PROJECT TITLE: FLIGHT DELAY ANALYSIS USING BIG DATA PROCESSING WITH MAPREDUCE

NAME: Tanushri Vijayakumar
COURSE NO: DS644
SECTION NO: 852
UCID: 31698861
EMAIL: tv233@njit.edu

DATASET DESCRIPTION
Dataset Name: Flight Delay and Cancellation Dataset (2019-2023)
Dataset Source:
https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023?resource=download&select=flights_sample_3m.csv

Dataset Description:
This dataset contains detailed records of flight operations over a period spanning from 2019 to 2023. The data is structured in a CSV format and includes key information such as:
Flight Date: The scheduled date of the flight.
Airline/CARRIER: Codes indicating the airline operating the flight.
Flight Number: The specific flight identifier.
Departure and Arrival Times: Scheduled and actual departure/arrival times to measure delays.
Delay Metrics: Information on the duration of delays (if any) for departures or arrivals.
Cancellation Indicators: Flags to indicate if a flight was cancelled, along with possible reasons for the cancellation.
Additional Operational Metrics: Other fields may include airport codes (origin/destination), taxi times, and other relevant performance indicators.

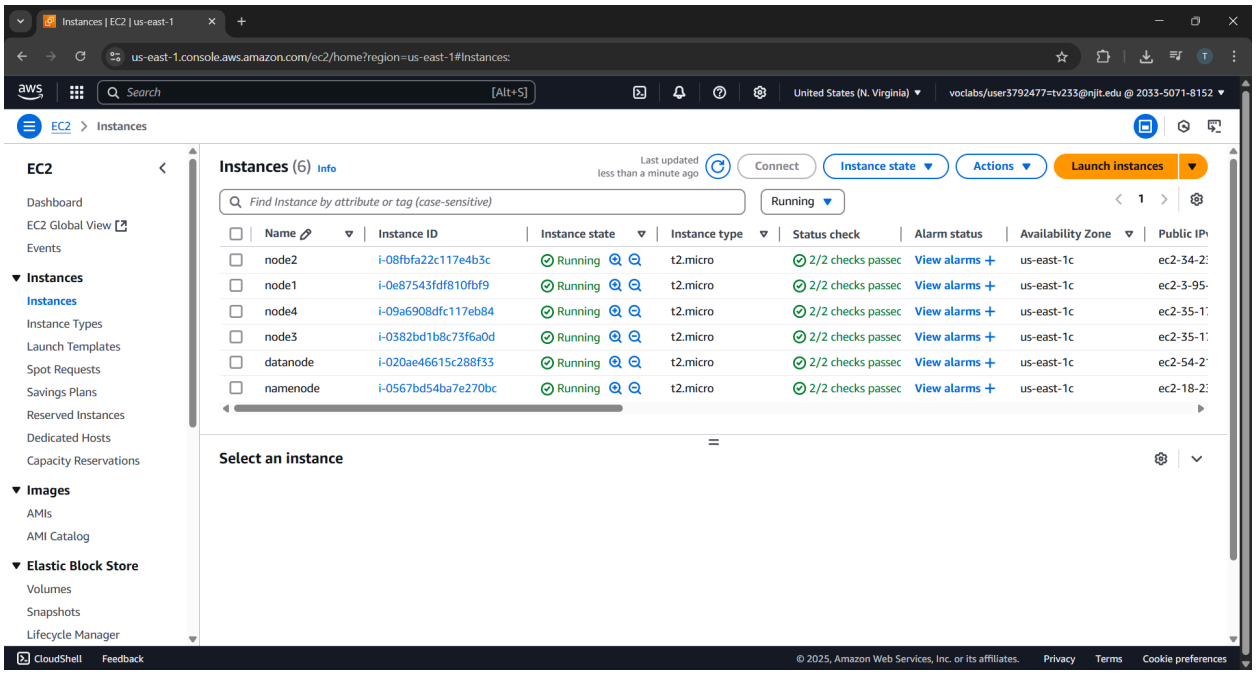
The dataset is organized in tabular form where each row represents a single flight record with multiple attributes that can be parsed using MapReduce.

Size & Format: File Size: The sample file, named “flights_sample_3m.csv,” contains approximately 3 million rows of flight records, and it’s size is : 585.7 MB
Format: CSV (Comma-Separated Values)

Reason for Selection: This dataset is especially useful for MapReduce processing due to its large volume of data, structured format. And the dataset provides practical insights into flight performance metrics.

