CS 644

INTRODUCTION TO BIG DATA

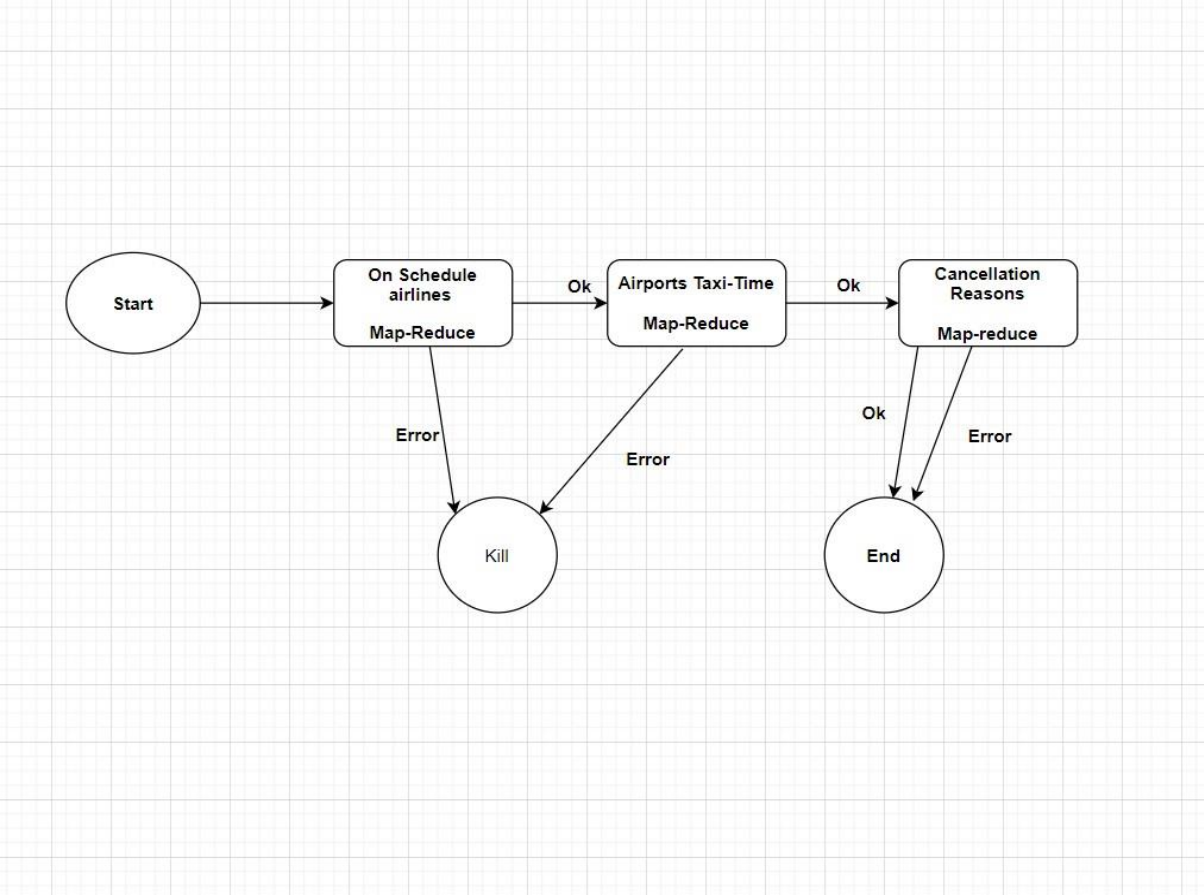
PROJECT – FLIGHT DATA ANALYSIS

BY:

Preethi Ravulapally – pr454

Karthik Ganeshula – kg496

i) Structure of Oozie Workflow:



ii) Algorithms:

a) First Map – Reduce: On Schedule Airlines

1. The Mapper starts first.

2. The mapper reads the data line by line but ignore the first line and the NA data. If the data of the ArrDelay is less than or equal to 10, it returns the output.

3. Now, the Reducer starts.

4. Reducer adds the values from the mapper of the same key and the sum will be the number of this aeroplanes of this airline on schedule. It then calculates the number of 0's and 1's and then calculate the on schedule probability of this airline.

5. The Reducer then uses the comparing functions to sort the data and then ouputs the 3 airlines with the highest and lowest probability.

6. If the data is NULL, then the output will be a statement stating that the operation cannot be done.

b) Second Map – Reduce : Airport Taxi- Time:

1. The Mapper starts first.

2. The Mapper reads the data line by line but ignores the first line. If the data of the TaxiIn or the TaxiOut column is NA, there will be no output.

3. Now, the reducer starts.

4. The Reducer adds the values from the mapper of the same key, calculates the total times the key is found. Then it does the calculation Normal/ All to calculate the Average time of each key.

5. The Reducer then uses the comparing functions to do the sorting. After sorting, it outputs the 3 airports with the shortest and longest waiting times.

6. If the data is NULL, the output will be a statement stating that there is no output.

c) Third Map – Reduce : Cancellation Reasons

1. The Mapper starts first.

2. The Mapper reads the data line by line but ignores the first line. If the value of the cancelled is 1 and the CancellationCode is NA, There is no output.

