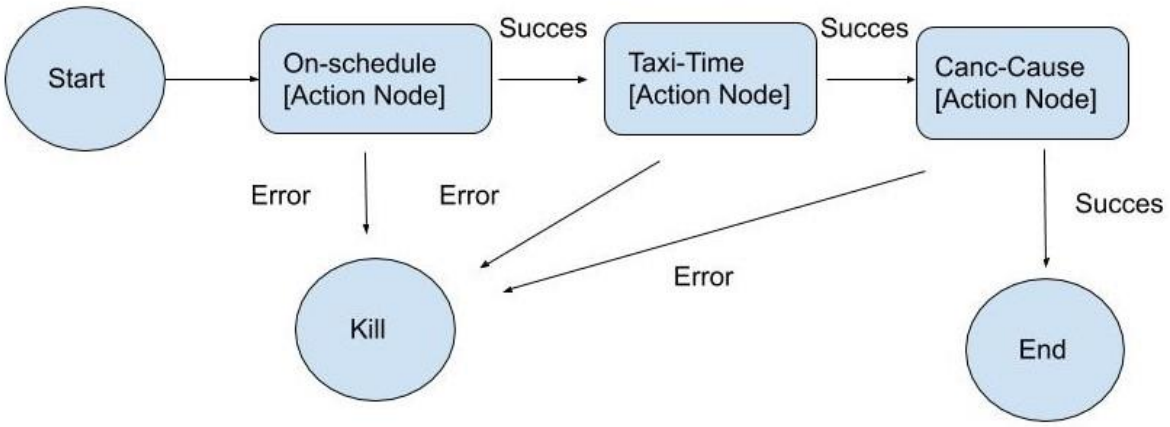
Project (Flight Data Analysis): CSS 644 Intro to Big Data
Date of assignment: December 15, 2019

Project Team:
Katie Perkins - kap92@njit.edu
Jamal Mirville - jrm93@njit.edu
Maha Faruque - mf454@njit.edu

Flight Data Analysis

22 years (1987-2008)

1. **Structure of Oozie Workflow**



2. **Algorithm descriptions:**

On Time Flights

- Define variables: delayThreshold = 10
- The Mapper, FlightScheduleMapper, computes the sum for each flight of Arrival Delay and Departure Delay as denoted by ArrDelay and DepDelay
- If the sum of ArrDelay & DepDelay is less than the delayThreshold, the flight is considered on schedule.
  Emit (UniqueCarrier, 1) to denote that this flight is on schedule.
- Else if the sum of ArrDelay & DepDelay is greater than the delayThreshold, the flight is considered as delayed.
  Emit (UniqueCarrier, 0) to denote that this flight is not on schedule.
- The Reducer, FlightScheduleReducer, uses the UniqueCarrier key to count the total number of values as denoted by totalCount.

- For each UniqueCarrier key from the Reducer, if value is 1, onSchedule increments by a count of one.
- Next the on-schedule probability is computed as onSchedule/totalCount.
- Add the (probability, UniqueCarrier) to an ArrayList.
- Steps 5-8 are repeated for each UniqueCarrier key.
- Once Reducer is complete, context cleanup will sort the arraylist in decreasing order of probabilities.
- An array list of values (UniqueCarrier, probability) is written to HDFS.

Average Taxi Time

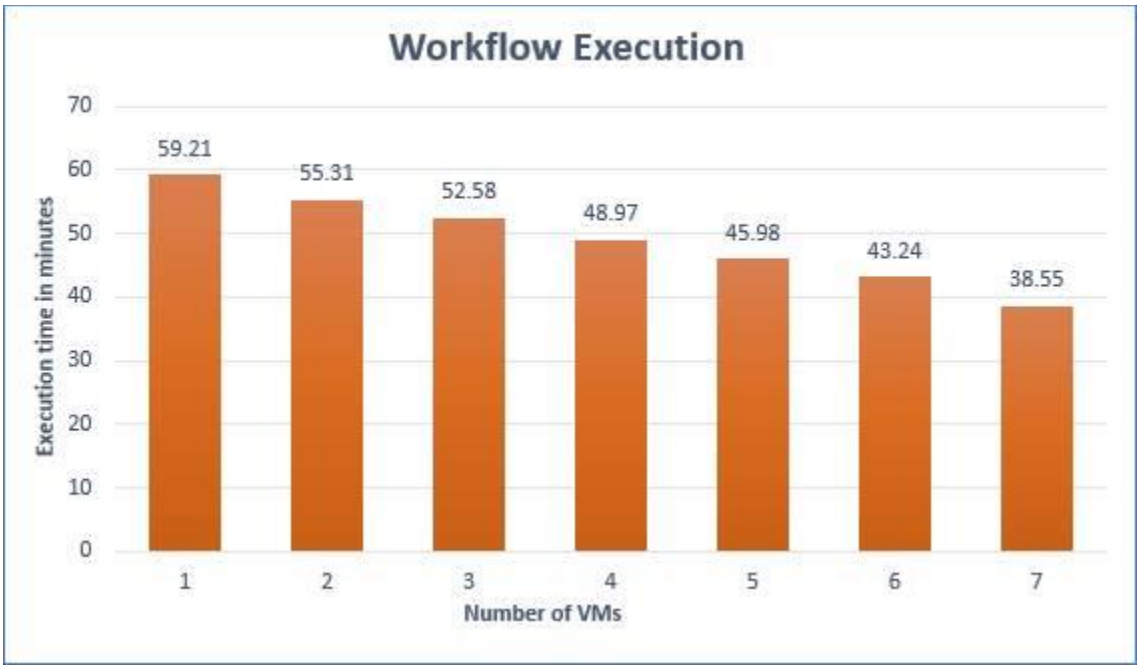
- The Mapper, AirportTaxiTimeMapper, emits (Origin, TaxiOut) & (Dest, TaxiIn) for each flight
- The Reducer, AirportTaxiTimeReducer, for each Origin/Dest key, will count the total number of values as totalCount
- For each Airport key from the Reducer, the total taxi time is computed by adding all the values (TaxiIn/TaxiOut).
- The Average Taxi Time for that Airport is computed by dividing the total taxi time obtained in above step by the totalCount.
- Add the (avgTaxiTime, Airport) to an ArrayList.
- Steps 3-5 are repeated for each Airport key.
- Once Reducer is complete, context cleanup will sort the array list in decreasing order of average taxi times.
- An array list values (Airport, avgTaxiTime) is written to HDFS.

