CS3563: Assignment 2

Deep Diwani (EE16BTECH11006)
Devansh Agarwal (ES16BTECH11009)
Ninad Akolekar (CS16BTECH11024)

March 15, 2019

Contents

1 Problem Statement

The aim of this assignment is to convert the ER diagram designed in the previous assignment to a relational schema and populate the database with data provided by the **Bureau of Transportation Statistics** [3]. We have used PostgrSQL to complete this assignment.

2 ER Diagram to Table

- For entity "airport" an airport table is created.
- For entity "aircraft" and relation "owns" an aircraft table is created.
- For entity "airline" an airline table is created.
- For entity "summary" and relation "summarizes" a summary table is created.
- For relations "Flights offered to", "Flights offered from", "Arrival","Departure", "Provides" and entity "flight" a table flight is created.
- For relation "diversion" a diversion table is created.
- No table for relation "Flight Trip" is created as we think this should be done based on queries because if we store all the connecting flights it will take an huge amount of space and most of the stored data will never be used.

3 Description of Tables

3.1 Airport

Long_term_id and sequence_id are used as primary key as the sequence_id changes shortly with time it needs to be coupled with Long_term_id to make a minimal set for primary key.

Name	Type	Description	Foreign Key
long_term_id	integer	PRIMARY KEY	-
sequence_id	integer	PRIMARY KEY	-
city_market_id	integer	NOT NULL	-
airport_code	character varying(5)	NOT NULL	-
airport_name	character varying(200)	NOT NULL	-
city_name	character varying(200)	NOT NULL	-
state_abr	character varying(10)	NOT NULL	-
state_fips	integer	NOT NULL	-
state_name	character varying(200)	NOT NULL	-
world_area_code	integer	NOT NULL	-

3.2 Airline

As mentioned on the BTS website[3] unqique_carrier_code is unique so it is made the primary key.

Name	Type	Description	Foreign Key
unique_carrier_code	character varying(10)	PRIMARY KEY	-
govt_id	integer	NOT NULL	-
other_org_id	character varying(10)	NOT NULL	-

3.3 Aircraft

For aircraft the primary key is unique_carrier_code and tail_number as each airline will have multiple aircrafts and one aircraft might have bee owned by multiple airlines at some point of time. unique_carrier_code acts as a foreign key for table airline.

Name	Type	Description	Foreign Key
unique_carrier_code	character varying(10)	PRIMARY KEY	airline.unique_carrier_code
tail_number	character varying(10)	PRIMARY KEY	-

3.4 Flight

Flight has 7 attributes as primary key, the main idea was to make a combination of flight number, time and origin airport. Primary key contains year, month, day_of_month, airline_carrier_code, flight_number, origin_airport_long_id and dept_time. This table has foreign keys for many other tables:

- Table: airline, Fk: airline_carrier_code
- Table: airport, Fk1:origin_airport_seq_id,origin_airport.long_term_id1; Fk2:dest_airport_seq_id,dest_airport.long_term_id1
- Table:cancellation_code, FK: cancel_code

Name	Type	Description	Foreign Key
year	integer	PRIMARY KEY	-
month	integer	PRIMARY KEY	-
day_of_month	integer	PRIMARY KEY	-
airline_carrier_code	character varying(10)	PRIMARY KEY	airline.unique_carrier_code
aircraft_tail_number	character varying(10)	NOT NULL	aircraft.tail_number
flight_number	integer	PRIMARY KEY	-
origin_airport_long_id	integer	PRIMARY KEY	airport.long_term_id1
origin_airport_seq_id	integer	NOT NULL	airport.sequence_id1
dest_airport_long_id	integer	NOT NULL	airport.long_term_id2
dest_airport_seq_id	integer	NOT NULL	airport.sequence_id2
crs_dept_time	character varying(4)	NOT NULL	-
dept_time	character varying(4)	PRIMARY KEY	-
crs_arr_time	character varying(4)	NOT NULL	-
arr_time	character varying(4)	NOT NULL	-
cancel_code	character varying(1)		cancellation_code.code
distance	numeric	NOT NULL	-
carrier_delay	numeric		-
weather_delay	numeric		-
nas_delay	numeric		-
security_delay	numeric		-
late_aircraft_delay	numeric		-
diverted_landings	integer	-	

3.5 Summary

Primary key for Summary is same as that of flight as each flight will have a related summary. Primary key contains flight_year, flight_month, flight_day_of_month, flight_airline_carrier_code, flight_flight_number, flight_origin_airport_long_id and flight_dept_time.

Name	Type	Description	Foreign Key
flight_dept_time	character varying(4)	PRIMARY KEY	-
flight_day_of_month	integer	PRIMARY KEY	-
flight_month	integer	PRIMARY KEY	-
flight_year	integer	PRIMARY KEY	-
flight_number	integer	PRIMARY KEY	-
flight_airline_carrier_code	character varying(10)	PRIMARY KEY	airline.unique_carrier_code
flight_origin_airport_long_id	integer	PRIMARY KEY	-
flight_crs_dept_time	character varying(4)	NOT NULL	-
dep_delay	numeric		-
dep_delay_new	numeric		-
dep_del15	numeric		-
taxi_out	numeric		-
wheels_off	character varying(4)		-
wheels_on	character varying(4)		-
taxi_in	numeric		-
crs_arr_time	character varying(4)		-
arr_time	character varying(4)		-
arr_delay	numeric		-
arr_del15	numeric		-
crs_elapsed_time	numeric		-
actual_elapsed_time	numeric		-
air_time	numeric		-
distance	numeric	-	

3.6 Diversion

Primary key for Diversion is same as that of flight as each flight may have a related diversion. Primary key contains flight_year, flight_month, flight_day_of_month, flight_airline_carrier_code, flight_flight_number, flight_origin_airport_long_id and flight_dept_time.

Name	Type	Description	Foreign Key
flight_dept_time	character varying(4)	PRIMARY KEY	-
flight_day_of_month	integer	PRIMARY KEY	-
flight_month	integer	PRIMARY KEY	-
flight_year	integer	PRIMARY KEY	-
flight_number	integer	PRIMARY KEY	-
flight_airline_carrier_code	character varying(10)	PRIMARY KEY	airline.unique_carrier_code
flight_origin_airport_long_id	integer	PRIMARY KEY	-
div_airport_landings	integer		-
div_reached_dest	numeric		-
div_actual_elapsed_time	numeric		-
div_arr_delay	numeric		-
div_distance	numeric		-
div1_airport	text		-
div1_airport_id	numeric		-
div1_airport_seq_id	numeric		-
div1_wheels_on	character varying(4)		-
div1_total_gtime	numeric		-
div1_wheels_off	character varying(4)		-
div1_tail_num	text		-
div2_airport	text		-
div2_airport_id	numeric		-
div2_airport_seq_id	numeric		-
div2_wheels_on	character varying(4)		-
div2_total_gtime	numeric		-
div2_wheels_off	character varying(4)		-
div2_tail_num	text		-
div3_airport	text		-
div3_airport_id	numeric		-
div3_airport_seq_id	numeric		-
div3_wheels_on	character varying(4)		-
div3_total_gtime	numeric		-
div3_wheels_off	character varying(4)		-
div3_tail_num	text		-
div4_airport	text		-
div4_airport_id	numeric		-
div4_airport_seq_id	numeric		-
div4_wheels_on	character varying(4)		-
div4_total_gtime	numeric		-
div4_wheels_off	character varying(4)		-
div4_tail_num	text		-
div5_airport	text		-
div5_airport_id	numeric		-
div5_airport_seq_id	numeric		-
div5_wheels_on	character varying(4)		-
div5_total_gtime	numeric		-
div5_wheels_off	character varying(4)		-
div5_tail_num	text	-	

3.7 Cancellation

Primary key = code.

Name	Type	Description	Foreign Key
code	character(1)	PRIMARY KEY	-
code_description	text	NOT NULL	-

4 Pre-processing data

The data crawled from the BTS[3] website needs to be re-structured before it can be used to populate the tables of our database. Hence, after creating the tables as described in the previous section, the following pre-processing steps were carried out:

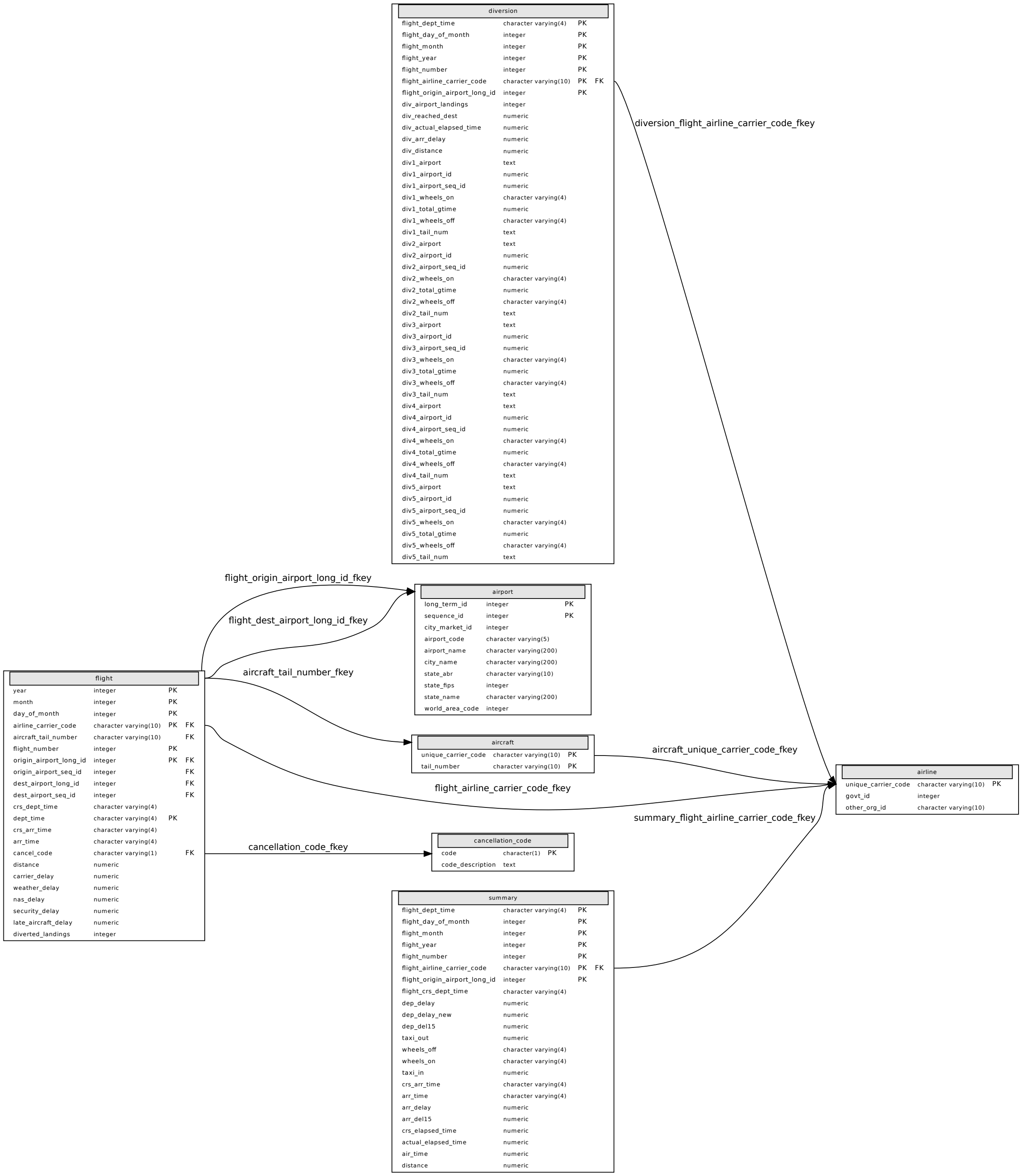
- It was observed that some of the flight records did not have aircraft tail number mentioned (missing values). To accommodate such records in our database, an additional record corresponding to every airline was inserted into the Aircraft table; having $tail_{n}umberattributeas"Unknown"$.

- It was also observed that the diverted flight data was very sparse. Out of all the flights between January to December 2017, only 13080 flights (0.002%) were diverted. To store these records in an efficient manner, we decided to make an new table for diversions that contains diversion details of only the flights that were diverted. This needed pre-processing of the diverted flight data downloaded from the BTS website[3] and removal of flight records that did not have any diversions.

We used the Pandas[1] library to pre-process and clean the data obtained from the BTS website[3].

5 Database Schema

A visual representation of the database schema is given on the next page. We used PostgreSQL Autodoc[2] to generate this visual representation.


diversion_flight_airline_carrier_code_fkey
flight_origin_airport_long_id_fkey
flight_dest_airport_long_id_fkey
aircraft_tail_number_fkey
aircraft_unique_carrier_code_fkey
flight_airline_carrier_code_fkey
summary_flight_airline_carrier_code_fkey
cancellation_code_fkey

diversion			
flight_dept_time	character varying(4)	PK	
flight_day_of_month	integer	PK	
flight_month	integer	PK	
flight_year	integer	PK	
flight_number	integer	PK	
flight_airline_carrier_code	character varying(10)	PK	FK
flight_origin_airport_long_id	integer	PK	
div_airport_landings	integer		
div_reached_dest	numeric		
div_actual_elapsed_time	numeric		
div_arr_delay	numeric		
div_distance	numeric		
div1_airport	text		
div1_airport_id	numeric		
div1_airport_seq_id	numeric		
div1_wheels_on	character varying(4)		
div1_total_gtime	numeric		
div1_wheels_off	character varying(4)		
div1_tail_num	text		
div2_airport	text		
div2_airport_id	numeric		
div2_airport_seq_id	numeric		
div2_wheels_on	character varying(4)		
div2_total_gtime	numeric		
div2_wheels_off	character varying(4)		
div2_tail_num	text		
div3_airport	text		
div3_airport_id	numeric		
div3_airport_seq_id	numeric		
div3_wheels_on	character varying(4)		
div3_total_gtime	numeric		
div3_wheels_off	character varying(4)		
div3_tail_num	text		
div4_airport	text		
div4_airport_id	numeric		
div4_airport_seq_id	numeric		
div4_wheels_on	character varying(4)		
div4_total_gtime	numeric		
div4_wheels_off	character varying(4)		
div4_tail_num	text		
div5_airport	text		
div5_airport_id	numeric		
div5_airport_seq_id	numeric		
div5_wheels_on	character varying(4)		
div5_total_gtime	numeric		
div5_wheels_off	character varying(4)		
div5_tail_num	text		

airport		
long_term_id	integer	PK
sequence_id	integer	PK
city_market_id	integer	
airport_code	character varying(5)	
airport_name	character varying(200)	
city_name	character varying(200)	
state_abr	character varying(10)	
state_fips	integer	
state_name	character varying(200)	
world_area_code	integer	

flight			
year	integer	PK	
month	integer	PK	
day_of_month	integer	PK	
airline_carrier_code	character varying(10)	PK	FK
aircraft_tail_number	character varying(10)		FK
flight_number	integer	PK	
origin_airport_long_id	integer	PK	FK
origin_airport_seq_id	integer		FK
dest_airport_long_id	integer		FK
dest_airport_seq_id	integer		FK
crs_dept_time	character varying(4)		
dept_time	character varying(4)	PK	
crs_arr_time	character varying(4)		
arr_time	character varying(4)		
cancel_code	character varying(1)		FK
distance	numeric		
carrier_delay	numeric		
weather_delay	numeric		
nas_delay	numeric		
security_delay	numeric		
late_aircraft_delay	numeric		
diverted_landings	integer		

aircraft		
unique_carrier_code	character varying(10)	PK
tail_number	character varying(10)	PK

airline		
unique_carrier_code	character varying(10)	PK
govt_id	integer	
other_org_id	character varying(10)	

cancellation_code		
code	character(1)	PK
code_description	text	

summary			
flight_dept_time	character varying(4)	PK	
flight_day_of_month	integer	PK	
flight_month	integer	PK	
flight_year	integer	PK	
flight_number	integer	PK	
flight_airline_carrier_code	character varying(10)	PK	FK
flight_origin_airport_long_id	integer	PK	
flight_crs_dept_time	character varying(4)		
dep_delay	numeric		
dep_delay_new	numeric		
dep_del15	numeric		
taxi_out	numeric		
wheels_off	character varying(4)		
wheels_on	character varying(4)		
taxi_in	numeric		
crs_arr_time	character varying(4)		
arr_time	character varying(4)		
arr_delay	numeric		
arr_del15	numeric		
crs_elapsed_time	numeric		
actual_elapsed_time	numeric		
air_time	numeric		
distance	numeric		

References

[1] Wes McKinney. Data structures for statistical computing in python.

[2] Rod Taylor. Postgresql autodoc.

[3] Washington DC U.S. Department of Transportation, Bureau of Transportation Statistics. Reporting carrier on-time performance. `https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time`.