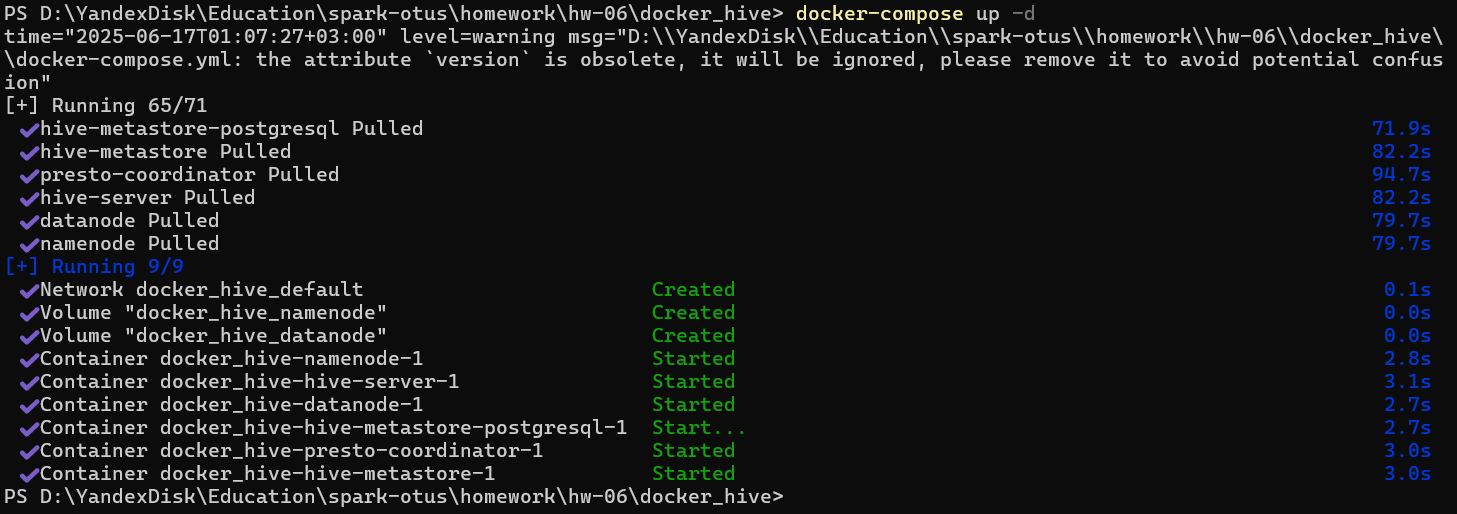
* Установка docker-container

docker-compose up -d

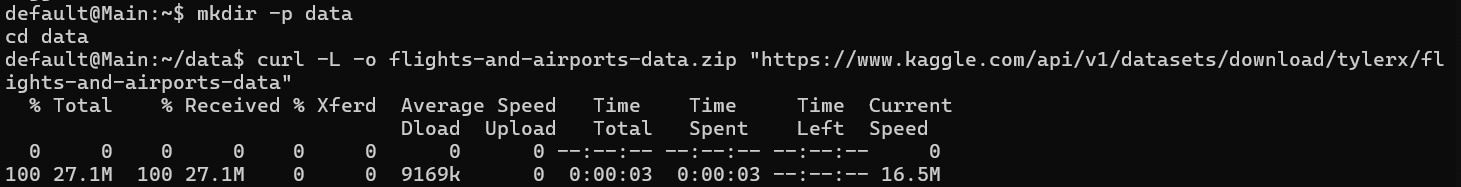


* Копирование данных с Kaggle

mkdir -p data

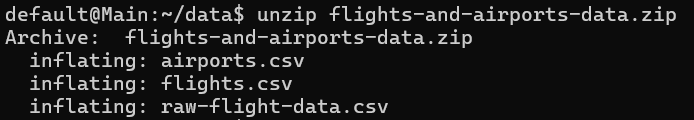
cd data

curl -L -o flights-and-airports-data.zip "https://www.kaggle.com/api/v1/datasets/download/tylerx/flights-and-airports-data"



