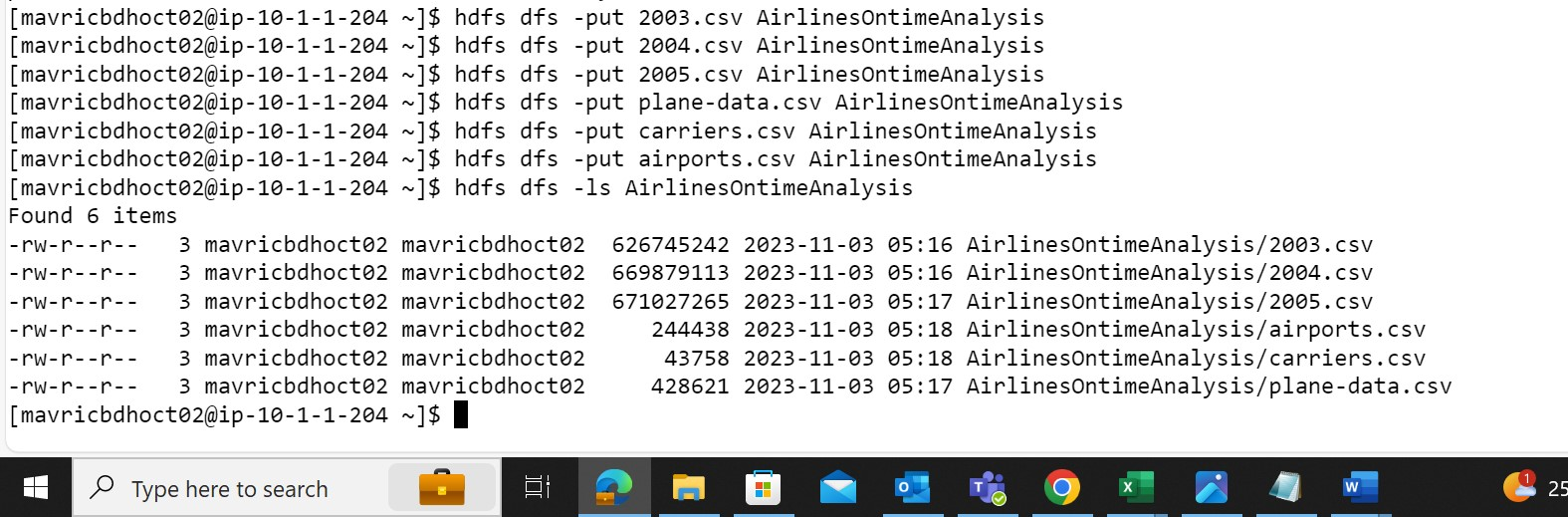
**Step 1: Copy the provided csv files in a directory named AirlinesOntimeAnalysis**

