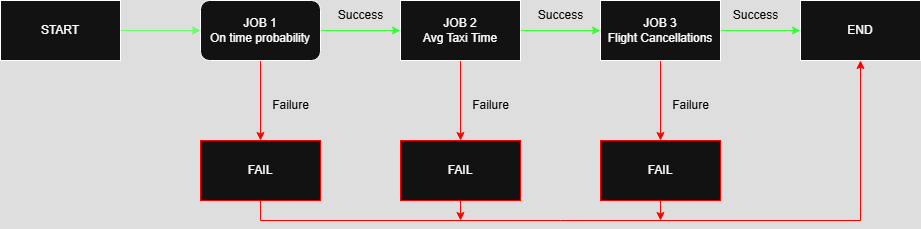
**Project – Airline Data Analysis**

DS644 – Introduction to Big Data

Dylan Da Costa (dd567), Omkar Naik (on36), Vaishnavi Gandhi (vg489)

1. **Oozie Workflow Diagram**



1. **Explanation of Code Logic for the 3 Jobs**

* On time Probability
* Avg Taxi Time
* Flight Cancellations

1. **Performance Measurement Plots**

* Increasing VMs vs Execution Time (Entire Dataset)

This plot shows how the total execution time changes when the number of virtual machines (VMs) is increased, while processing the constant data size of 22 years.

The plot indicates that execution time decreases significantly when increasing from 2 to 4 VMs, highlighting the benefits of parallel processing. However, beyond 4 or 5 VMs, the reduction in execution time becomes marginal. This pattern reflects diminishing returns, where the overhead of managing more VMs limits further performance improvements. It suggests there is an optimal range for VM usage, beyond which additional resources do not yield proportional gains.

A line graph with numbers and a line

AI-generated content may be incorrect.

* Varying Dataset Size vs Execution Time

This plot illustrates how execution time increases as the amount of data, measured in years, increases while using a fixed number of VMs.

Initially, from 1 to around 10 years, the increase in execution time is gradual, suggesting efficient scaling for smaller datasets. However, as the number of years continues to grow, especially beyond 15, the execution time begins to rise more steeply. This indicates that the system begins to face performance bottlenecks due to resource constraints such as CPU, memory, or I/O limits. The steeper slope at the higher data volumes reflects the increasing complexity and load on the system.

A line graph with orange lines

AI-generated content may be incorrect.