* Распаковка архива

unzip flights-and-airports-data.zip

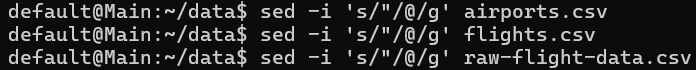


* Исправление кавычек в файлах

sed -i 's/"/@/g' airports.csv

sed -i 's/"/@/g' flights.csv

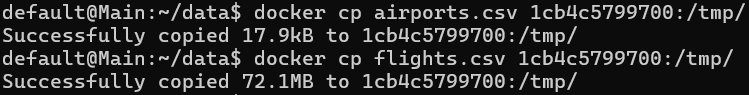
sed -i 's/"/@/g' raw-flight-data.csv



* Копирование файлов в контейнер namenode

docker cp airports.csv 1cb4c5799700:/tmp/

docker cp flights.csv 1cb4c5799700:/tmp/



* Создание директорий в HDFS

docker exec -it 1cb4c5799700 bash

hdfs dfs -mkdir -p /user/hive/warehouse/flight\_analysis.db/airports

hdfs dfs -mkdir -p /user/hive/warehouse/flight\_analysis.db/flights



* Загрузка данных в HDFS

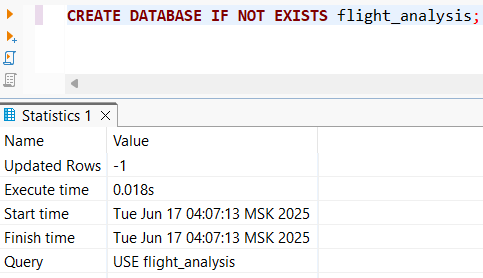
hdfs dfs -put /tmp/airports.csv /user/hive/warehouse/flight\_analysis.db/airports/

hdfs dfs -put /tmp/flights.csv /user/hive/warehouse/flight\_analysis.db/flights/



* Создание БД «flight\_analysis»

**CREATE** **DATABASE** **IF** **NOT** **EXISTS** flight\_analysis;



* Создание таблицы «airports»

**CREATE** **EXTERNAL** **TABLE** airports (

airport\_id **INT**,

city **STRING**,

state **STRING**,

name **STRING**

)

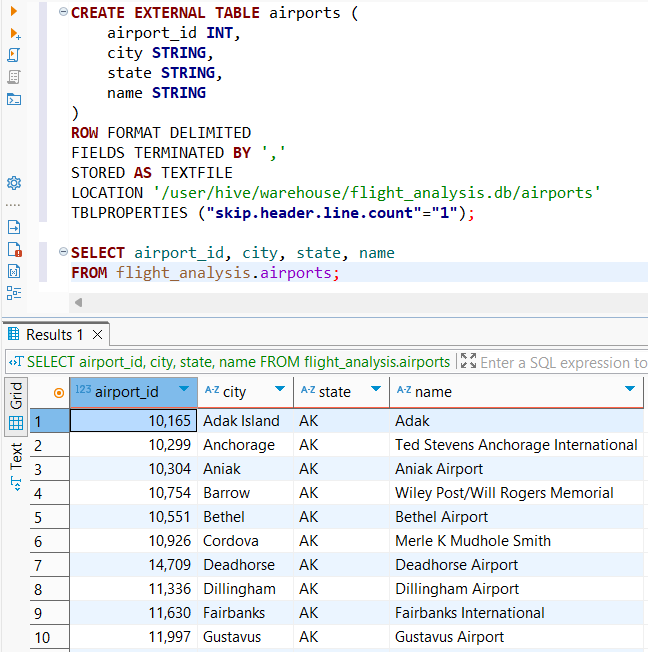
**ROW** FORMAT DELIMITED

FIELDS TERMINATED **BY** ','

STORED **AS** TEXTFILE

LOCATION '/user/hive/warehouse/flight\_analysis.db/airports'

TBLPROPERTIES (**"skip.header.line.count"**=**"1"**);



* Создание таблицы «flights»

**CREATE** **EXTERNAL** **TABLE** flights (

day\_of\_month **INT**,

day\_of\_week **INT**,

carrier **STRING**,

origin\_airport\_id **INT**,

dest\_airport\_id **INT**,

dep\_delay **INT**,

arr\_delay **INT**

)

