**Innovative Assignment**

**Flight Data Analysis using Hive**

**Date:** 2 Nov, 2023

**Course Code and Name:** 2CS702 Big Data Analysis

|  |  |
| --- | --- |
| **Sr No** | **Title** |
| 1 | Problem Statement |
| 2 | Introduction |
| 3 | Hive |
| 4 | Implementation |
| 5 | Results |
| 6 | Applications |
| 7 | Conclusion |

**BY:**

**20BCE204**

**20BCE206**

**Problem Statement :**

The objective of this comprehensive analysis is to glean meaningful insights from aviation datasets, focusing on airports, airlines, and routes. The analysis aims to address specific queries to provide a nuanced understanding of the aviation landscape. The outlined tasks are as follows:

**A. Identification of Airports in India:**

- Extract and compile a detailed list of airports operating within the sovereign boundaries of India. This information will be crucial for various stakeholders, including government agencies, airlines, and infrastructure planners.

**B. Compilation of Zero-Stop Flights:**

- Identify and present a comprehensive list of airlines offering zero-stop flights. This data will be instrumental in understanding direct flight options and can be leveraged by travelers and airline companies alike.

**C. Analysis of Airlines with Code Share Agreements:**

- Compile a detailed list of airlines engaged in code-share agreements. This analysis will shed light on collaborative practices within the aviation industry and aid in understanding the dynamics of airline partnerships.

**D. Determination of Country/Territory with Highest Airports:**

- Investigate and ascertain the country or territory boasting the highest number of airports. This information will be pivotal for global aviation planning, government policies, and strategic investments in airport infrastructure.

**E. Compilation of Active Airlines in the United States:**

- Identify and present a comprehensive list of active airlines currently operating in the United States. This data is crucial for market research, competition analysis, and strategic decision-making within the U.S. aviation sector.

This problem statement sets the stage for a thorough exploration of the aviation data landscape, utilizing advanced data processing techniques. The outcomes of this analysis will contribute valuable insights for stakeholders ranging from government bodies to industry participants, fostering a deeper understanding of the intricate dynamics within the aviation domain.

**Overview:**

The Airlines Analysis Hadoop project aims to extract insights from datasets on airports, airlines, and routes. Using Hadoop, MapReduce, and Hive, the project focuses on tasks such as listing Indian airports, identifying zero-stop flights, listing code-share airlines, finding the country with the most airports, and listing active U.S. airlines. Challenges include handling large datasets and ensuring data quality. The project concludes with summarizing key findings for a better understanding of the aviation landscape.

**Introduction**

In the dynamic landscape of aviation, understanding the operational intricacies of airports, airlines, and routes is crucial for informed decision-making and strategic planning. The Airlines Analysis Hadoop project embarks on a comprehensive exploration of datasets encompassing airports, airlines, and routes, employing cutting-edge technologies such as Hadoop, MapReduce, and Hive for efficient and distributed data processing.

This report encapsulates the project's objective, dataset description, and analysis tasks, offering a glimpse into the vast realm of the aviation industry. By delving into the details of airports, airlines, and routes, the project aims to uncover valuable insights that can shape the future of air travel. Through the use of Hadoop-based tools, the report outlines the technology stack employed, the expected outputs, and the challenges encountered during the analytical journey.

The subsequent sections provide a detailed overview of the datasets, the analysis tasks at hand, the chosen technology stack, and the anticipated conclusions. As we navigate through the complexities of airport distribution, airline operations, and route characteristics, the report sets the stage for a compelling exploration of the aviation data landscape.

