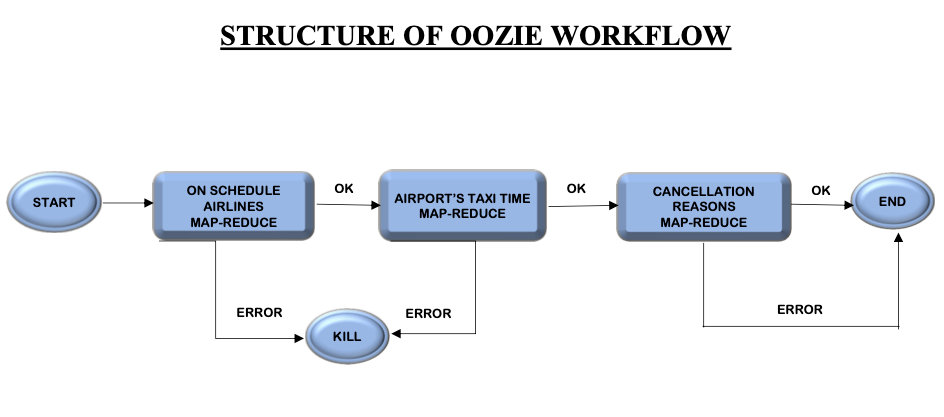
**Project of CS 644 : Intro to Big Data Flight Data Analysis**

**Project Report**

**Ketki Ambekar   
Parth Patel**





**Algorithm:**

**1. The 3 airlines with the highest and lowest probability, respectively, for being on schedule**

**Step 1:** The Mapper first splits all of the file elements. The value for a particular key will be given as “1”, if the combination of arrival and departure delays, otherwise the combination is “0” as one context. It also gives a value of “1” for the total values as another context.

**Step 2:** The Reducer then sums up all the values for the keys. Then the probability is calculated by using the following formula → **Probability = Count of No. Delays / Total Number of Flying**

**Step 3:** TheMapper then gets the input from the Reducer from the previous step and swaps the key values accordingly. The DecreasingComparator will sort the keys(Probability) in Descending order. This will give us the highest and lowest probability when moved back to the Reducer.

**Step 4:** The reducer delivers the final output according to requirements.

**2. The Three Airports With The Longest And Shortest Average Taxi Time Per Flight**

**Taxi-in:**

The Mapper to analyze taxi-in time:

* Step 1- Split the data by comma (,)
* Step 2- Select the column Taxi-in and analyse the data.
* Step 3- If the value is “NA” don’t do anything.
* Step 4- If the value is not “NA” then take Key as Destination Airport(Dest) and Value will be the taxi in (Taxi-in)   
    
  The Reducer to analyze taxi-in time.
* Step 1- Take the Sum of Taxi-in time.
* Step 2- Count the number of Flights.
* Step 3- Divide the Sum by the Count.

**Taxi-out:**

The Mapper to analyze taxi-out time:

* Step 1- Split the data by comma (,)
* Step 2- Select the column Taxi-out and analyse the data.
* Step 3- If the value is “NA” don’t do anything.
* Step 4- If the value is not “NA” then take Key as Origin Airport(Origin) and Value will be the taxi out (Taxiout)

The Reducer to analyze taxi-out time.

* Step 1- Take the Sum of Taxi-out time.
* Step 2- Count the number of Flights.
* Step 3- Divide the Sum by the Count.

Find the longest and shortest time by sorting.

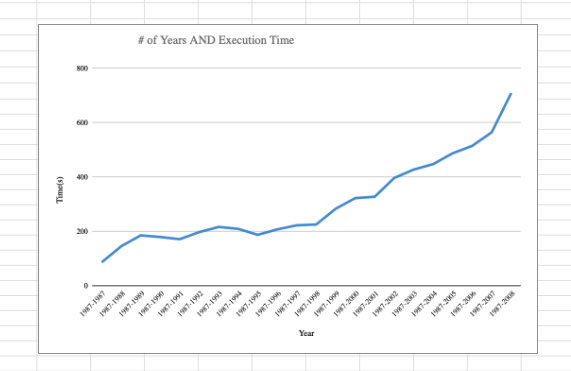
**3. Find The Most Common Reason For Flight Cancellations.**

**Step 1:** The Mapper splits all of the file elements first. The value for a key will be given “1” for every cancellation with its code as key.

**Step 2:** The Reducer will sum up all the values from the Mapper Phase and find the highest of all the values and context write the highest value with its cancel code as key.

**Time Vs Datasets:**

* As we can see in the chart below, the time increases with an increase in the number of datasets for the process. This proves that when the amount of data increases, the time to process also increases with it. Time is directly proportional towards the data size.



**Time Vs Number of Instances:**

* As we can see in the chart below, the time decreases with an increase in the number of instances to process our data. This proves that when the instances to process increases, the time to process our data decreases. The time is inversely proportional to our instance size.