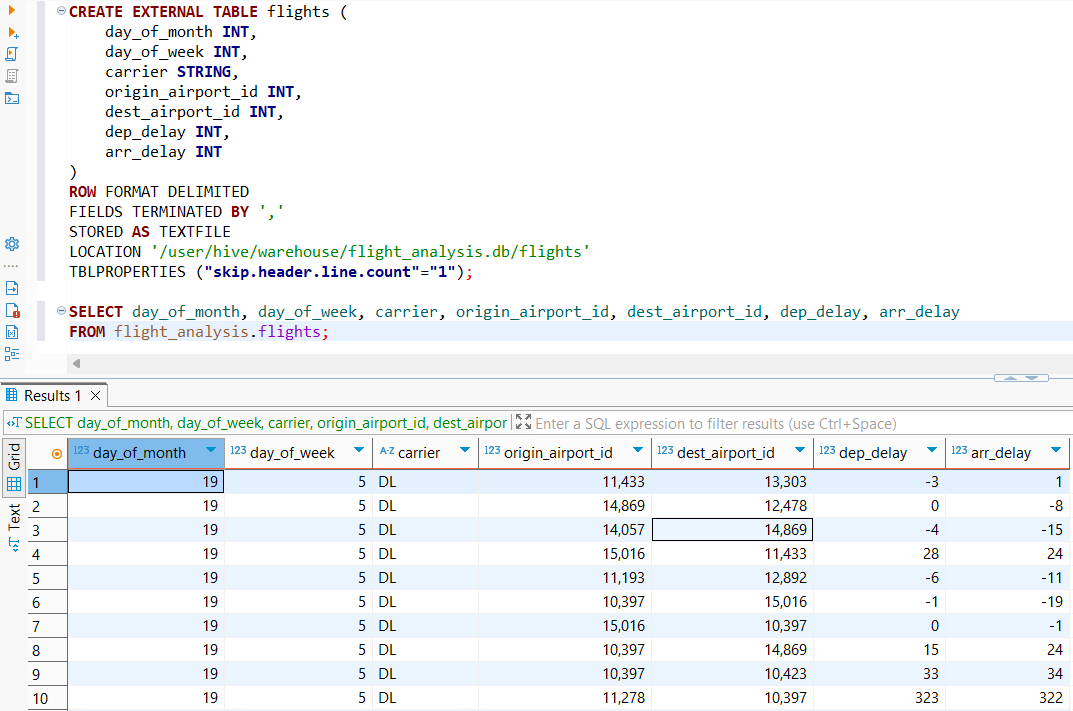
**ROW** FORMAT DELIMITED

FIELDS TERMINATED **BY** ','

STORED **AS** TEXTFILE

LOCATION '/user/hive/warehouse/flight\_analysis.db/flights'

TBLPROPERTIES (**"skip.header.line.count"**=**"1"**);



* Создание витрины с количеством рейсов по авиакомпаниям

**CREATE** VIEW airline\_flight\_counts **AS**

**SELECT**

carrier,

**COUNT**(\*) **as** flight\_count

**FROM**

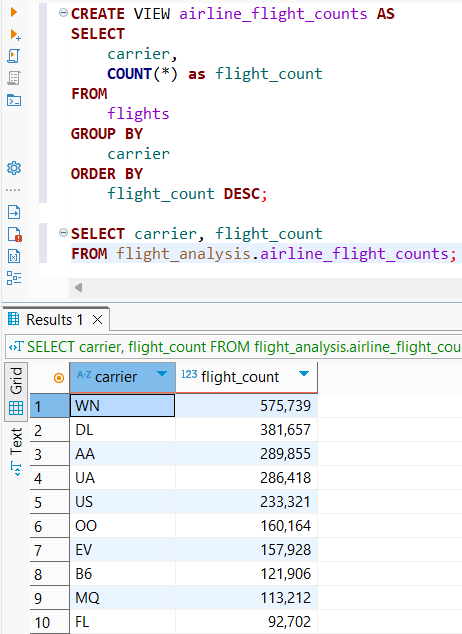
flights

**GROUP** **BY**

carrier

**ORDER** **BY**

flight\_count **DESC**;



