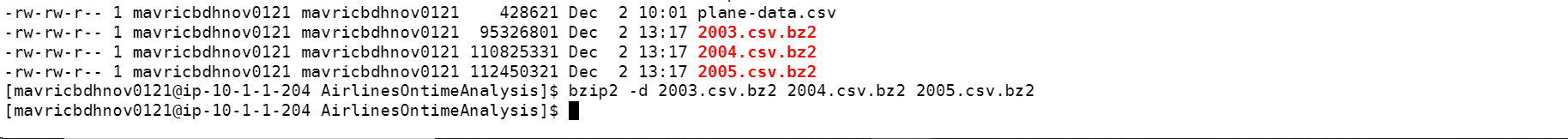
Step 1: Copy the provided csv files in a directory named AirlinesOntimeAnalysis

A black text with red and white text

Description automatically generated

Step 2: Unzip the contents for the bz2 files



Step 3: Create the directories in hdfs where your data will reside

Hdfs dfs -mkdir -p rawdata/airline/flights

Hdfs dfs -mkdir -p rawdata/airline/planeinfo

Hdfs dfs -mkdir -p rawdata/airline/airport

Hdfs dfs -mkdir -p rawdata/airline/carriers

A close-up of a paper

Description automatically generated

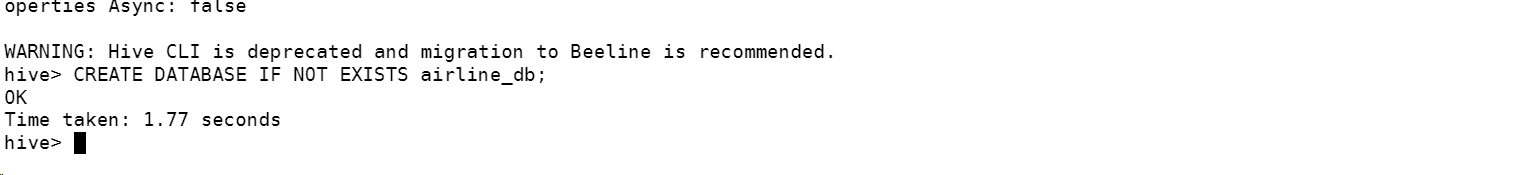
Step 4: copy the respective files to the above created hdfs structure

A close up of a white paper

Description automatically generated

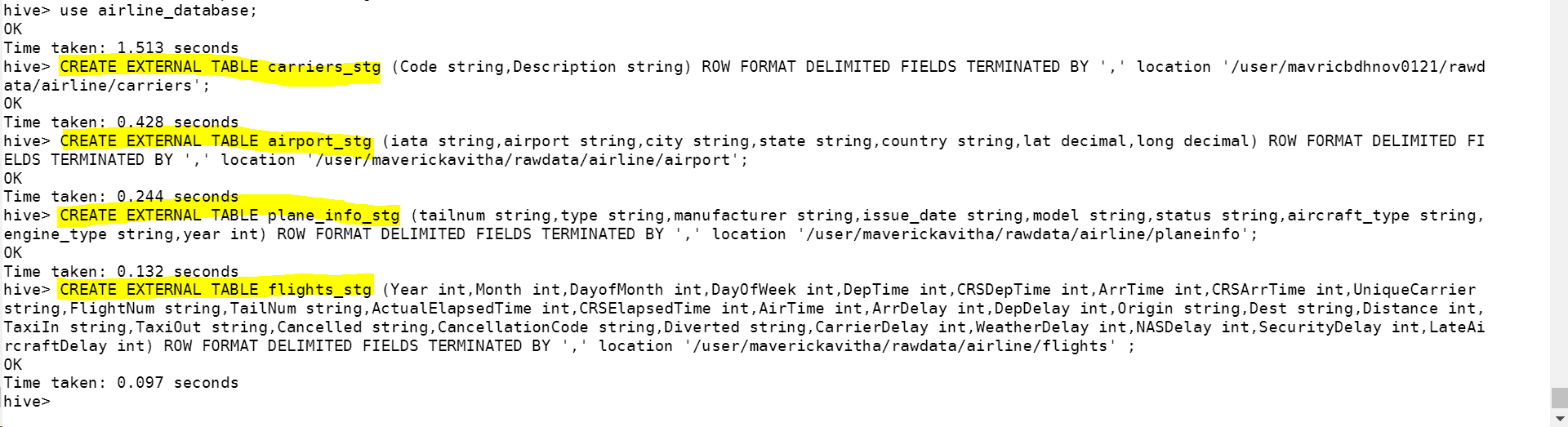
**2. Create tables using Hive**

**Create all the tables in a different database in Hive(airline\_db)**



**3.Create tables using Hive**

Create the required tables as External (flights, airports, carriers, plane\_info) in the staging area. The schema of each of the tables is provided above. Also point these tables to the hdfs locations already created above Note: Name every table name at the end with \_stg. Create all the tables in a different database in Hive(airline\_db)