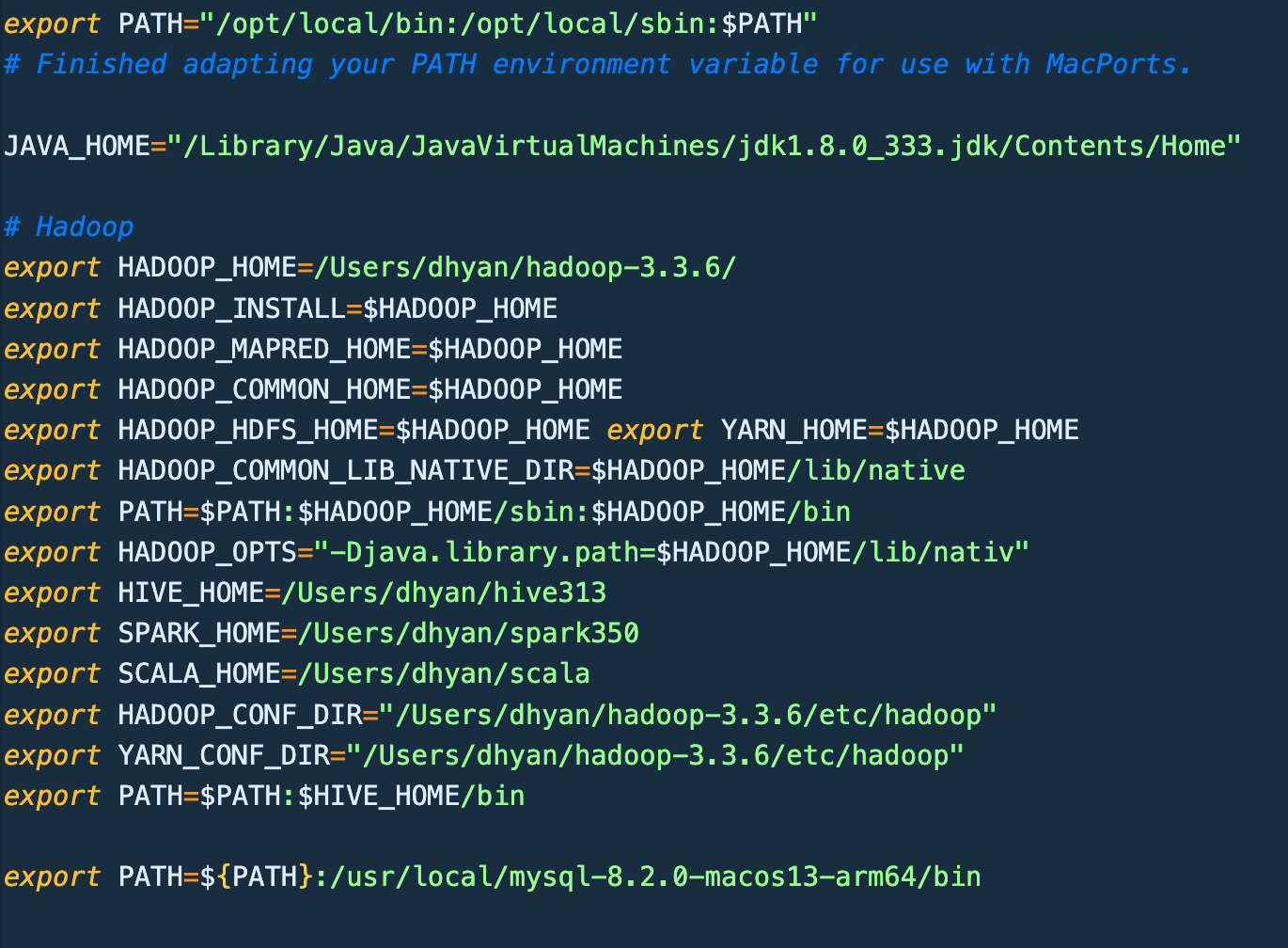
**Steps to Install Hive:**

1. Download and Untar Hive:

$ wget https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz

$ tar xzf apache-hive-3.1.2-bin.tar.gz

1. **Configure Hive Environment Variables (zprofile):**



1. **Connecting MySQL and Setting up JDBC Driver for MySQL**

(base) dhyan@Dhyans-MacBook-Pro lib % cp /Users/dhyan/Downloads/mysql-connector-j-8.2.0.tar.gz .

