

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.

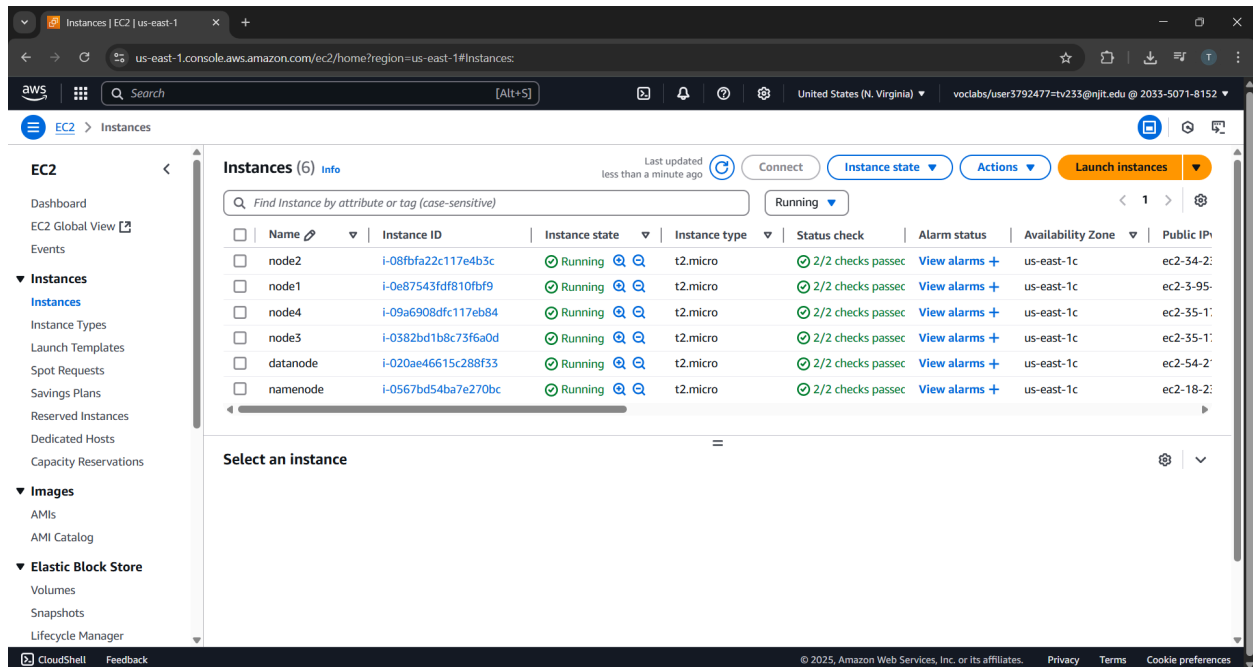
Size & Format: File Size: The sample file, named "flights_sample_3m.csv," contains approximately 3 million rows of flight records, and its size is : 585.7 MB

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.

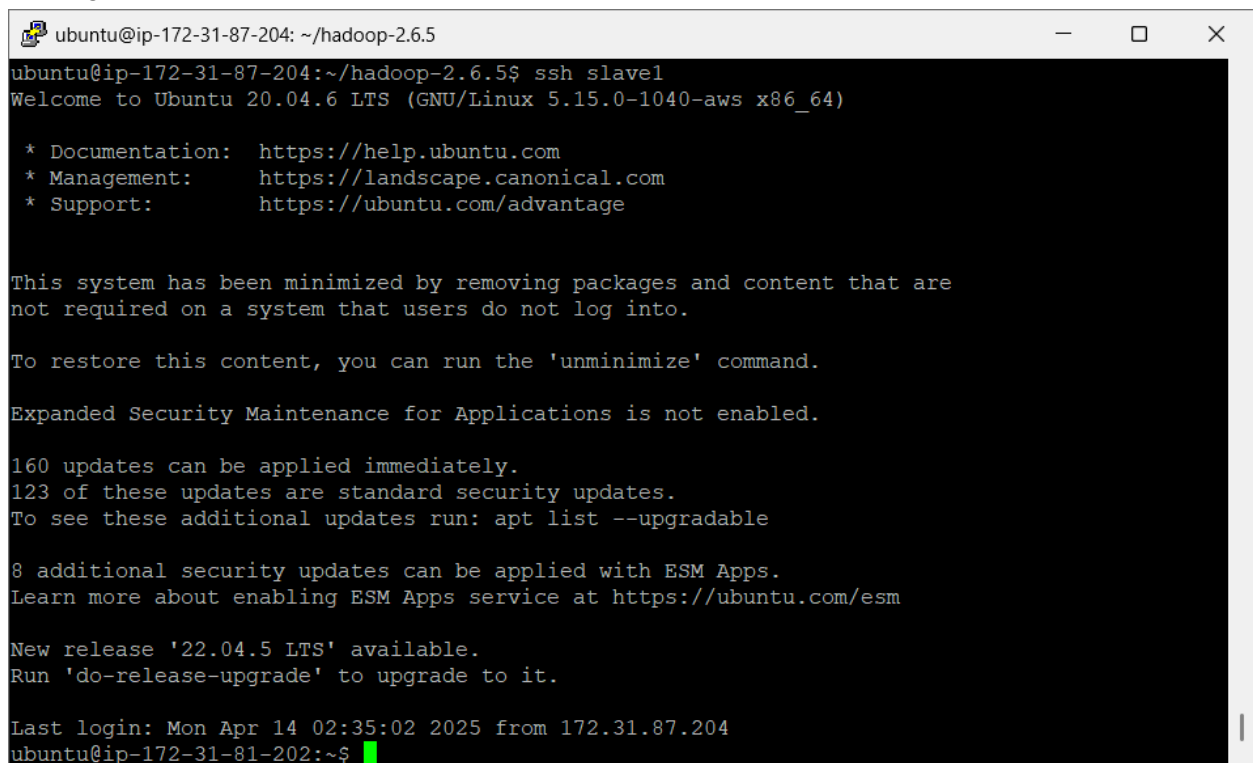
Hadoop Cluster Setup Successful Screenshots:

