- Create hive table, *flight_data*:

```
CREATE TABLE flight_data(
   year INT,
   month INT,
   day INT,
   day_of_week INT,
   dep_time INT,
   crs_dep_time INT,
   arr_time INT,
   crs_arr_time INT,
   unique_carrier STRING,
   flight_num INT,
   tail_num STRING,
   actual_elapsed_time INT,
   crs_elapsed_time INT,
   air_time INT,
   arr_delay INT,
   dep_delay INT,
   origin STRING,
   dest STRING,
   distance INT,
   taxi_in INT,
   taxi_out INT,
   cancelled INT,
   cancellation_code STRING,
   diverted INT,
   carrier_delay STRING,
   weather_delay STRING,
   nas_delay STRING,
   security_delay STRING,
   late_aircraft_delay STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

- Load the data into the table:

```
LOAD DATA LOCAL INPATH '2008.csv' OVERWRITE INTO TABLE flight_data;
```

- Ensure the table got created and loaded fine:

```
SHOW TABLES;
SELECT
   *
FROM
   flight_data
LIMIT 10;
```

- Query the table. Find average arrival delay for all flights departing SFO in January:

```
SELECT
   avg(arr_delay)
FROM
   flight_data
WHERE
   month=1
   AND origin='SFO';
```

- On hive shell: create the airports table

```
CREATE TABLE airports(
   name STRING,
   country STRING,
   area_code INT,
   code STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

- Load data into airports table:

```
LOAD DATA LOCAL INPATH 'cloudcon-hive/airports.csv' OVERWRITE INTO TABLE airports;
```

- On hive shell, list some rows from the airports table:

```
SELECT
   *
FROM
   airports
LIMIT 10
```

- On hive shell: run a join query to find the average delay in January 2008 for each airport and to print out the airport's name:

```
SELECT
   name,
   AVG(arr_delay)
FROM
   flight_data f
   INNER JOIN airports a
   ON (f.origin=a.code)
WHERE
   month=1
GROUP BY
   name;
```