(base) dhyan@Dhyans-MacBook-Pro lib % tar xvf mysql-connector-j-8.2.0.tar.gz

1. **Setting up Hive.xml file**

(base) dhyan@Dhyans-MacBook-Pro hive313 % ls

LICENSE binary-package-licenses jdbc

NOTICE conf lib

RELEASE\_NOTES.txt examples scripts

bin hcatalog

(base) dhyan@Dhyans-MacBook-Pro hive313 % cd conf

(base) dhyan@Dhyans-MacBook-Pro conf % ls

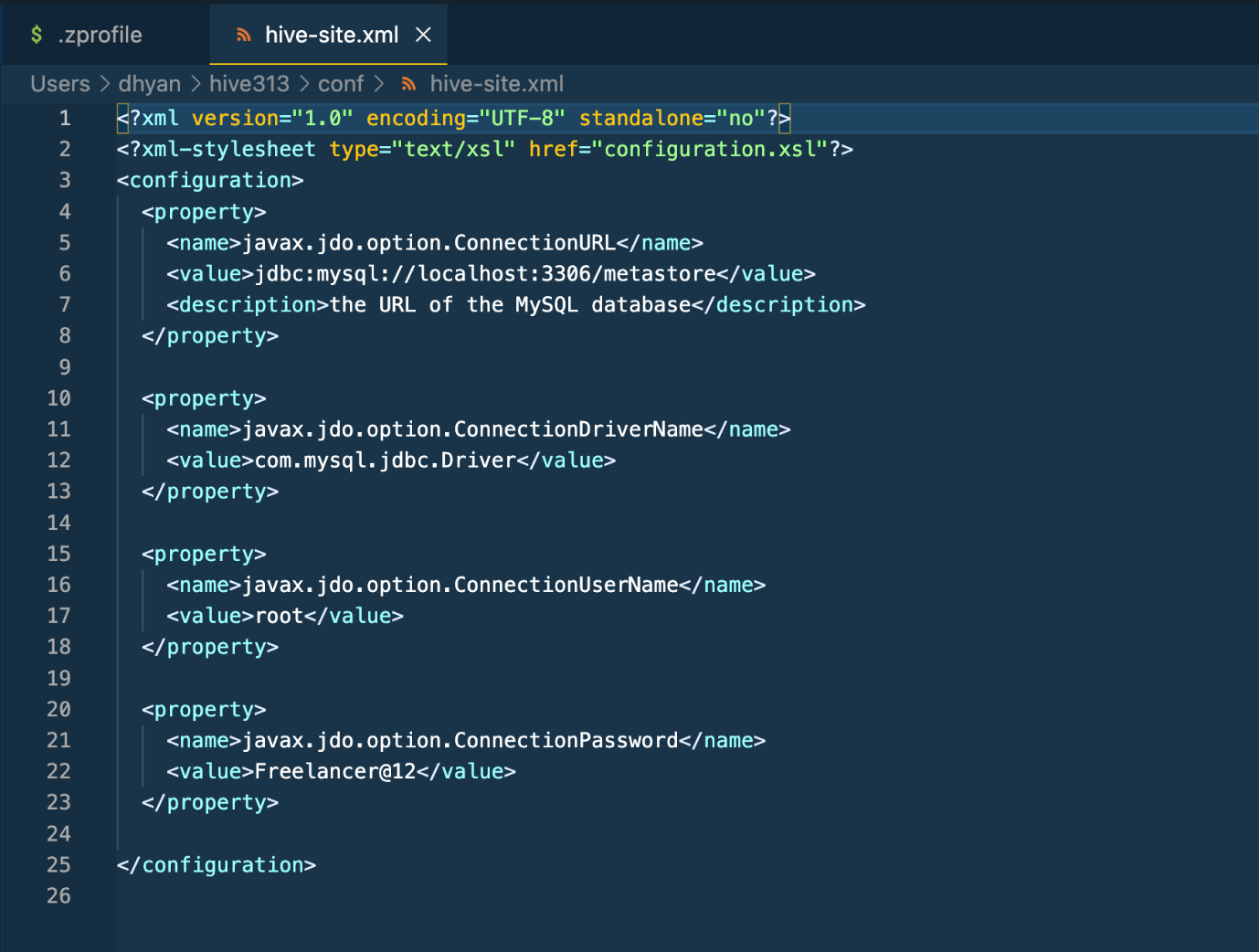
beeline-log4j2.properties.template ivysettings.xml

hive-default.xml.template llap-cli-log4j2.properties.template

hive-env.sh.template llap-daemon-log4j2.properties.template

(base) dhyan@Dhyans-MacBook-Pro conf % cp hive-default.xml.template hive-site.xml

