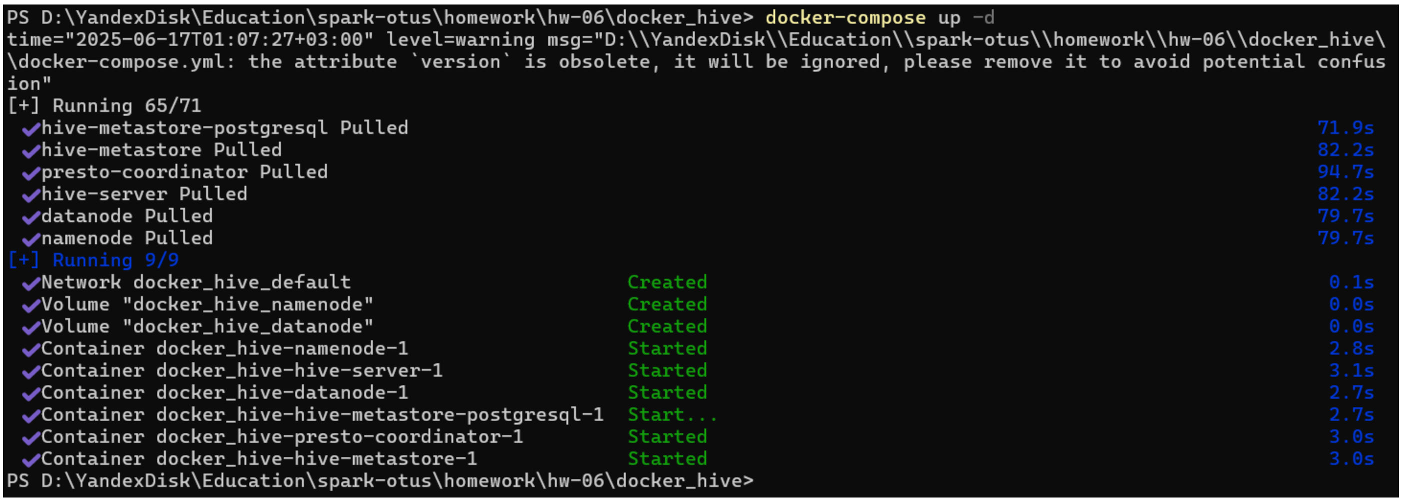
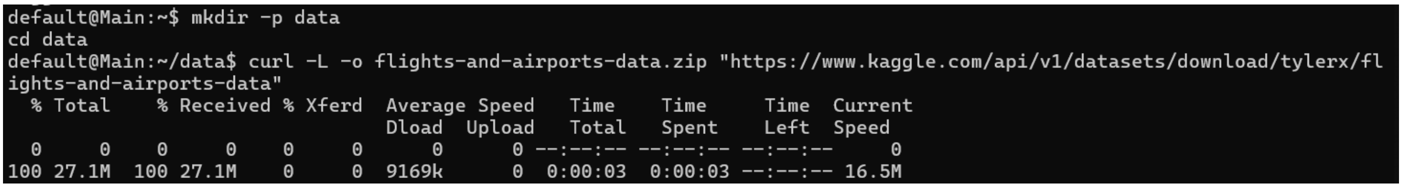
- Установка docker-container
  `docker-compose up -d`



```
PS D:\YandexDisk\Education\spark-otus\homework\hw-06\docker_hive> docker-compose up -d
time="2025-06-17T01:07:27+03:00" level=warning msg="D:\\YandexDisk\\Education\\spark-otus\\homework\\hw-06\\docker_hive\
\docker-compose.yml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confus
ion"
[+] Running 65/71
 ✔hive-metastore-postgresql Pulled                                                                              71.9s
 ✔hive-metastore Pulled                                                                                         82.2s
 ✔presto-coordinator Pulled                                                                                     94.7s
 ✔hive-server Pulled                                                                                            82.2s
 ✔datanode Pulled                                                                                               79.7s
 ✔namenode Pulled                                                                                               79.7s
[+] Running 9/9
 ✔Network docker_hive_default                      Created                                                       0.1s
 ✔Volume "docker_hive_namenode"                    Created                                                       0.0s
 ✔Volume "docker_hive_datanode"                    Created                                                       0.0s
 ✔Container docker_hive-namenode-1                 Started                                                       2.8s
 ✔Container docker_hive-hive-server-1              Started                                                       3.1s
 ✔Container docker_hive-datanode-1                 Started                                                       2.7s
 ✔Container docker_hive-hive-metastore-postgresql-1  Start...                                                    2.7s
 ✔Container docker_hive-presto-coordinator-1       Started                                                       3.0s
 ✔Container docker_hive-hive-metastore-1           Started                                                       3.0s
PS D:\YandexDisk\Education\spark-otus\homework\hw-06\docker_hive>
```

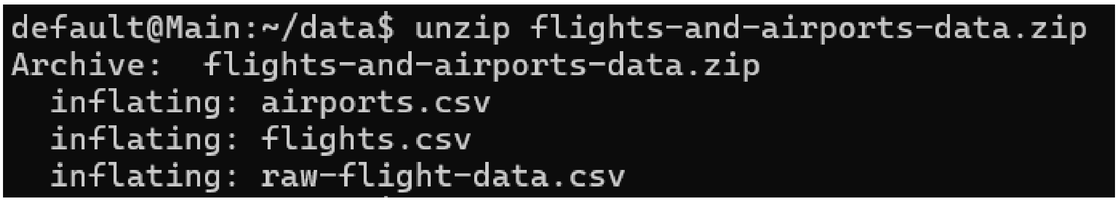
- Копирование данных с Kaggle
  ```
  mkdir -p data
  cd data
  curl -L -o flights-and-airports-data.zip
  "https://www.kaggle.com/api/v1/datasets/download/tylerx/flights-and-airports-data"
  ```



```
default@Main:~$ mkdir -p data
cd data
default@Main:~/data$ curl -L -o flights-and-airports-data.zip "https://www.kaggle.com/api/v1/datasets/download/tylerx/fl
ights-and-airports-data"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
100 27.1M  100 27.1M    0     0  9169k      0  0:00:03  0:00:03 --:--:-- 16.5M
```

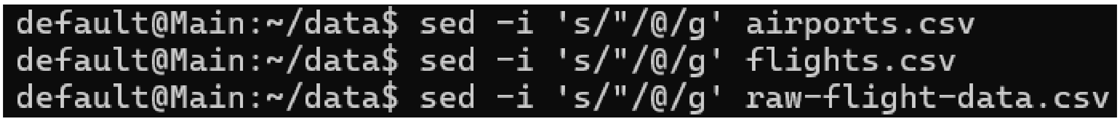
- Распаковка архива

  `unzip flights-and-airports-data.zip`



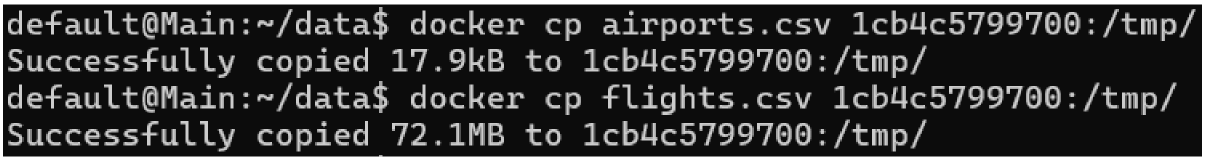
```
default@Main:~/data$ unzip flights-and-airports-data.zip
Archive:  flights-and-airports-data.zip
  inflating: airports.csv
  inflating: flights.csv
  inflating: raw-flight-data.csv
```

- Исправление кавычек в файлах
  ```
  sed -i 's/"/@/g' airports.csv
  sed -i 's/"/@/g' flights.csv
  sed -i 's/"/@/g' raw-flight-data.csv
  ```



```
default@Main:~/data$ sed -i 's/"/@/g' airports.csv
default@Main:~/data$ sed -i 's/"/@/g' flights.csv
default@Main:~/data$ sed -i 's/"/@/g' raw-flight-data.csv
```

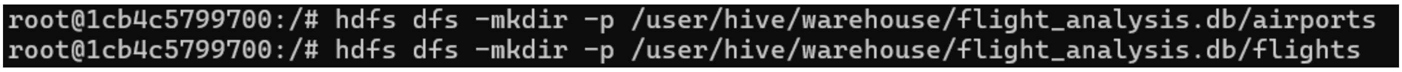
- Копирование файлов в контейнер namenode
  ```
  docker cp airports.csv 1cb4c5799700:/tmp/
  docker cp flights.csv 1cb4c5799700:/tmp/
  ```



```
default@Main:~/data$ docker cp airports.csv 1cb4c5799700:/tmp/
Successfully copied 17.9kB to 1cb4c5799700:/tmp/
default@Main:~/data$ docker cp flights.csv 1cb4c5799700:/tmp/
Successfully copied 72.1MB to 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
```

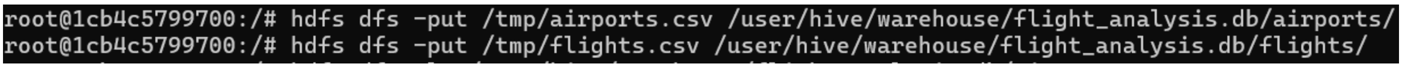


```
root@1cb4c5799700:/# hdfs dfs -mkdir -p /user/hive/warehouse/flight_analysis.db/airports
root@1cb4c5799700:/# hdfs dfs -mkdir -p /user/hive/warehouse/flight_analysis.db/flights
```

- Загрузка данных в HDFS

hdfs dfs -put /tmp/airports.csv /user/hive/warehouse/flight_analysis.db/airports/

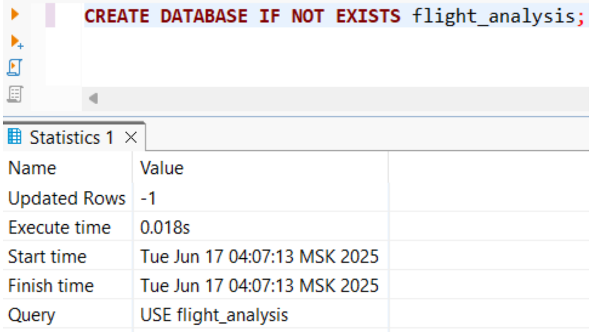
hdfs dfs -put /tmp/flights.csv /user/hive/warehouse/flight_analysis.db/flights/



```
root@1cb4c5799700:/# hdfs dfs -put /tmp/airports.csv /user/hive/warehouse/flight_analysis.db/airports/
root@1cb4c5799700:/# hdfs dfs -put /tmp/flights.csv /user/hive/warehouse/flight_analysis.db/flights/
```

- Создание БД «flight_analysis»

CREATE DATABASE IF NOT EXISTS flight_analysis;

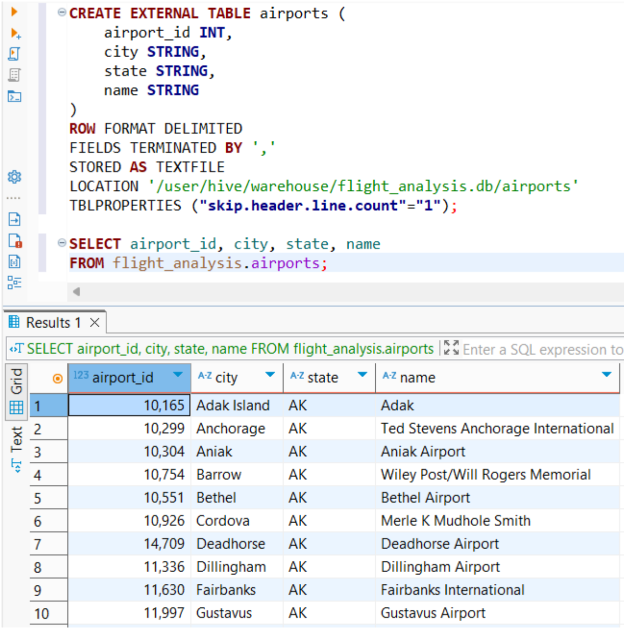


```
CREATE DATABASE IF NOT EXISTS flight_analysis;
```

Name	Value
Updated Rows	-1
Execute time	0.018s
Start time	Tue Jun 17 04:07:13 MSK 2025
Finish time	Tue Jun 17 04:07:13 MSK 2025
Query	USE 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");
```



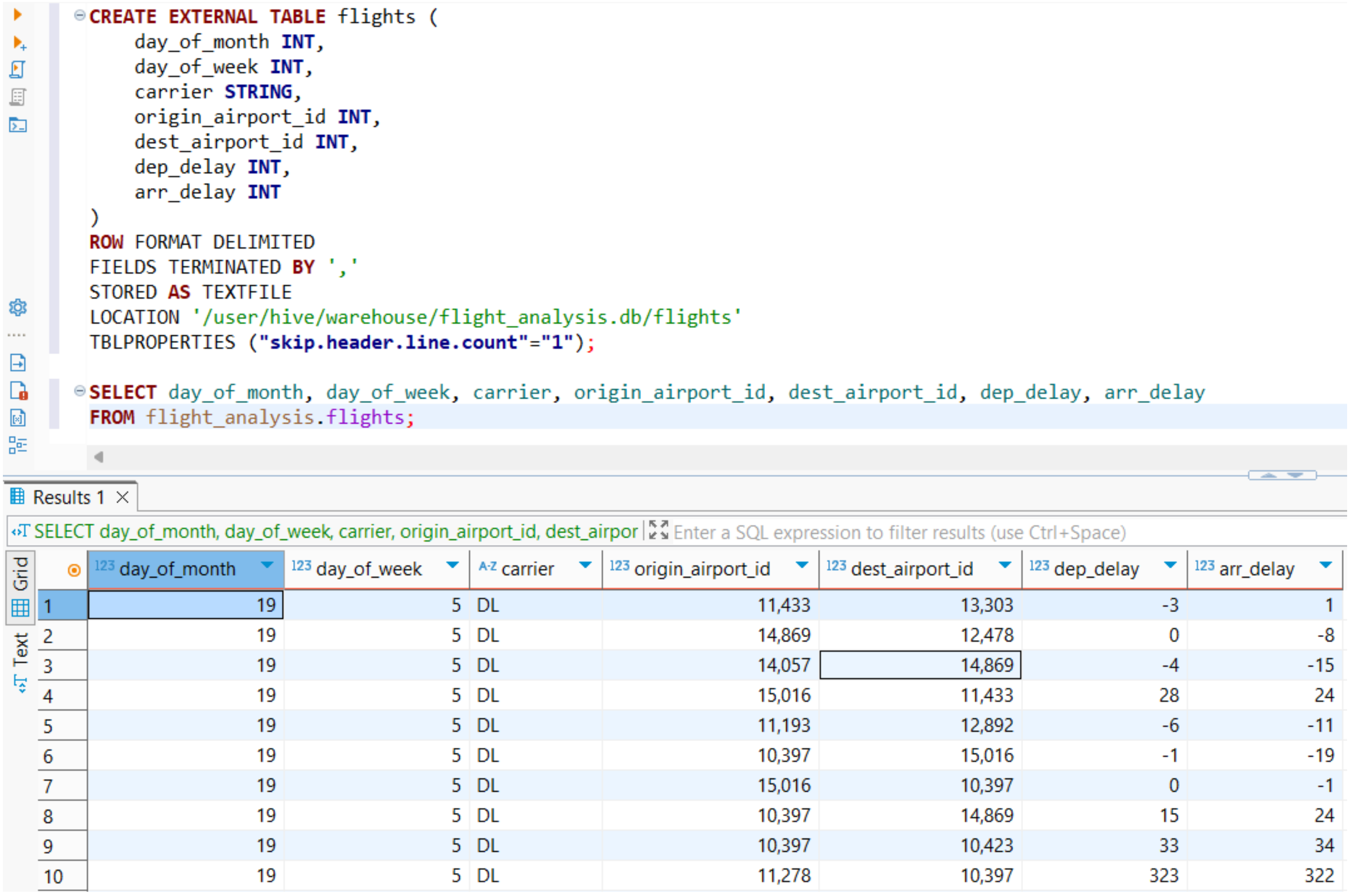
```
CREATE EXTERNAL TABLE airports (
    airport_id INT,
    city STRING,
    state STRING,
    name STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/flight_analysis.db/airports'
TBLPROPERTIES ("skip.header.line.count"="1");

SELECT airport_id, city, state, name
FROM flight_analysis.airports;
```

Results 1
SELECT airport_id, city, state, name FROM flight_analysis.airports
Enter a SQL expression to
Grid
Text

	airport_id	city	state	name
1	10,165	Adak Island	AK	Adak
2	10,299	Anchorage	AK	Ted Stevens Anchorage International
3	10,304	Aniak	AK	Aniak Airport
4	10,754	Barrow	AK	Wiley Post/Will Rogers Memorial
5	10,551	Bethel	AK	Bethel Airport
6	10,926	Cordova	AK	Merle K Mudhole Smith
7	14,709	Deadhorse	AK	Deadhorse Airport
8	11,336	Dillingham	AK	Dillingham Airport
9	11,630	Fairbanks	AK	Fairbanks International
10	11,997	Gustavus	AK	Gustavus Airport

- 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");
```



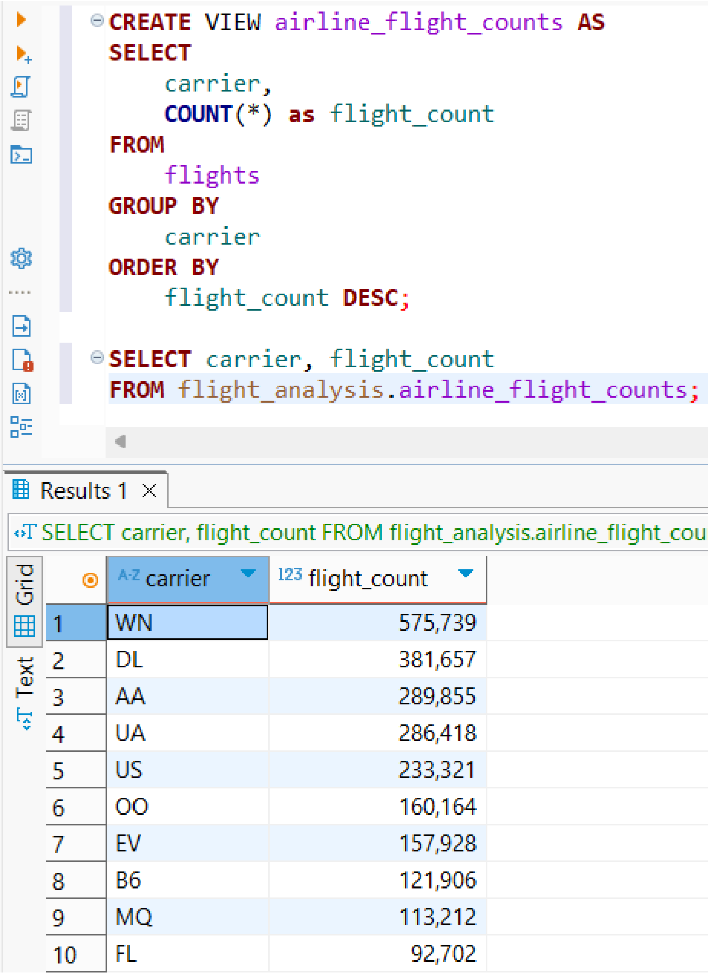
```
CREATE EXTERNAL TABLE flights (
    day_of_month INT,
    day_of_week INT,
    carrier STRING,
    origin_airport_id INT,
    dest_airport_id INT,
    dep_delay INT,
    arr_delay INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/flight_analysis.db/flights'
TBLPROPERTIES ("skip.header.line.count"="1");

SELECT day_of_month, day_of_week, carrier, origin_airport_id, dest_airport_id, dep_delay, arr_delay
FROM flight_analysis.flights;
```

Results 1
SELECT day_of_month, day_of_week, carrier, origin_airport_id, dest_airpor
Enter a SQL expression to filter results (use Ctrl+Space)
Grid
Text

	day_of_month	day_of_week	carrier	origin_airport_id	dest_airport_id	dep_delay	arr_delay
1	19	5	DL	11,433	13,303	-3	1
2	19	5	DL	14,869	12,478	0	-8
3	19	5	DL	14,057	14,869	-4	-15
4	19	5	DL	15,016	11,433	28	24
5	19	5	DL	11,193	12,892	-6	-11
6	19	5	DL	10,397	15,016	-1	-19
7	19	5	DL	15,016	10,397	0	-1
8	19	5	DL	10,397	14,869	15	24
9	19	5	DL	10,397	10,423	33	34
10	19	5	DL	11,278	10,397	323	322

- 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;
```



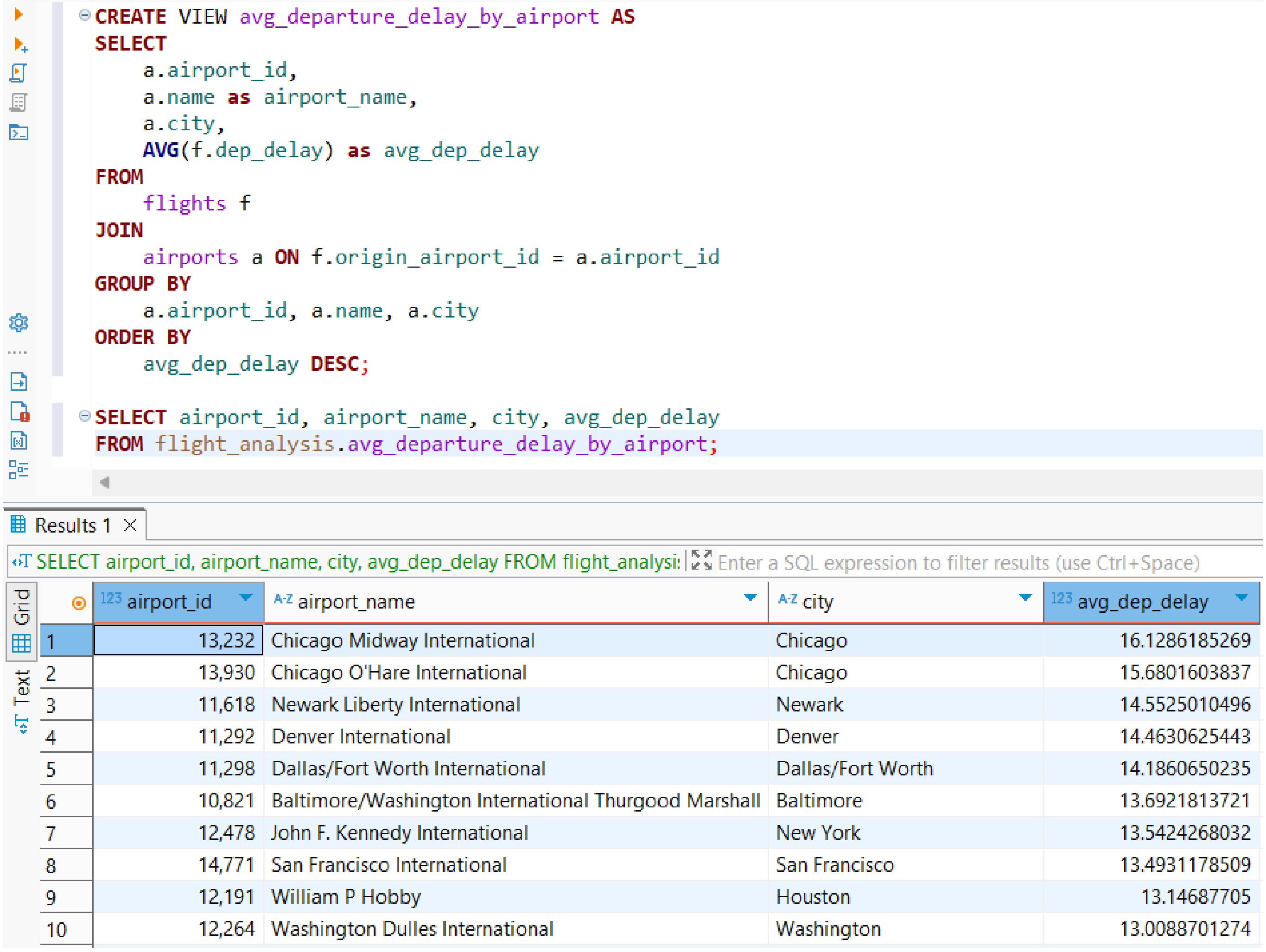
```
CREATE VIEW airline_flight_counts AS
SELECT
    carrier,
    COUNT(*) as flight_count
FROM
    flights
GROUP BY
    carrier
ORDER BY
    flight_count DESC;

SELECT carrier, flight_count
FROM flight_analysis.airline_flight_counts;
```

SELECT carrier, flight_count FROM flight_analysis.airline_flight_cou

	carrier	flight_count
1	WN	575,739
2	DL	381,657
3	AA	289,855
4	UA	286,418
5	US	233,321
6	OO	160,164
7	EV	157,928
8	B6	121,906
9	MQ	113,212
10	FL	92,702

- 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;
```



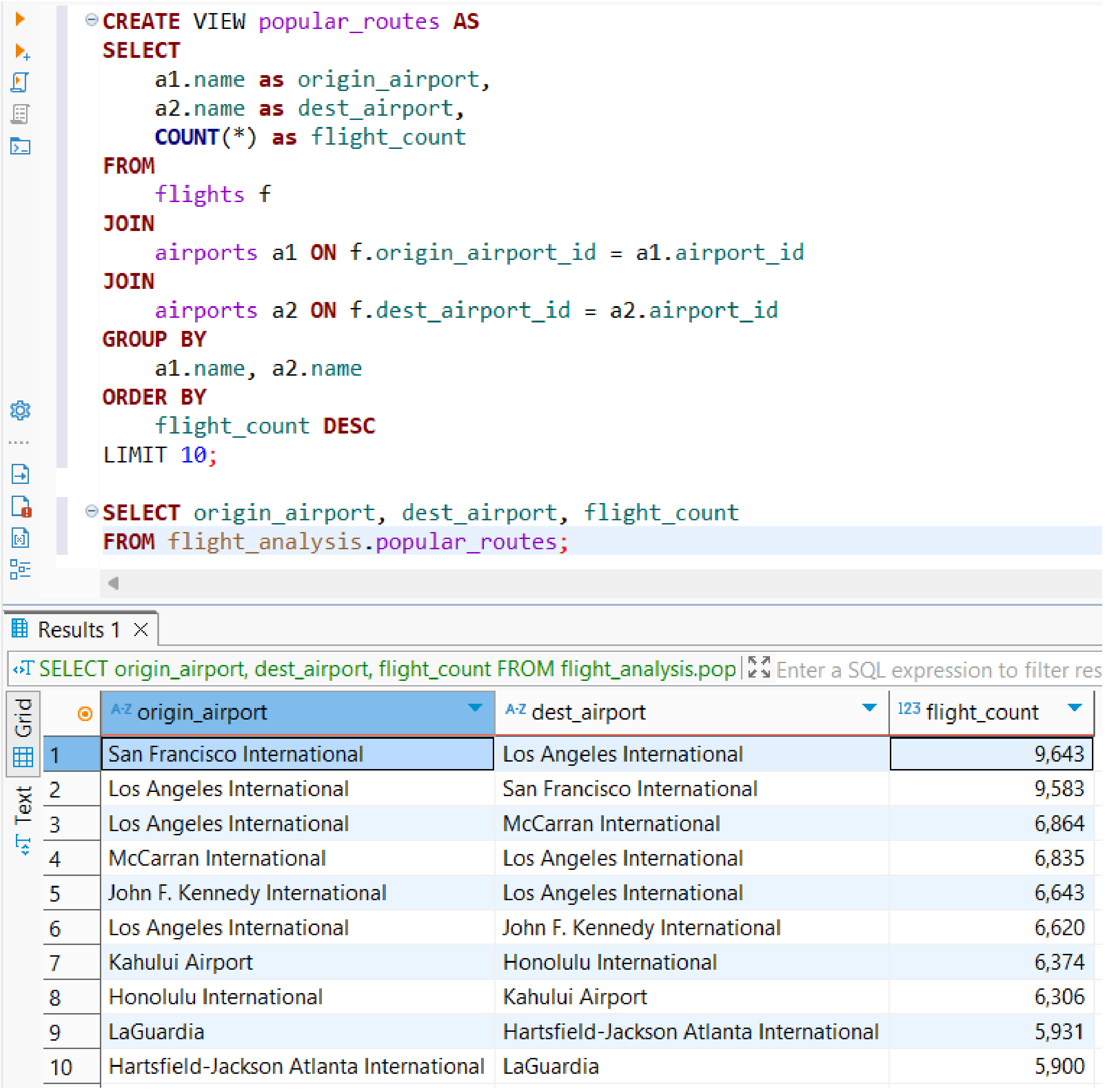
```
CREATE VIEW avg_departure_delay_by_airport AS
SELECT
    a.airport_id,
    a.name as airport_name,
    a.city,
    AVG(f.dep_delay) as avg_dep_delay
FROM
    flights f
JOIN
    airports a ON f.origin_airport_id = a.airport_id
GROUP BY
    a.airport_id, a.name, a.city
ORDER BY
    avg_dep_delay DESC;

SELECT airport_id, airport_name, city, avg_dep_delay
FROM flight_analysis.avg_departure_delay_by_airport;
```

Results 1
SELECT airport_id, airport_name, city, avg_dep_delay FROM flight_analysi:
Enter a SQL expression to filter results (use Ctrl+Space)
Grid
Text

	airport_id	airport_name	city	avg_dep_delay
1	13,232	Chicago Midway International	Chicago	16.1286185269
2	13,930	Chicago O'Hare International	Chicago	15.6801603837
3	11,618	Newark Liberty International	Newark	14.5525010496
4	11,292	Denver International	Denver	14.4630625443
5	11,298	Dallas/Fort Worth International	Dallas/Fort Worth	14.1860650235
6	10,821	Baltimore/Washington International Thurgood Marshall	Baltimore	13.6921813721
7	12,478	John F. Kennedy International	New York	13.5424268032
8	14,771	San Francisco International	San Francisco	13.4931178509
9	12,191	William P Hobby	Houston	13.14687705
10	12,264	Washington Dulles International	Washington	13.0088701274

- 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;
```



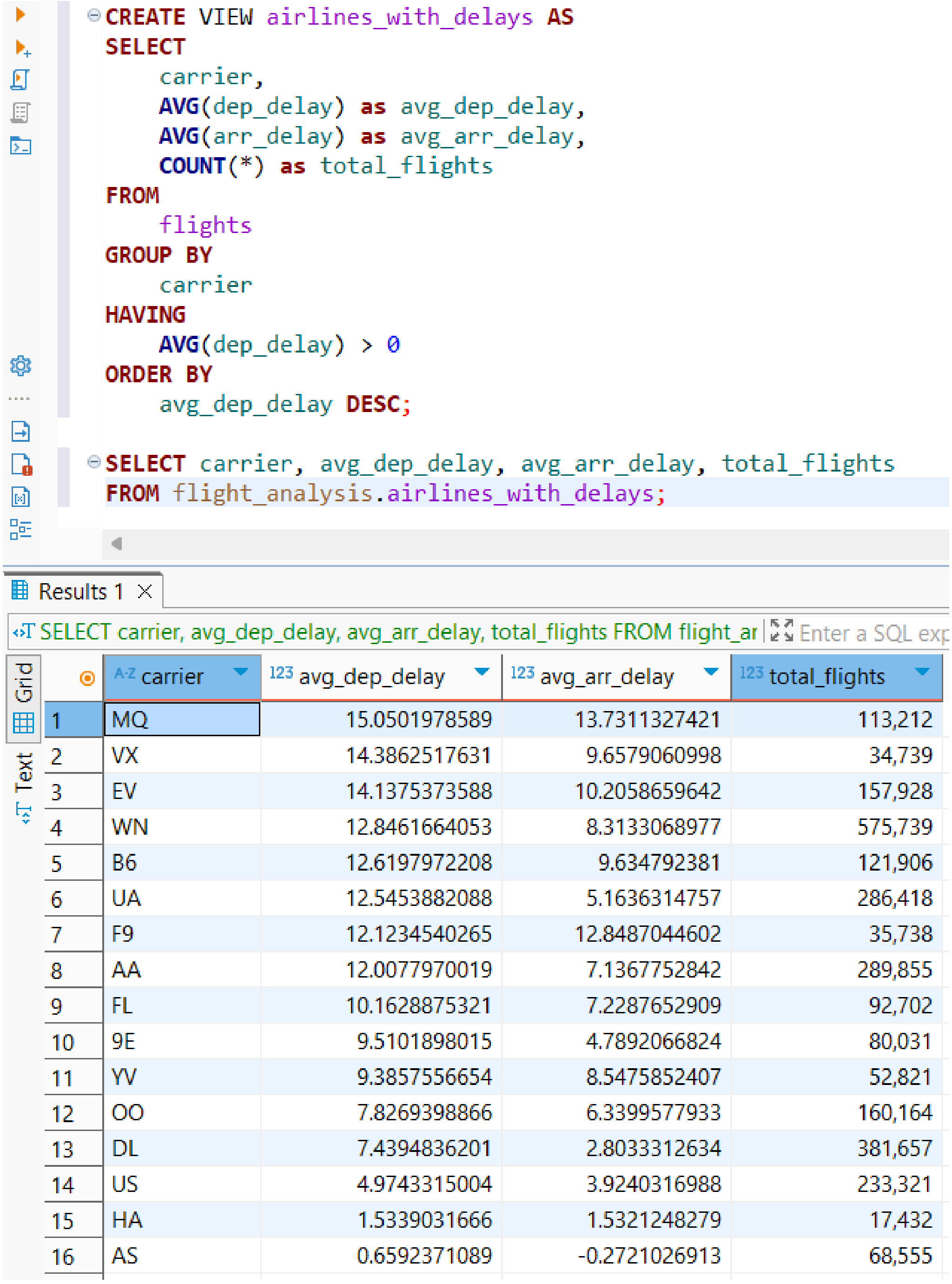
```
CREATE VIEW popular_routes AS
SELECT
    a1.name as origin_airport,
    a2.name as dest_airport,
    COUNT(*) as flight_count
FROM
    flights f
JOIN
    airports a1 ON f.origin_airport_id = a1.airport_id
JOIN
    airports a2 ON f.dest_airport_id = a2.airport_id
GROUP BY
    a1.name, a2.name
ORDER BY
    flight_count DESC
LIMIT 10;

SELECT origin_airport, dest_airport, flight_count
FROM flight_analysis.popular_routes;
```

SELECT origin_airport, dest_airport, flight_count FROM flight_analysis.pop

	origin_airport	dest_airport	flight_count
1	San Francisco International	Los Angeles International	9,643
2	Los Angeles International	San Francisco International	9,583
3	Los Angeles International	McCarran International	6,864
4	McCarran International	Los Angeles International	6,835
5	John F. Kennedy International	Los Angeles International	6,643
6	Los Angeles International	John F. Kennedy International	6,620
7	Kahului Airport	Honolulu International	6,374
8	Honolulu International	Kahului Airport	6,306
9	LaGuardia	Hartsfield-Jackson Atlanta International	5,931
10	Hartsfield-Jackson Atlanta International	LaGuardia	5,900

- 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;
```



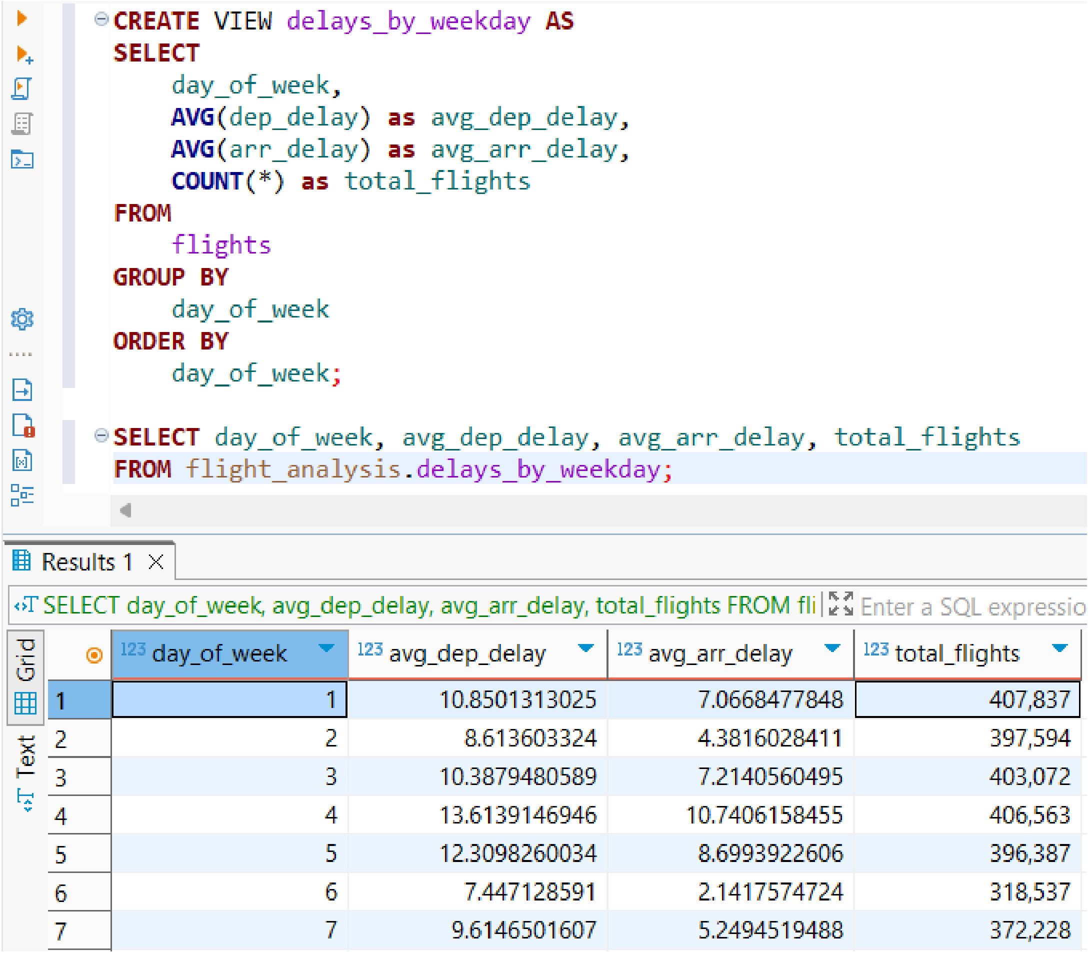
```
CREATE VIEW airlines_with_delays AS
SELECT
    carrier,
    AVG(dep_delay) as avg_dep_delay,
    AVG(arr_delay) as avg_arr_delay,
    COUNT(*) as total_flights
FROM
    flights
GROUP BY
    carrier
HAVING
    AVG(dep_delay) > 0
ORDER BY
    avg_dep_delay DESC;

SELECT carrier, avg_dep_delay, avg_arr_delay, total_flights
FROM flight_analysis.airlines_with_delays;
```

SELECT carrier, avg_dep_delay, avg_arr_delay, total_flights FROM flight_ar

	carrier	avg_dep_delay	avg_arr_delay	total_flights
1	MQ	15.0501978589	13.7311327421	113,212
2	VX	14.3862517631	9.6579060998	34,739
3	EV	14.1375373588	10.2058659642	157,928
4	WN	12.8461664053	8.3133068977	575,739
5	B6	12.6197972208	9.634792381	121,906
6	UA	12.5453882088	5.1636314757	286,418
7	F9	12.1234540265	12.8487044602	35,738
8	AA	12.0077970019	7.1367752842	289,855
9	FL	10.1628875321	7.2287652909	92,702
10	9E	9.5101898015	4.7892066824	80,031
11	YV	9.3857556654	8.5475852407	52,821
12	OO	7.8269398866	6.3399577933	160,164
13	DL	7.4394836201	2.8033312634	381,657
14	US	4.9743315004	3.9240316988	233,321
15	HA	1.5339031666	1.5321248279	17,432
16	AS	0.6592371089	-0.2721026913	68,555

- 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;
```



```
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;
```

Results 1
SELECT day_of_week, avg_dep_delay, avg_arr_delay, total_flights FROM fli
Enter a SQL expressio
Grid
Text

	day_of_week	avg_dep_delay	avg_arr_delay	total_flights
1	1	10.8501313025	7.0668477848	407,837
2	2	8.613603324	4.3816028411	397,594
3	3	10.3879480589	7.2140560495	403,072
4	4	13.6139146946	10.7406158455	406,563
5	5	12.3098260034	8.6993922606	396,387
6	6	7.447128591	2.1417574724	318,537
7	7	9.6146501607	5.2494519488	372,228