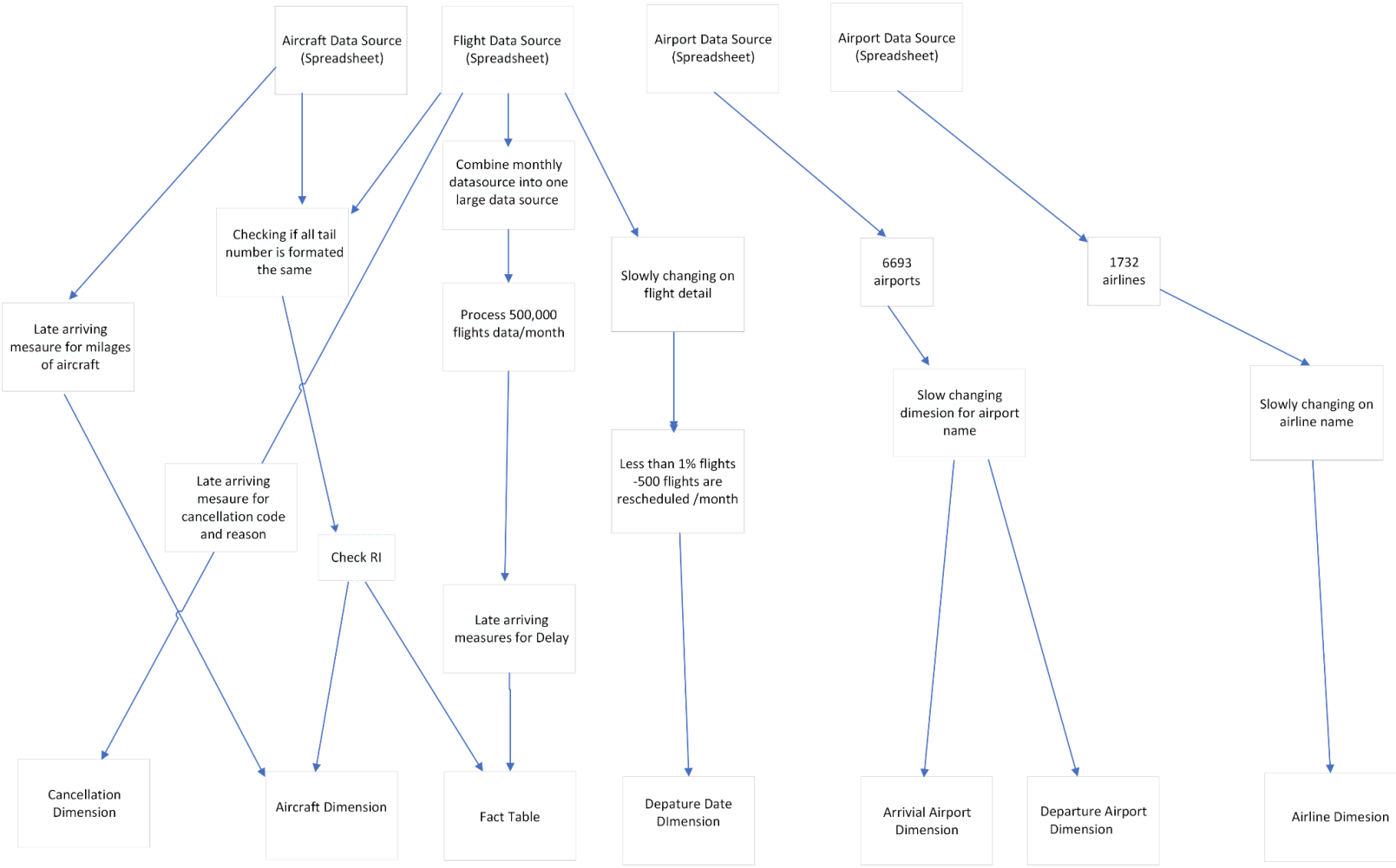
High-level planning



Handling inconsistency and null value in initial data source

Attribute Name	How it was handled
Null Values	
Flight Measures(e.g Air Time, Distance Traveled, Actual Time Elapsed, and Schedule Time Elapsed)	In the initial flight dataset prior to any transformation, it contains many statistics about the flights duration and distance traveled and those columns were always null if the flights were canceled, so after discovering that these values are null because the flight never occurred, we ended replacing these null value with 0 because the flight was canceled and never took placed