(base) dhyan@Dhyans-MacBook-Pro conf % code hive-site.xml



1. **Starting hive**

(base) dhyan@Dhyans-MacBook-Pro ~ % schematool -dbType mysql -initSchema

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/Users/dhyan/hive313/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/Users/dhyan/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Metastore connection URL: jdbc:mysql://localhost:3306/metastore

Metastore Connection Driver : com.mysql.jdbc.Driver

Metastore connection User: root

Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.

Starting metastore schema initialization to 3.1.0

Initialization script hive-schema-3.1.0.mysql.sql

Initialization script completed

schemaTool completed

(base) dhyan@Dhyans-MacBook-Pro ~ % hive

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/Users/dhyan/hive313/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/Users/dhyan/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Hive Session ID = 96402e96-147e-40f6-bd07-3964e25604dc

Logging initialized using configuration in jar:file:/Users/dhyan/hive313/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true

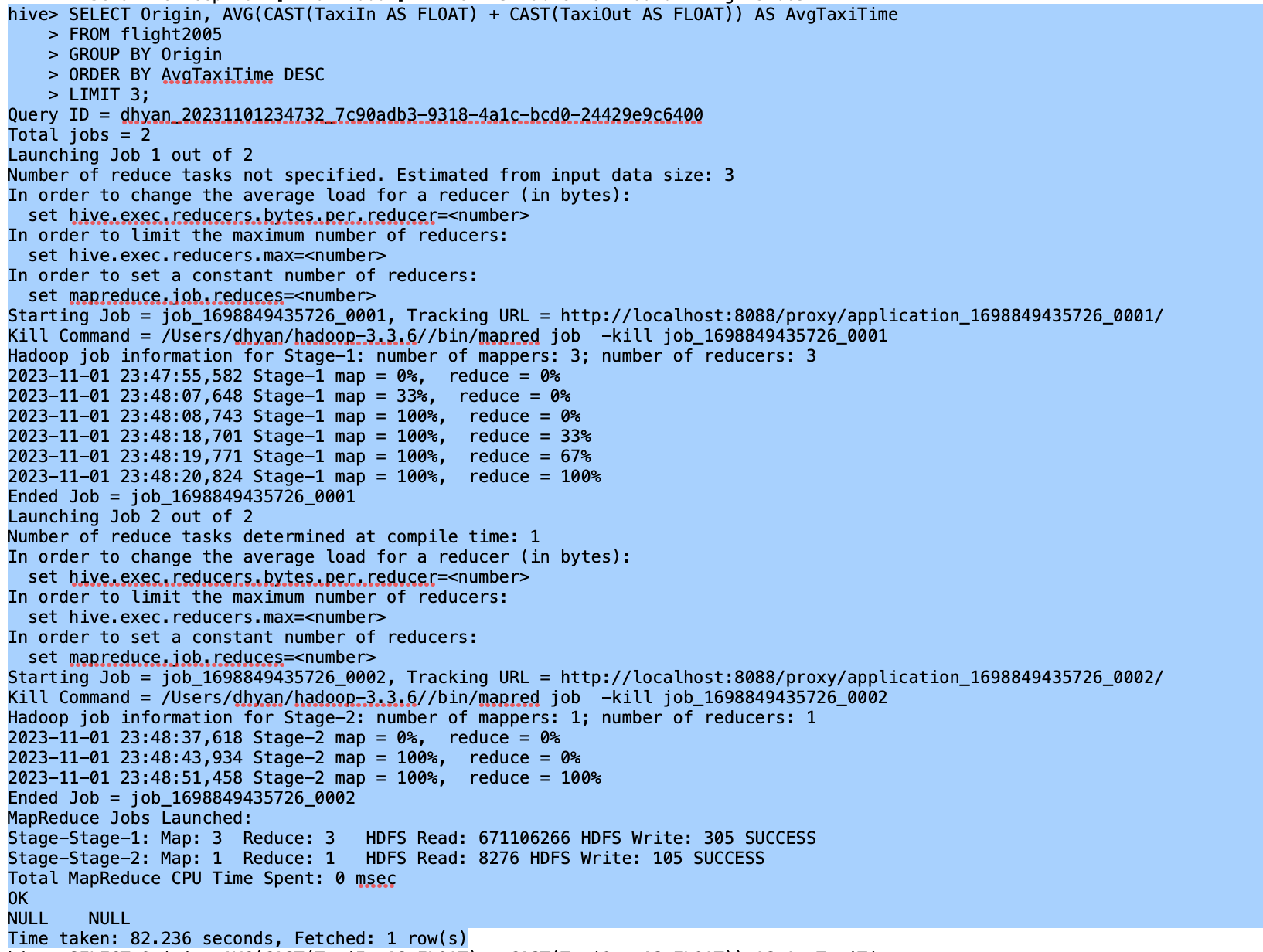
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.