* Создание витрины со средней задержкой по аэропортам отправления

**CREATE** VIEW avg\_departure\_delay\_by\_airport **AS**

**SELECT**

a.airport\_id,

a.name **as** airport\_name,

a.city,

**AVG**(f.dep\_delay) **as** avg\_dep\_delay

**FROM**

flights f

**JOIN**

airports a **ON** f.origin\_airport\_id = a.airport\_id

**GROUP** **BY**

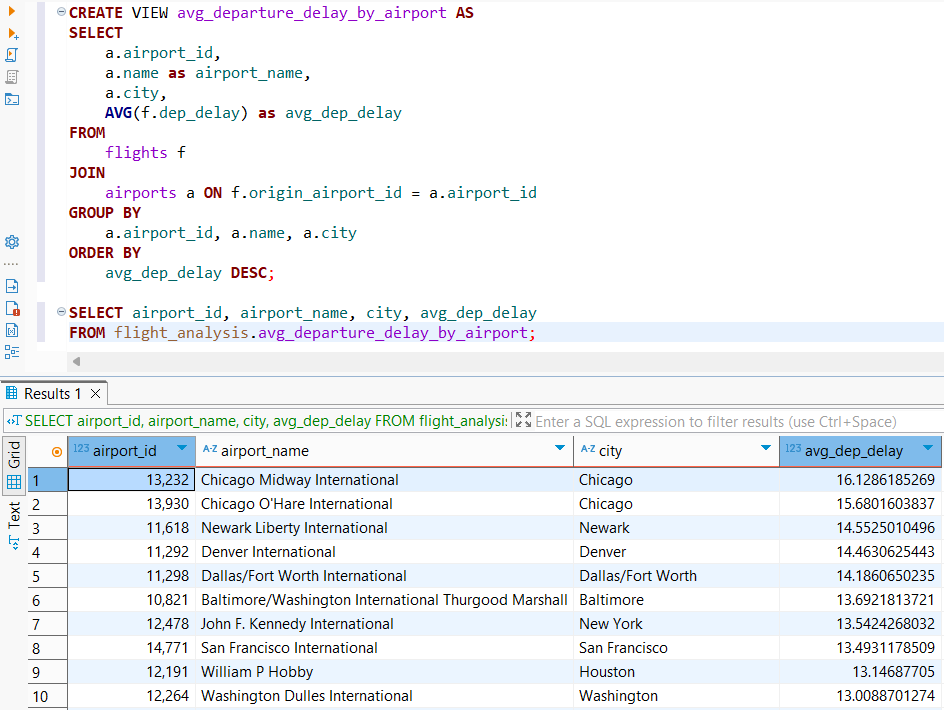
a.airport\_id, a.name, a.city

**ORDER** **BY**

avg\_dep\_delay **DESC**;