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:

```
ubuntu@ip-172-31-81-202: ~  
ubuntu@ip-172-31-81-202:~$ ssh master  
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1040-aws x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:       https://ubuntu.com/advantage  
  
This system has been minimized by removing packages and content that are  
not required on a system that users do not log into.  
  
To restore this content, you can run the 'unminimize' command.  
  
Expanded Security Maintenance for Applications is not enabled.  
  
160 updates can be applied immediately.  
123 of these updates are standard security updates.  
To see these additional updates run: apt list --upgradable  
  
8 additional security updates can be applied with ESM Apps.  
Learn more about enabling ESM Apps service at https://ubuntu.com/esm  
  
New release '22.04.5 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Mon Apr 14 02:35:10 2025 from 172.31.81.202  
ubuntu@ip-172-31-87-204:~$  
ubuntu@ip-172-31-87-204:~$
```

Ssh connection closed:

```
ubuntu@ip-172-31-87-204: ~  
  
This system has been minimized by removing packages and content that are  
not required on a system that users do not log into.  
  
To restore this content, you can run the 'unminimize' command.  
  
Expanded Security Maintenance for Applications is not enabled.  
  
161 updates can be applied immediately.  
124 of these updates are standard security updates.  
To see these additional updates run: apt list --upgradable  
  
8 additional security updates can be applied with ESM Apps.  
Learn more about enabling ESM Apps service at https://ubuntu.com/esm  
  
New release '22.04.5 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Mon Apr 14 02:21:21 2025 from 71.250.80.81  
ubuntu@ip-172-31-81-202:~$ exit  
logout  
Connection to slave1 closed.  
ubuntu@ip-172-31-87-204:~$
```

```
ubuntu@ip-172-31-81-202: ~  
  
This system has been minimized by removing packages and content that are  
not required on a system that users do not log into.  
  
To restore this content, you can run the 'unminimize' command.  
  
Expanded Security Maintenance for Applications is not enabled.  
  
161 updates can be applied immediately.  
124 of these updates are standard security updates.  
To see these additional updates run: apt list --upgradable  
  
8 additional security updates can be applied with ESM Apps.  
Learn more about enabling ESM Apps service at https://ubuntu.com/esm  
  
New release '22.04.5 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Mon Apr 14 02:20:53 2025 from 71.250.80.81  
ubuntu@ip-172-31-87-204:~$ exit  
logout  
Connection to master closed.  
ubuntu@ip-172-31-81-202:~$
```

Output of the "jps" command

Master node:

```
ubuntu@ip-172-31-87-204: ~/hadoop-2.6.5
\master: starting namenode, logging to /home/ubuntu/hadoop-2.6.5/logs/hadoop-ubuntu-namenode-
ip-172-31-87-204.out
slavel: starting datanode, logging to /home/ubuntu/hadoop-2.6.5/logs/hadoop-ubuntu-datanode-i
p-172-31-81-202.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:kibj+BboleMejD9Pid9D4EJMjCqLvwQFSG1M3MtvMl4.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/hadoop-2.6.5/logs/hadoop-ubuntu-
secondarynamenode-ip-172-31-87-204.out
ubuntu@ip-172-31-87-204:~/hadoop-2.6.5$ sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/ubuntu/hadoop-2.6.5/logs/yarn-ubuntu-resourcemanager-ip-172-31-87-204.out
slavel: starting nodemanager, logging to /home/ubuntu/hadoop-2.6.5/logs/yarn-ubuntu-nodemanager-ip-172-31-81-202.out
ubuntu@ip-172-31-87-204:~/hadoop-2.6.5$ jps
6784 Jps
6230 NameNode
6426 SecondaryNameNode
6542 ResourceManager
ubuntu@ip-172-31-87-204:~/hadoop-2.6.5$
```

Slave node:

```
ubuntu@ip-172-31-81-202: ~
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logos/
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logos/maven-feather.png
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logos/build-by-maven-black.png
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logos/build-by-maven-white.png
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/maven-logo-2.gif
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logo_apache.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/h3.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/icon_error_sml.gif
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/bg.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/h5.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/banner.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/breadcrumbs.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/logo_maven.jpg
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/icon_info_sml.gif
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/external.png
hadoop-2.6.5/share/doc/hadoop/hadoop-minikdc/images/collapsed.gif
ubuntu@ip-172-31-81-202:~$ vim ~/.bash_profile
ubuntu@ip-172-31-81-202:~$ source ~/.bash_profile
ubuntu@ip-172-31-81-202:~$ jps
6193 DataNode
6419 Jps
6325 NodeManager
ubuntu@ip-172-31-81-202:~$
```