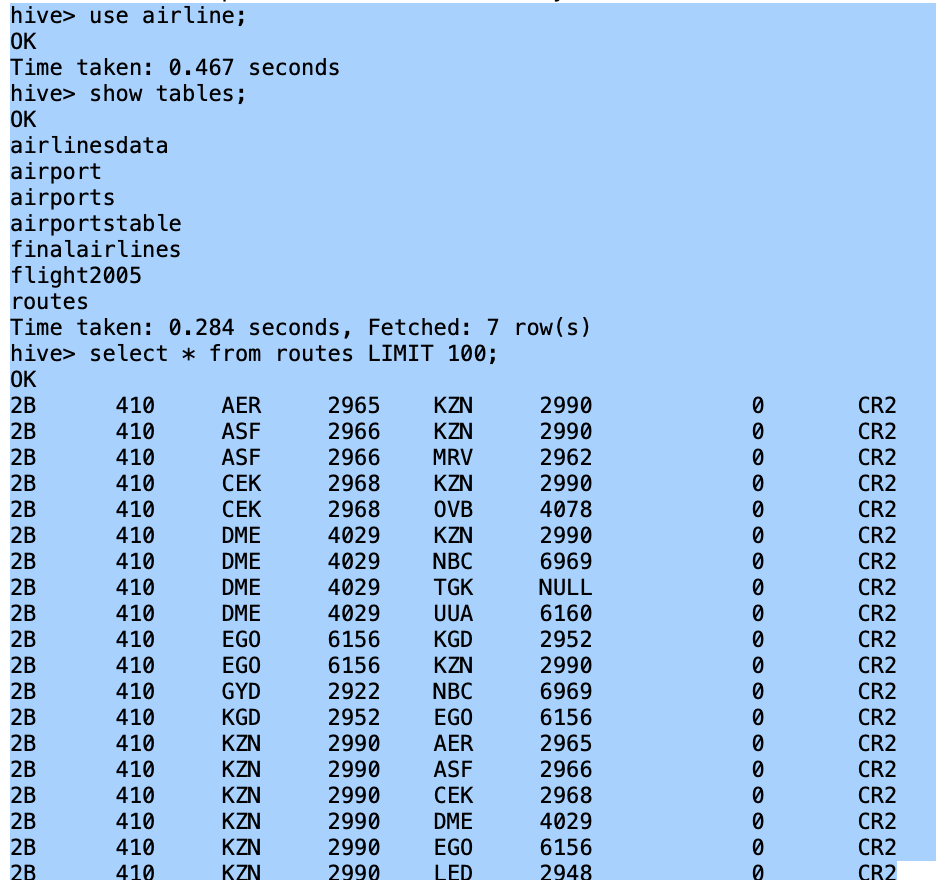
Hive Session ID = 0e94329a-b989-461b-8eec-2027a2fcccdb

Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

hive>

**Implementation :**

****

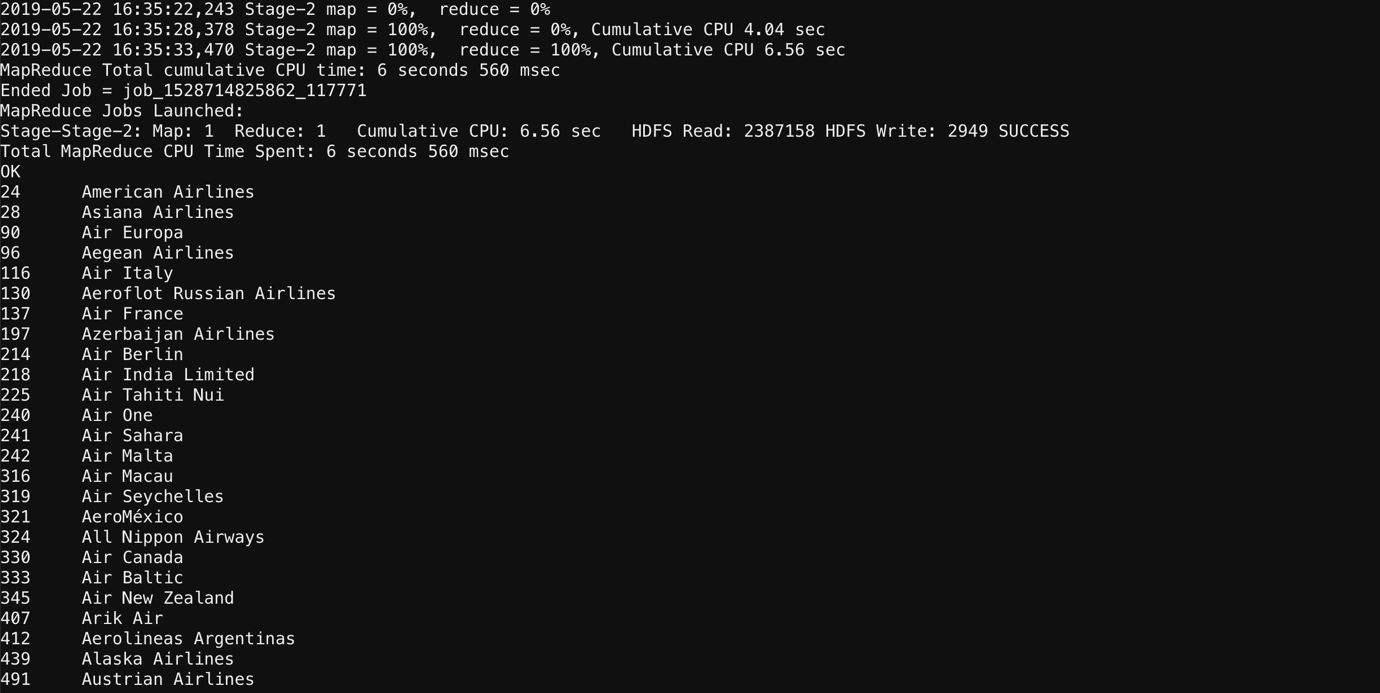
****

**select distinct(a.airlineid),a.name**

**from finalairlines a join routes on**

**a.airlineid=routes.airlineid**

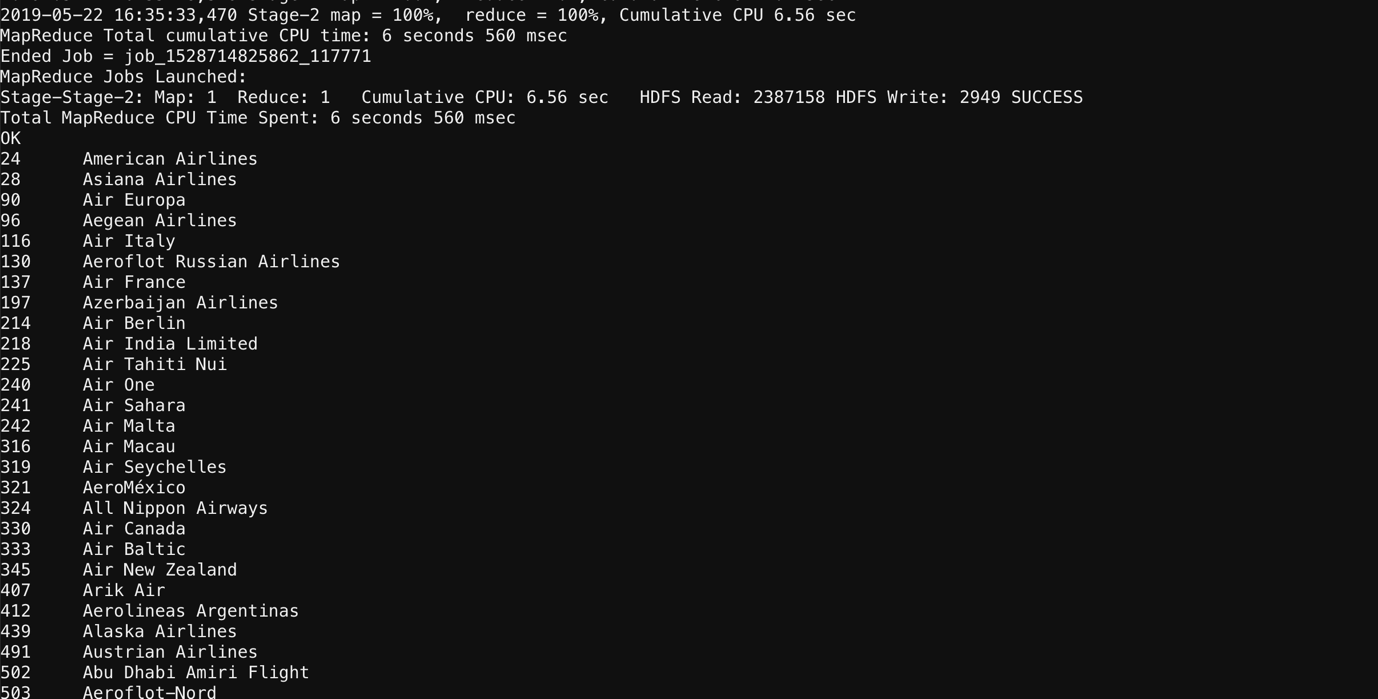
**where stops=0**



**select airlineid, name**

**from finalairlines**

**where country="United States" AND active="Y"**



**Applications:**

**1. Airline Operations Optimization:**

- The analysis of airports, airlines, and routes can aid in optimizing airline operations. Insights into zero-stop flights and code-share agreements can inform decisions to streamline routes and improve overall efficiency.

**2. Market Research for Airlines:**

- Understanding the active airlines in the United States and identifying key players provides valuable market research data. Airlines can use this information to assess competition and strategize market expansion.

**3. Government Planning and Infrastructure Development:**

- Knowledge of the distribution of airports and the country with the highest number of airports is essential for government planning and infrastructure development. It can guide decisions on where to invest in airport facilities and transportation networks.

**4. Travel and Tourism Industry Insights:**

- The data analysis can offer insights into travel patterns, helping the travel and tourism industry understand popular routes, airlines, and destinations. This information can guide marketing strategies and service offerings.

**Conclusion:**

In conclusion, the Airlines Analysis Hadoop project provides a comprehensive exploration of aviation data, leveraging the capabilities of Hadoop, MapReduce, and Hive. By addressing specific tasks such as listing airports in India, identifying zero-stop flights, and analyzing active airlines in the United States, the project aims to offer valuable insights for various stakeholders in the aviation industry.

The use of Hadoop technologies ensures efficient processing of large datasets, enabling a deeper understanding of the complexities within the aviation landscape. As the report unfolds, it delves into the challenges faced during the analysis, the technology stack employed, and the expected outputs. Ultimately, the project seeks to contribute meaningfulinformation forinformed decision-making, strategic planning, and optimization in the dynamic and competitive field of aviation.