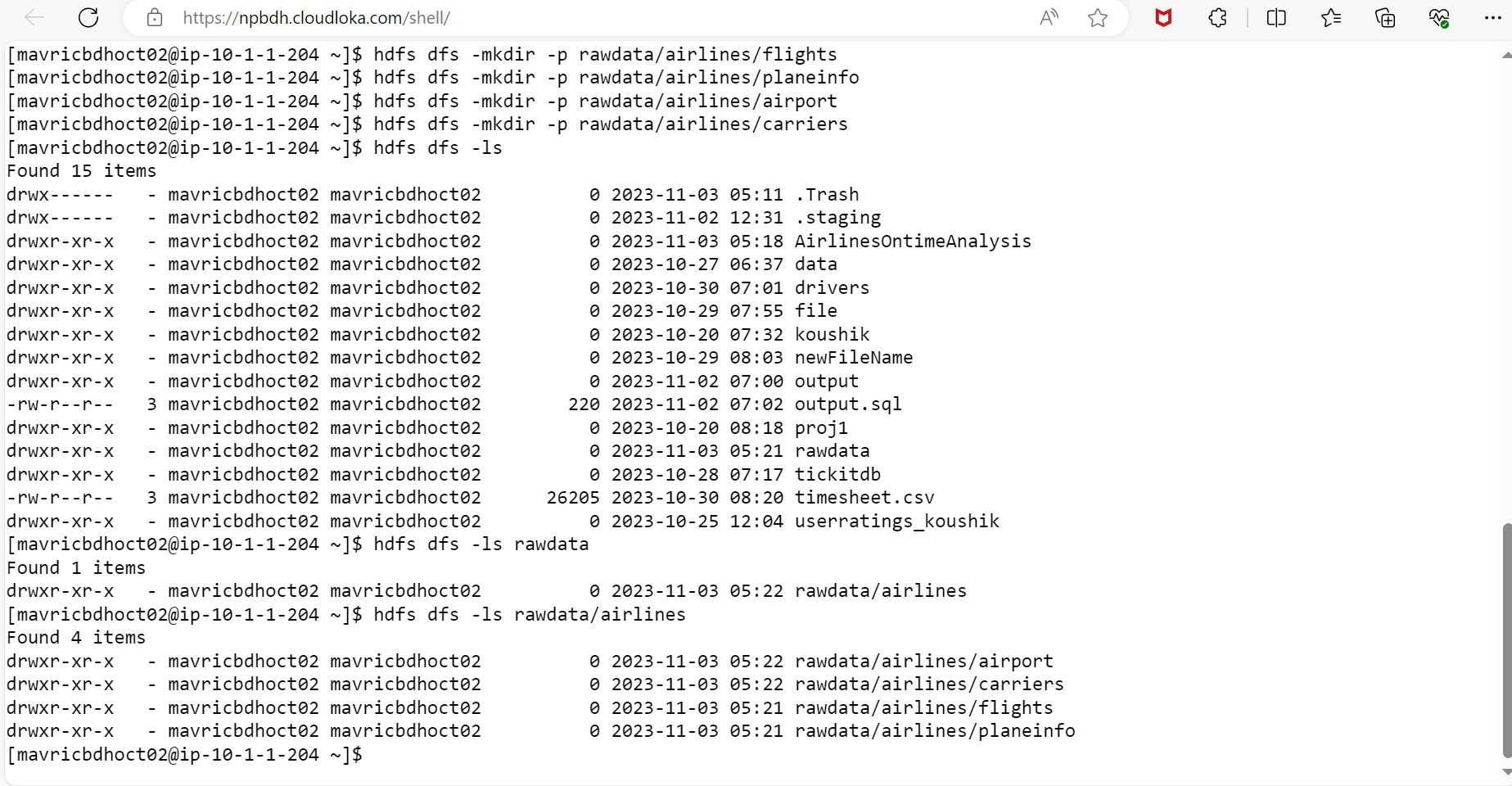


**Step 2: Unzip the contents for the bz2 files.**

A screen shot of a computer

Description automatically generated

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

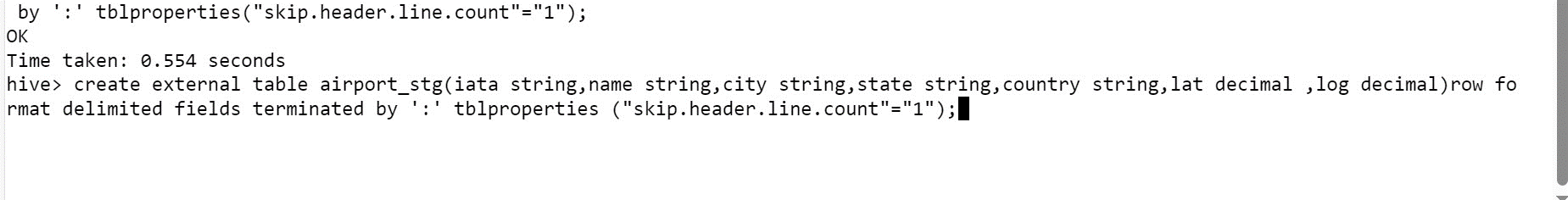


**Step 4: copy the respective files to the above created hdfs structure**

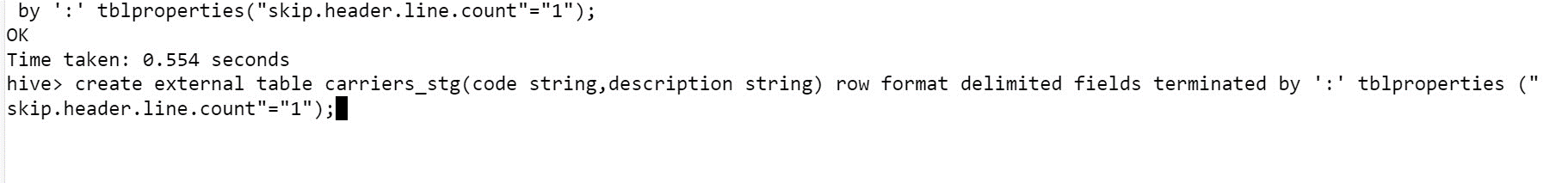
A close-up of a text

Description automatically generated

**Airports**



**Carrier**

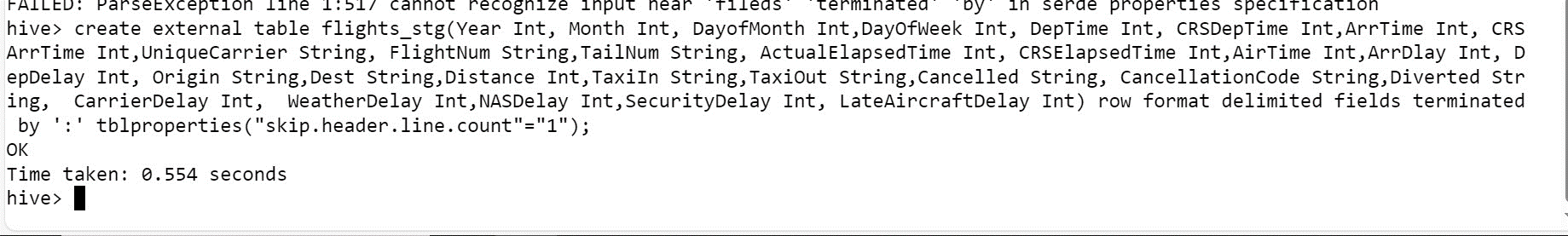


**Plane Info**

A close-up of a text

Description automatically generated

**Flights**



**Loading the data to the tables**

1. Airports

load data inpath 'AirlinesOntimeAnalysis/airports.csv' into table airport\_stg;

1. **Carriers**

load data inpath 'AirlinesOntimeAnalysis/carriers.csv' into table carriers\_stg;

1. Plane info

load data inpath 'AirlinesOntimeAnalysis/plane-data.csv' into table planeinfo\_stg;

1. **Flights**

load data inpath 'AirlinesOntimeAnalysis/2003.csv' into table flights\_stg;

load data inpath 'AirlinesOntimeAnalysis/2004.csv' into table flights\_stg;

load data inpath 'AirlinesOntimeAnalysis/2005.csv' into table flights\_stg;

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 white background with black text

Description automatically generated

A close-up of a computer screen

Description automatically generated

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

create external table flights\_pq(Year Int, Month Int, DayofMonth Int,DayOfWeek Int, DepTime Int, CRSDepTime Int,ArrTime Int, CRSA

rrTime Int,UniqueCarrier String, FlightNum String,TailNum String, ActualElapsedTime Int, CRSElapsedTime Int,AirTime Int,ArrDlay Int, De

pDelay Int, Origin String,Dest String,Distance Int,TaxiIn String,TaxiOut String,Cancelled String, CancellationCode String,Diverted Stri

ng, CarrierDelay Int, WeatherDelay Int,NASDelay Int,SecurityDelay Int, LateAircraftDelay Int) row format delimited fields terminated

by ':' stored as parquet location ‘user/mavricbdhoct02/output/airline/pq\_flight’ tblproperties("skip.header.line.count"="1");

insert overwrite table flight\_pq select \* from flight\_stg;

select month, avg(arrdelay) as avg\_arrdela

y, avg(depdelay) as avg\_depdelay from flights\_pq where year = 2004 group by month

**Is there a difference between the response times for the queries(2 and 3)**

A screenshot of a computer

Description automatically generated