**SELECT** airport\_id, airport\_name, city, avg\_dep\_delay

**FROM** flight\_analysis.avg\_departure\_delay\_by\_airport;



* Создание витрины с 10-ю самых популярных маршрутов

**CREATE** VIEW popular\_routes **AS**

**SELECT**

a1.name **as** origin\_airport,

a2.name **as** dest\_airport,

**COUNT**(\*) **as** flight\_count

**FROM**

flights f

**JOIN**

airports a1 **ON** f.origin\_airport\_id = a1.airport\_id

**JOIN**

airports a2 **ON** f.dest\_airport\_id = a2.airport\_id

**GROUP** **BY**

a1.name, a2.name

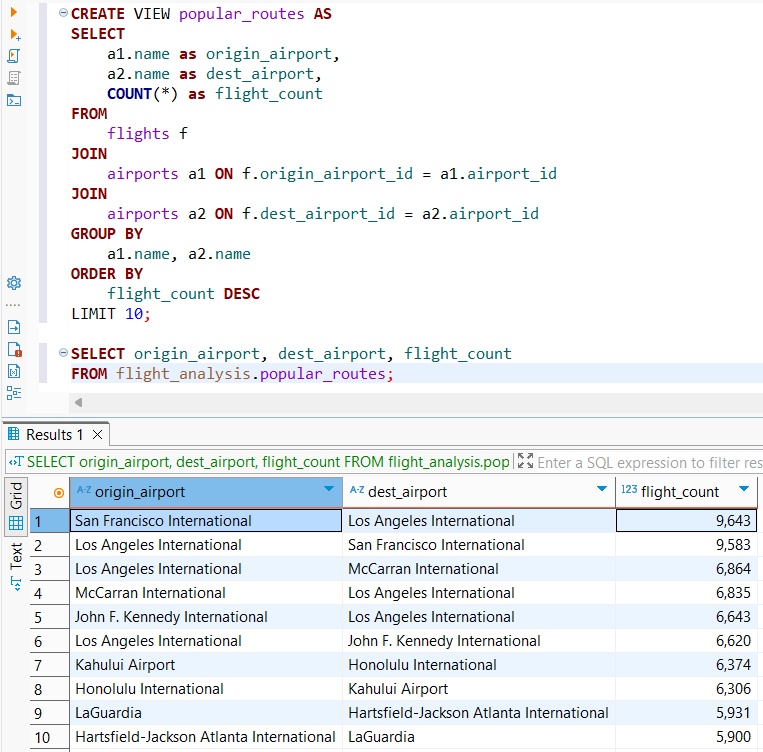
**ORDER** **BY**

flight\_count **DESC**

LIMIT 10;

**SELECT** origin\_airport, dest\_airport, flight\_count

**FROM** flight\_analysis.popular\_routes;



* Создание витрины с авиакомпаниями, к которых наибольшие задержки

**CREATE** VIEW airlines\_with\_delays **AS**

**SELECT**

carrier,

**AVG**(dep\_delay) **as** avg\_dep\_delay,

**AVG**(arr\_delay) **as** avg\_arr\_delay,

**COUNT**(\*) **as** total\_flights

**FROM**

flights

**GROUP** **BY**

carrier

**HAVING**

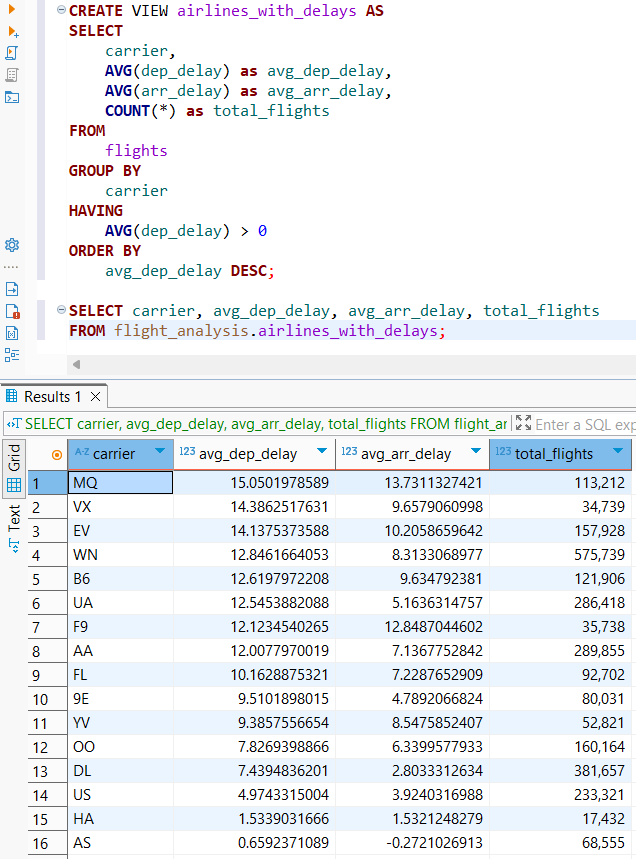
**AVG**(dep\_delay) > 0

**ORDER** **BY**

avg\_dep\_delay **DESC**;

**SELECT** carrier, avg\_dep\_delay, avg\_arr\_delay, total\_flights

**FROM** flight\_analysis.airlines\_with\_delays;



* Создание витрины с задержками по дням недели

**CREATE** VIEW delays\_by\_weekday **AS**

**SELECT**

day\_of\_week,

**AVG**(dep\_delay) **as** avg\_dep\_delay,

**AVG**(arr\_delay) **as** avg\_arr\_delay,

**COUNT**(\*) **as** total\_flights

**FROM**

flights

**GROUP** **BY**

day\_of\_week

**ORDER** **BY**

day\_of\_week;

**SELECT** day\_of\_week, avg\_dep\_delay, avg\_arr\_delay, total\_flights

**FROM** flight\_analysis.delays\_by\_weekday;