Skipping headers

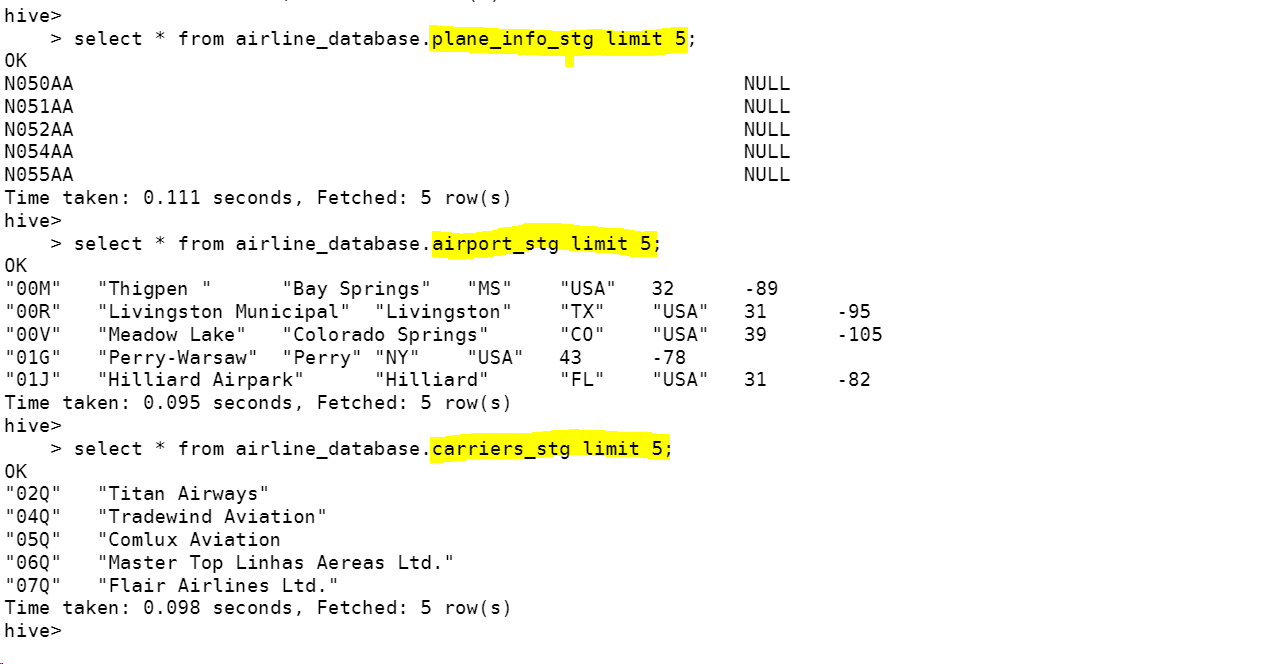
A screenshot of a computer code

Description automatically generated

Checking data:

A screenshot of a computer

Description automatically generated



**What is the average arrival delay and average departure delay in each month of the year 2004?**

A close-up of a computer screen

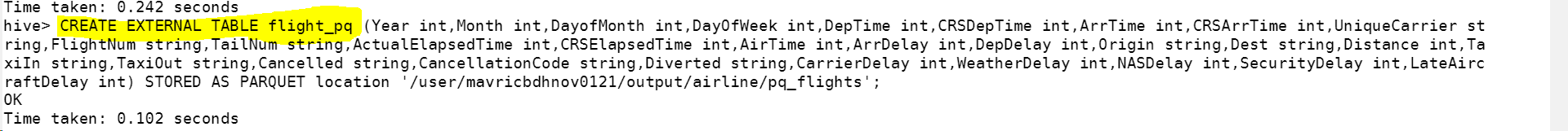
Description automatically generated

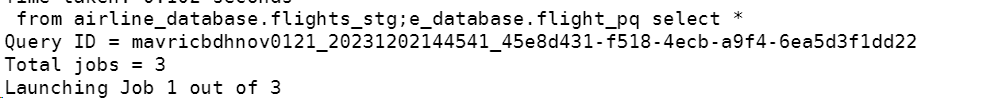
A computer screen with black text

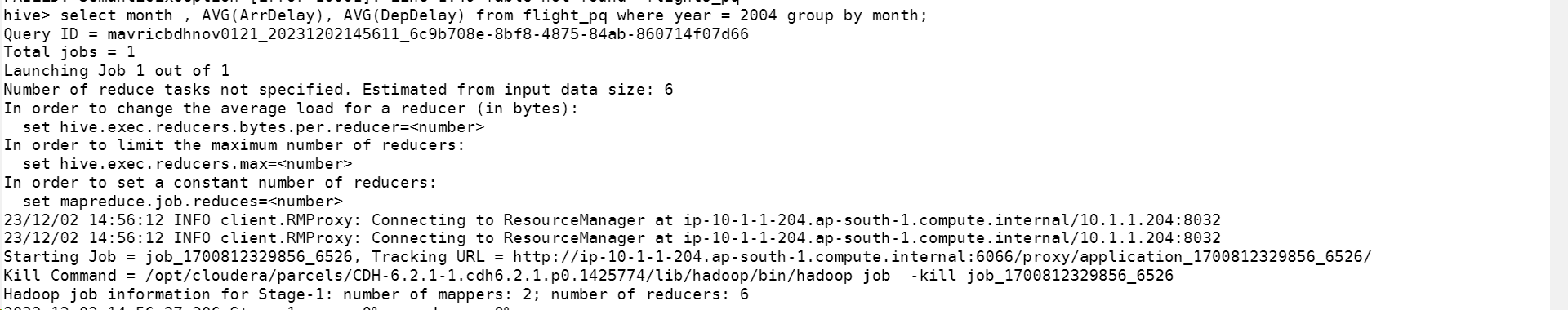
Description automatically generated

**3. Improving the performance of Hive tables - Parquet format**

Create the parquet table as external tables for flights from the above staging table. Name the new table with \_pq as a suffix: Ex : flights\_pq Point all the table location output/airline/pq\_flights







**The time is reduced from 4+ min to 1min 21 seconds.**

A white screen with black text

Description automatically generated