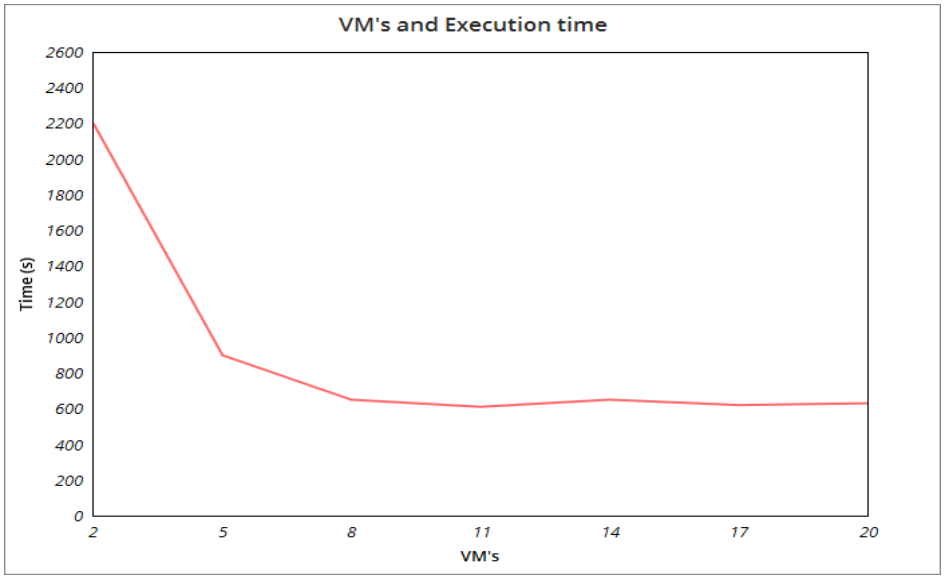
3. Now, the Reducer starts.

4. Reducer adds the values of the mapper of the same key.

5. The reducer then uses the comparing functions to do the sorting. After sorting, it outputs the most common reasons for the cancellations.

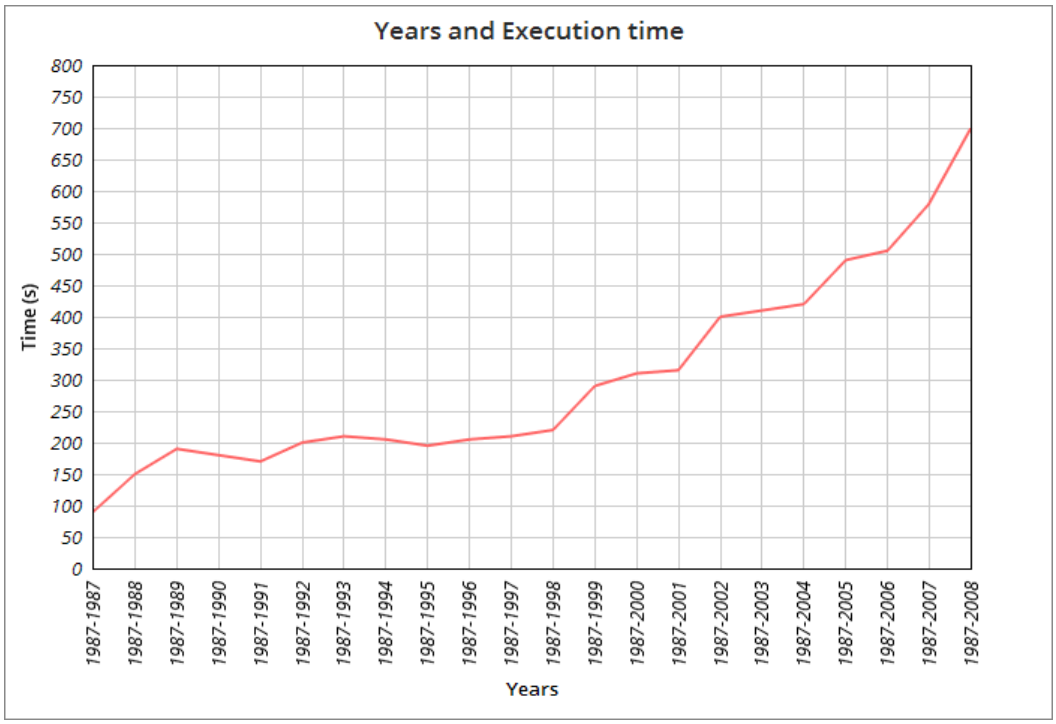
6. If the data is NULL, the output will be a statement that states that there is no output.

iii) Increasing number of VM's (Entire data set)



According to the above figure, along with the increasing number of VM's, the workflow execution time will decrease. By increasing the number of VM's the processing ability of the cluster also increases which is obvious because the number of machines working on the dataset is more than the previous number, every time we increase. But this will not be the case forever because as we increase the number of VM's at a point, the processing ability become stable. Though the ability keeps increasing, the interaction between the data nodes also increases with the number, which makes the cluster slower.

iv) Increasing size of data (20 Vm's)



According to the above figure, along with the increasing data size, the execution time of the workflow will always increase too. As the data increases, the time consumption also increases. As the trend goes by, we can observe that the time has a gradual increase until 1998 because, the amount of data generated after 1998 is bigger than the previous years and the time consumed is also high. This also tells us that the number of people preferring airways has also increased after 1998.