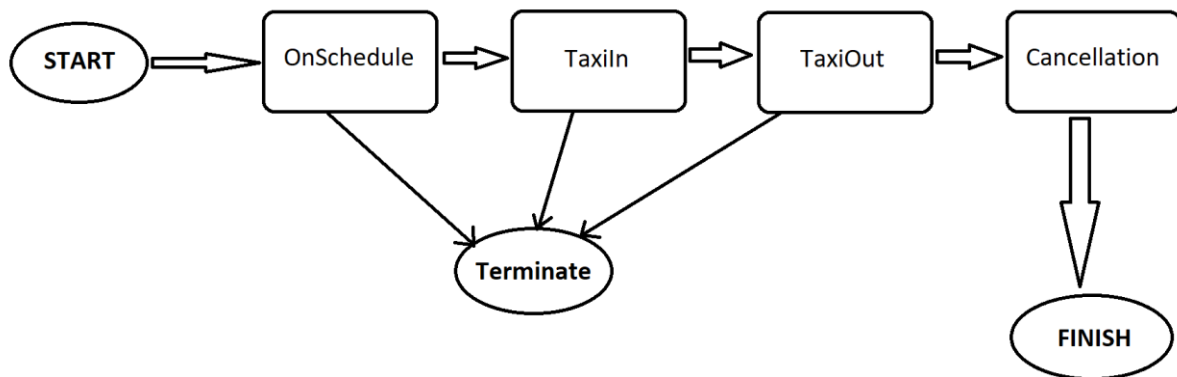


Oozie Workflow Diagram:



Algorithm:

1. The 3 Airlines with the highest and lowest probability, each, to be on schedule
 - For the flight to be on schedule, a benchmark of 10 has been set to determine if the flight is on schedule or not. If the Arrival Delay is greater than 10, it is classified as being late, i.e., not on schedule. Or else, the flight is on schedule.
 - The Mapper here takes the 'UniqueCarrier' column of the data as the key to write it to the context for the Reducer phase.
 - For each row, a small String "All: " is concatenated before the key and 1 is assigned as its value. And if the 'ArrivalDelay' column has a reading more than 10, a small String "Delayed: " is concatenated before the key. This is done simply to help the addition of each in the Reducer phase and calculate the probability.
 - The Reducer stores 6 instance variables for max & min (3 each) to keep record of the 3 maximum and 3 minimum flight carriers.
 - To get the total count, the Reducer checks if the key has the string "All: " in it and if it does, it calculates the total sum for that key. This is similarly done for the keys with the "Delayed: " string and the probability is taken out by dividing the two.
 - Now the 3 max and 3 min variables are updated according to the probability calculated for each row by a simple method call to the 6 methods.
 - Finally everything is written in to the cleanup function.

2. The 3 Airports with the longest and shortest taxi time per flight, each.
 - This program is done in 2 separate jobs, one each for Taxi In time and Taxi Out time. But the Reducer remains the same for both.
 - The Mapper function for Taxi In job does nothing but writes the “Dest” column of the data as the key to the context and “TaxiIn” column as the corresponding value.
 - Similarly, the Taxi Out job has “Origin” column as the key and “TaxiOut” column as the corresponding value.
 - The Reducer maintains the 3 max and 3 min values in the same way as the 1st Program, which is by maintaining 3 max and 3 min instance variables.
 - The Counter is run for each key and is updated accordingly.
 - Finally, the cleanup function is called, and all the data is written in to it for output.

3. The most common reason for flight cancellation.
 - This is a simple word count type of job.
 - The Mapper for this job writes into context the “CancellationCode” column of the data as the key and 1 as its value.
 - The Reducer does the counter for each key’s value and does the sum. The ‘maxCancellationCode’ instance variable maintains the maximum value as the code progresses.
 - The cleanup function in the end writes the maximum value in the context for the output.