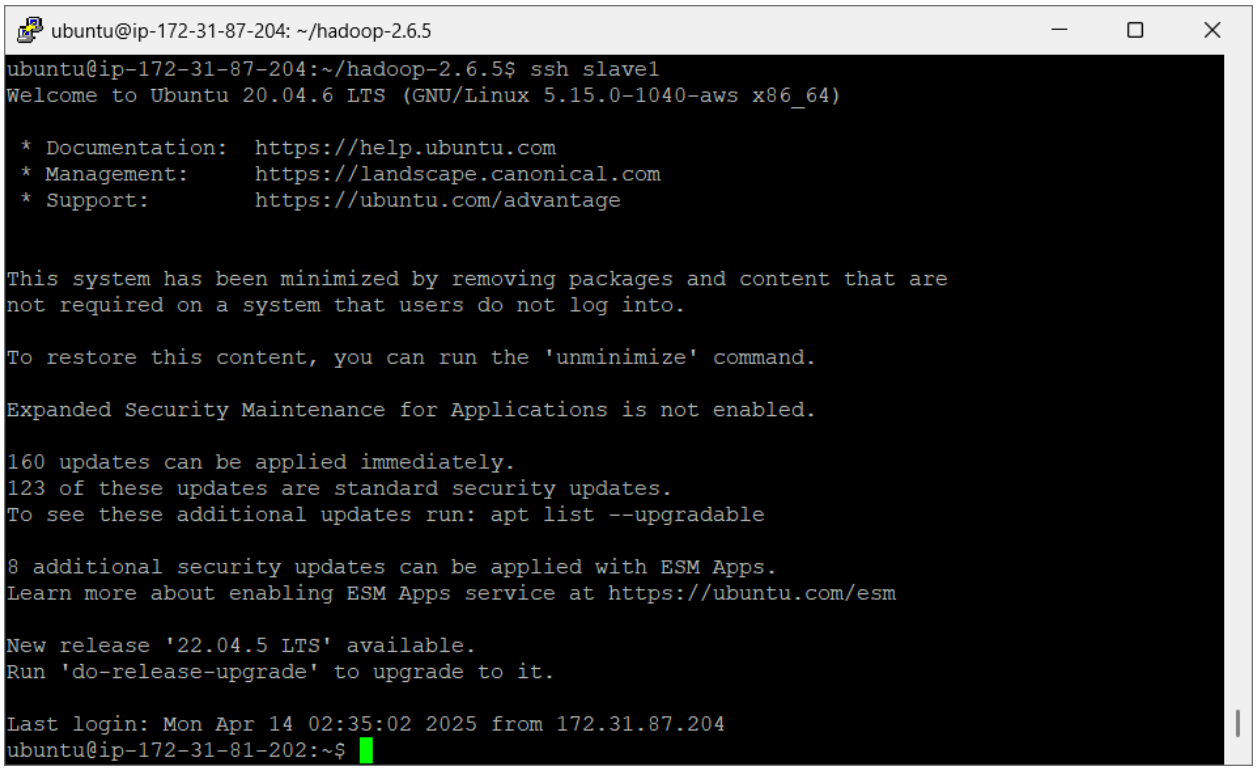
Hadoop Cluster Setup Successful Screenshots:

Namenode and datanode initialised and running successfully and passed 2/2 status checks:

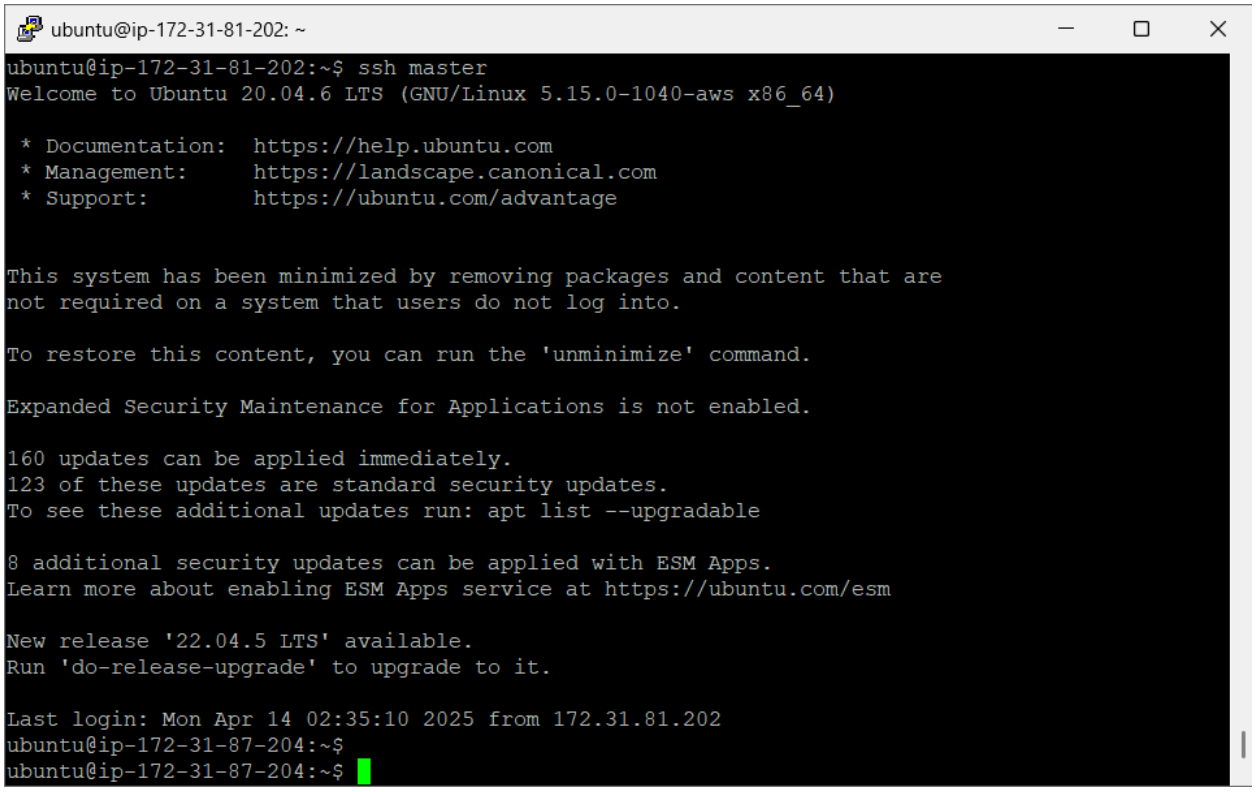


Passphraseless SSH login:

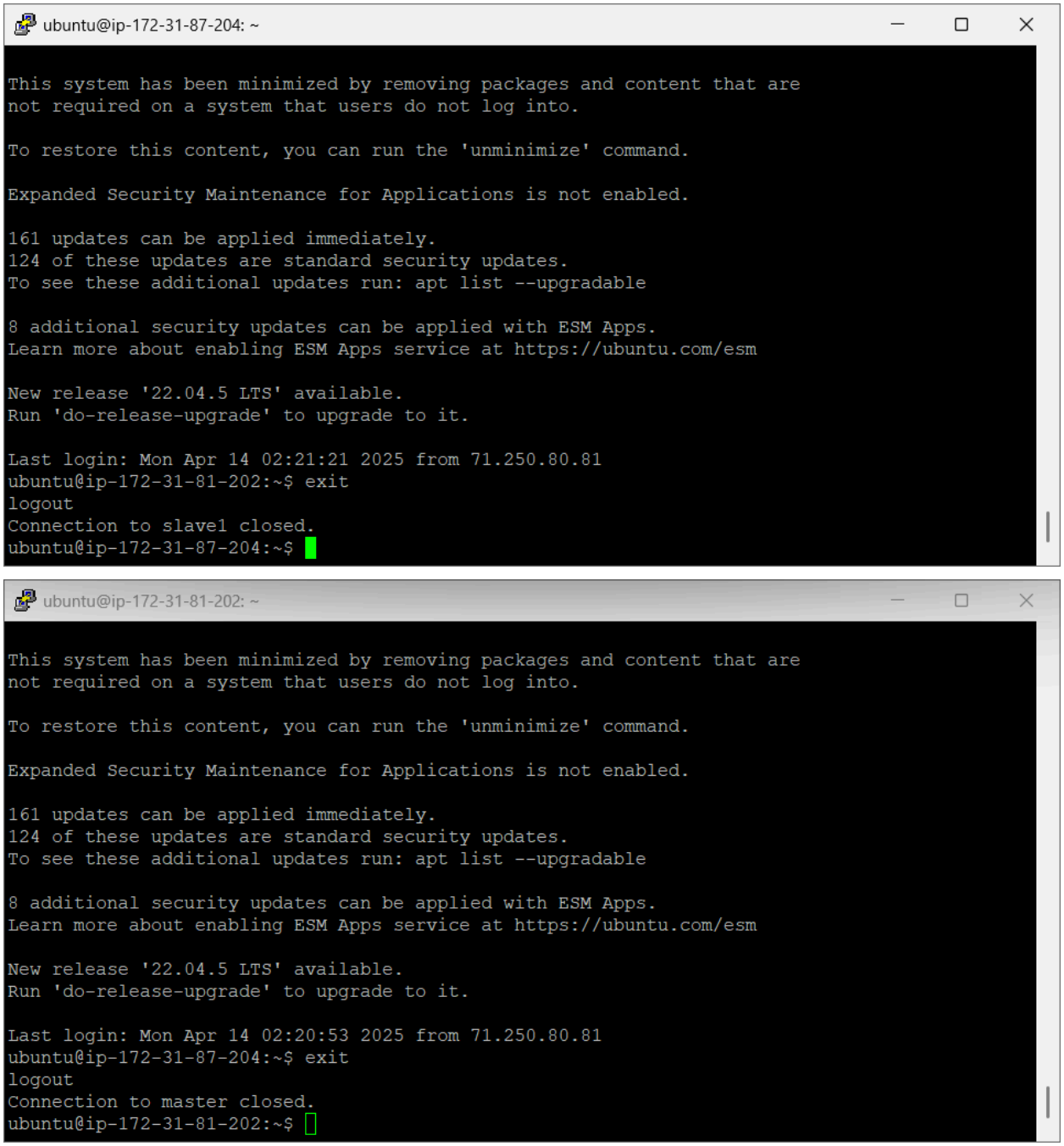
Ssh login to slave from master node:



Ssh login to master node from slave node:

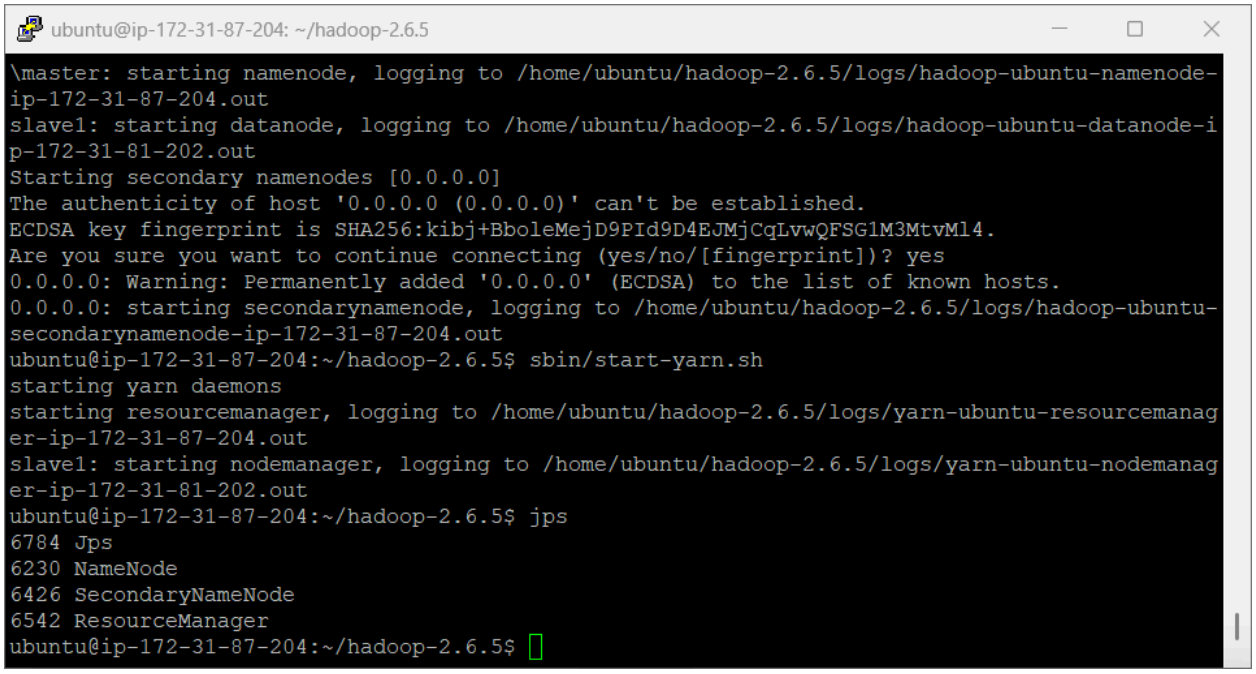


Ssh connection closed:



Output of the "jps" command

Master node:



Slave node: