Name: Mohan Sai Bandarupalli
UCID: mb2279
Professor: Ravneet Kaur
CS 644 004 Introduction to Big Data

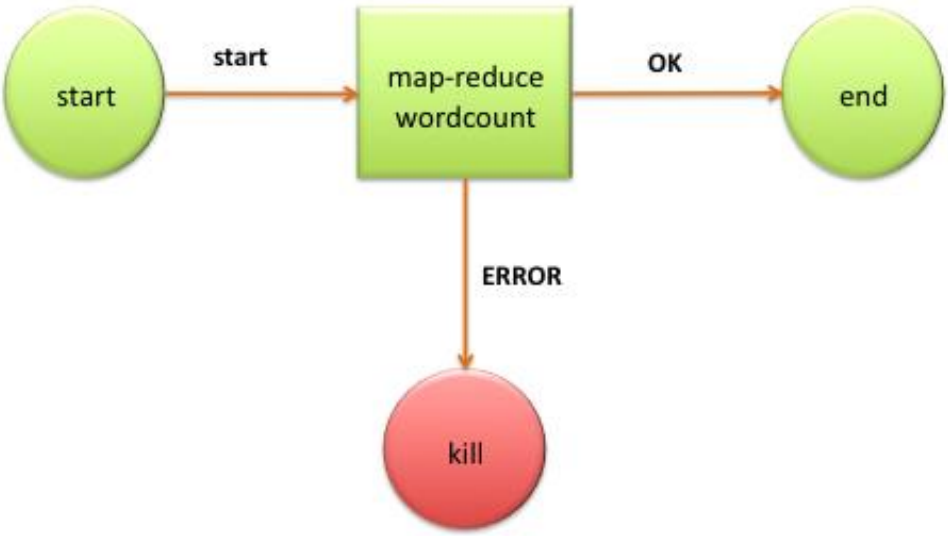
Project: Flight Data Analysis

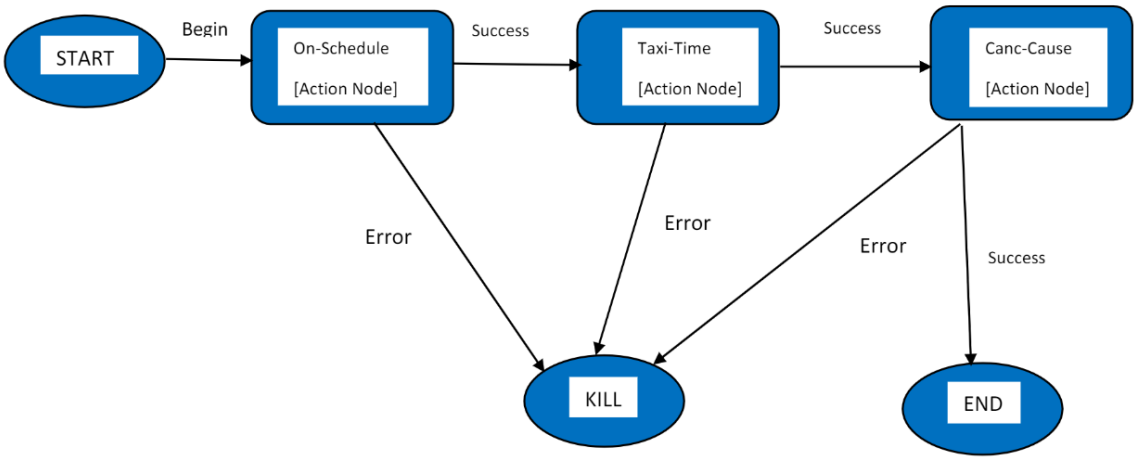
Introduction:

In this project, we analyze the Airline On-time Performance dataset spanning from October 1987 to April 2008. The objective is to use big data tools on AWS virtual machines (VMs), in Oozie, to find insights into the operational efficiency of airports and airlines and cancellations of flights.

Data Set: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

- **Workflow in Oozie Hadoop**





1.Finding Top 3 Airlines with the Highest and Lowest Probability of Being on Schedule:

Objective: Calculate the probability that each airline's flights are on time.

MapReduce Job 1:

Mapper: Each mapper reads a flight record and extracts information about the airline and whether the flight was on time or delayed.
Reducer: Receives the output from the mappers and calculates the probability of each airline being on schedule by dividing the number of on-time flights by the total number of flights for that airline.
Output: Provides key-value pairs where the key represents the airline code and the value represents the calculated probability of being on schedule.
MapReduce Job 2:

Mapper: Identifies the top 3 airlines with the highest probability of being on schedule by sorting the key-value pairs from Job 1.
Reducer: Selects the top 3 airlines with the highest probabilities based on the sorted data.
Output: Presents key-value pairs of the top 3 airlines with the highest probability of being on schedule.

MapReduce Job 3:

Mapper: Identifies the top 3 airlines with the lowest probability of being on schedule by sorting the key-value pairs from Job 1.
Reducer: Selects the top 3 airlines with the lowest probabilities based on the sorted data.
Output: Provides key-value pairs of the top 3 airlines with the lowest probability of being on schedule.

2. Finding Top 3 Airports with the Longest and Shortest Average Taxi Time per Flight:

Objective: Determine the average taxi-in and taxi-out times for each airport.

MapReduce Job 4:

Mapper: Extracts airport and taxi time data from input records.
Reducer: Calculates the average taxi time for each airport for both inbound and outbound flights.
Output: Yields key-value pairs where the key represents the airport code and the value represents the average taxi time.

MapReduce Job 5:

Mapper: Identifies the top 3 airports with the longest average taxi time by sorting the key-value pairs from Job 4.
Reducer: Selects the top 3 airports with the longest average taxi time based on the sorted data.
Output: Presents key-value pairs of the top 3 airports with the longest average taxi time.

MapReduce Job 6:

Mapper: Identifies the top 3 airports with the shortest average taxi time by sorting the key-value pairs from Job 4.
Reducer: Selects the top 3 airports with the shortest average taxi time based on the sorted data.
Output: Provides key-value pairs of the top 3 airports with the shortest average taxi time.

3.Finding the Most Common Reason for Flight Cancellations:

Objective: Identify the most common reasons for flight cancellations.

MapReduce Job 7:

Mapper: Extracts cancellation reason data from input records.
Reducer: Counts the occurrences of each cancellation reason.
Output: Produces key-value pairs where the key represents the cancellation reason and the value represents the count.

MapReduce Job 8:

Mapper: Identifies the cancellation reason with the highest count.
Reducer: Selects the cancellation reason with the highest count.
Output: Provides a key-value pair of the most common cancellation reason.