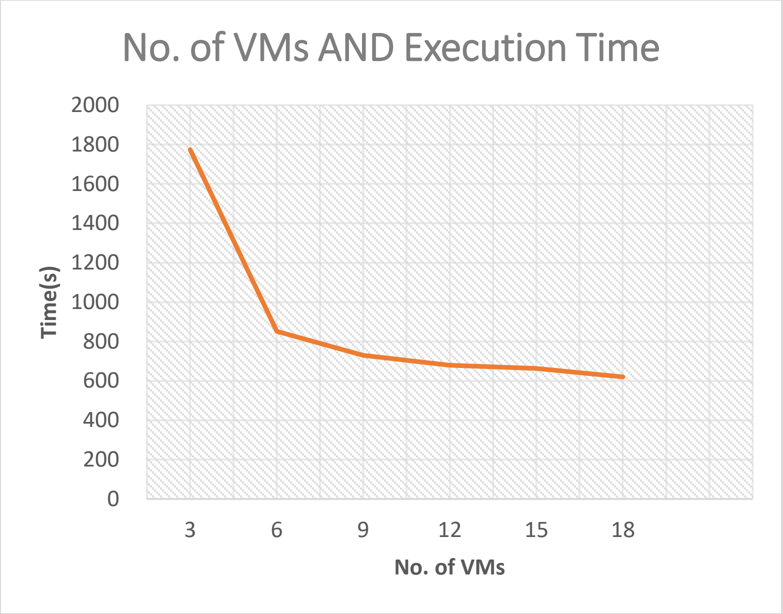


Figure2

As shown in above Figure, as the number of VM's increases, the execution time of OOZIE workflow will decrease. By increasing the number of the VMs, the processing ability of the hadoop cluster will also increase, because the data can be dealed with in parallel on more datanodes. Then the execution time of every map-reduce job will be shorter than before, thus the execution time of the oozie workflow will be shorter than before too. However, the execution time of deal with the same data size will not always decrease by increasing the number of VMs. When the execution time decrease to a certain range, although trying to increase the number of VM, the execution time will no longer decreasing anymore. The reason is more VMs means more information interaction time between the datanodes of a hadoop cluster. Information interaction time of a hadoop cluster increases when the number of VMs increases.

Increasing data size with same number of VMs

This is a performance measure plot that compares the workflow 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.

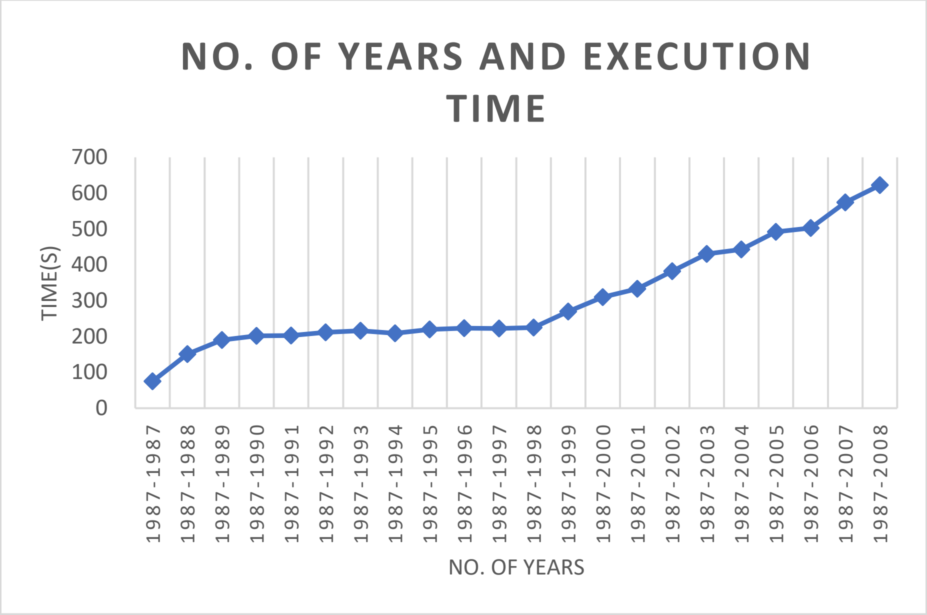


Figure3

As shown in above Figure, along with the increasing data size, the execution time of the oozie workflow will always increase too. In the beginning, the time-consuming increase with the increase in the amount of data, but the time-consuming increasing is slow, this is because the data increasing of first few years is not that much. On the contrary, after year 1998, the time-consuming increase very fast, the slope is become much steep compare to the first few years. The reason is the flight data between year 1998 to year 2008 is increase faster than the previous years. It also shows more and more people choose traveling by plane.

REFERENCES

AWS EC2 Setup:
Slides and support provided by the TA – Wuji Liu

Hadoop Installation & Setup:
https://www.youtube.com/watch?v=lRQqR0Fm1oE&list=PLWsYJ2ygHmWhOvtHIPxGJDtki9SJlINCw&index=4

Oozie Installation:
Slides and support provided by the TA – Wuji Liu