Most Common Cancellation Cause

- The Mapper, FlightCancellationMapper, emits (CancellationCode, 1) if the flight is cancelled (Cancelled =1)
- The Reducer, FlightCancellationReducer, for each CancellationCode key, will add all the values to get the totalCount.
- (CancellationCode, totalCount) is written to HDFS.
- Steps 2-3 are repeated for each CancellationCode key.

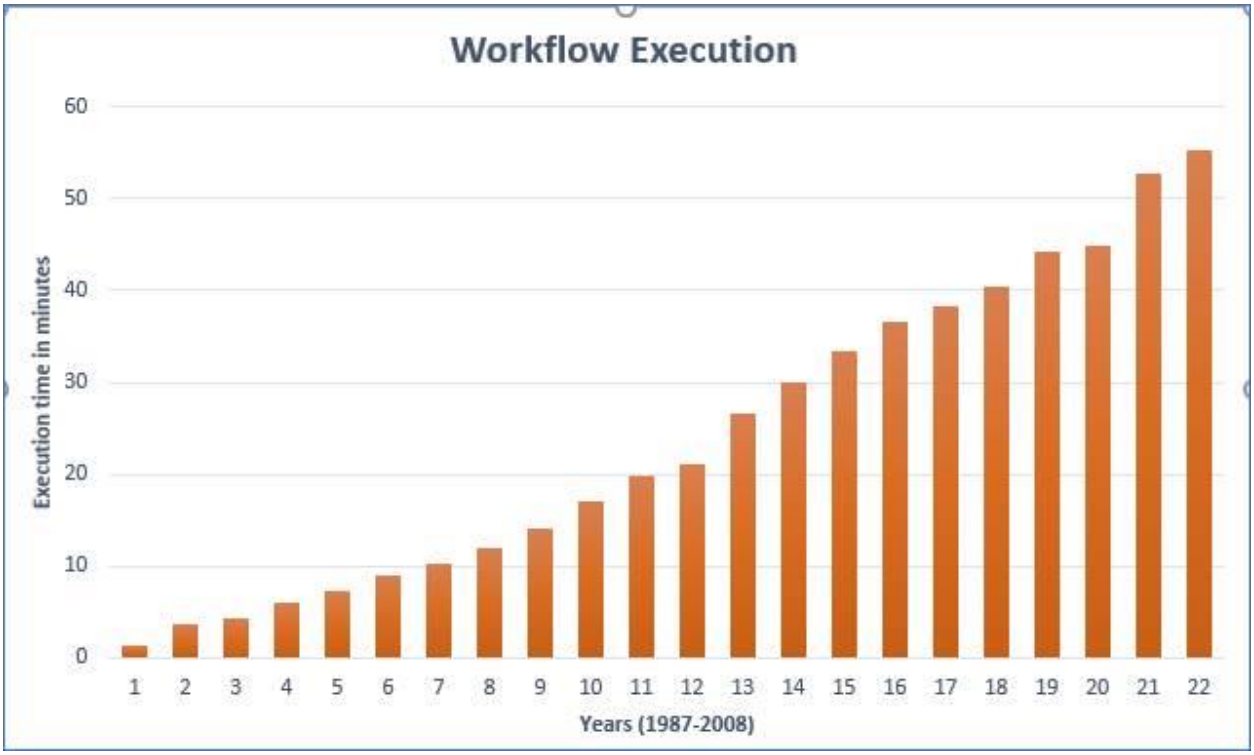
3. Performance Measurements:

Time response to an increase of VMs



This performance measurement plot compares the workflow execution time in response to an increasing number of virtual machines used for processing the entire data set of 22 years of flight data. Here you can see as we increased the virtual machines used, the execution time decreased. We may conclude that performance and number of resources used are directly proportional. This means our workflow improves it's time efficiency as we increase our use of virtual machines.

Time response to an increase of data size



This performance measurement plot compares the workflow execution time in response to an increasing data size (from 1 year to 22 years). In this experiment, we used two virtual machines, and gradually added each year of flight data. Here you can see that with one year of flight data, the virtual machines processed the input file in a little amount of time. We observed as the input data increases, the execution time will increase. We may conclude performance time and input data size are inversely proportional. This means we will need to increase the number of virtual machines to see a better performance output.