Flight Delays Attributes	In the initial flight dataset, a record(row) of a flight can have many types of delay that all have their own value(such as Weather, Traffic, Security, etc), rarely do flights have all 5 types of delay at once, which causes some of the delay value to be defaulted as null. After we found out that null value delays have no impact on the flight itself, we decided to replace the null value with 0 since that won't impact future analysis of delay
Tail Number(Aircraft ID) for flight and aircraft dataset	In the flight dataset, some of the values of the tail number were null. After examining the rows that contained a null tail number, we discovered that those tail numbers were null because the flight was canceled, thereby no aircraft were assigned to those flights. Therefore, we ended up replacing the null values to a new surrogate key that maps to the term "unknown"
Inconsistent Format for value	
Tail Number(Aircraft ID) for the flight	In the Tail Number(Aircraft Id used to merge the aircraft and the flight dataset) column, there was a very inconsistent format of the value. In the flight dataset, some of the values started with 'N' while others didn't, whereas in the aircraft dataset none of the Tail numbers started with a 'N. This made it very difficult to merge the two dataset without having any null values after joining. After this discovering we created a function that would change the Tail Number in the flight dataset to remove 'N' if it's the first character in order to match the Tail Number from the flight dataset to the aircraft dataset for future joining

Basically, we used indexes to generate surrogate keys for dimensions and the fact table. Before importing data to DBMS, dimension tables and the fact table are created.

Cancellation Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_cancellation
(
    "Cancellation_ID" integer NOT NULL,
    "Code" "char",
    "Description" character varying COLLATE pg_catalog."default",
```

```
    CONSTRAINT dim_cancellation_pkey PRIMARY KEY ("Cancellation_ID")
)
```

Aircraft Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_aircraft
(
    "Aircraft ID" integer NOT NULL,
    "TAIL_NUM" character varying COLLATE pg_catalog."default",
    "NAME" character varying COLLATE pg_catalog."default",
    CONSTRAINT dim_aircraft_pkey PRIMARY KEY ("Aircraft ID")
)
```

Origin Airport Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_origin_airport
(
    "Origin Airport Surrogate Key" integer NOT NULL,
    "ORIGIN_AIRPORT_ID" integer,
    "State" character varying COLLATE pg_catalog."default",
    "City" character varying COLLATE pg_catalog."default",
    "Name" character varying COLLATE pg_catalog."default",
    CONSTRAINT dim_origin_airport_pkey PRIMARY KEY ("Origin Airport Surrogate Key")
)
```

Destination Airport Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_destination_airport
(
    "Destination Airport Surrogate Key" integer NOT NULL,
    "DEST_AIRPORT_ID" integer,
    "State" character varying COLLATE pg_catalog."default",
    "City" character varying COLLATE pg_catalog."default",
    "Name" character varying COLLATE pg_catalog."default",
    CONSTRAINT dim_destination_airport_pkey PRIMARY KEY ("Destination Airport Surrogate
Key")
)
```

Flight Date Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_flight_date
(
    "Flight Date ID" integer NOT NULL,
    "YEAR" integer,
    "MONTH" integer,
    "DAY_OF_MONTH" integer,
    "FL_DATE" character varying COLLATE pg_catalog."default",
```

```
    CONSTRAINT dim_flight_date_pkey PRIMARY KEY ("Flight Date ID")
)
```

Airline Dimension:

```
CREATE TABLE IF NOT EXISTS "Staging".dim_airline
(
    "Airline ID" integer NOT NULL,
    "OP_CARRIER_AIRLINE_ID" integer,
    "Description" character varying COLLATE pg_catalog."default",
    CONSTRAINT dim_airline_pkey PRIMARY KEY ("Airline ID")
)
```

Fact Table:

```
CREATE TABLE IF NOT EXISTS "Staging".fact_table
(
    "Fact Table ID" bigint NOT NULL,
    "Cancellation Surrogate Key" integer,
    "Flight Date Surrogate Key" integer,
    "Airline Surrogate Key" integer,
    "Aircraft Surrogate Key" numeric,
    "Origin Airport Surrogate Key" integer,
    "Destination Airport Surrogate Key" integer,
    "DISTANCE" numeric,
    "CRS_ELAPSED_TIME" numeric,
    "ACTUAL_ELAPSED_TIME" numeric,
    "AIR_TIME" numeric,
    "CARRIER_DELAY" numeric,
    "WEATHER_DELAY" numeric,
    "NAS_DELAY" numeric,
    "SECURITY_DELAY" numeric,
    "LATE_AIRCRAFT_DELAY" numeric,
    CONSTRAINT fact_table_pkey PRIMARY KEY ("Fact Table ID"),
    CONSTRAINT fact_table_aircraft_fkey FOREIGN KEY ("Aircraft Surrogate Key")
        REFERENCES "Staging".dim_aircraft ("Aircraft ID") MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID,
    CONSTRAINT fact_table_airline_fkey FOREIGN KEY ("Airline Surrogate Key")
        REFERENCES "Staging".dim_airline ("Airline ID") MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID,
    CONSTRAINT fact_table_canclellation_fkey FOREIGN KEY ("Cancellation Surrogate Key")
        REFERENCES "Staging".dim_cancellation ("Cancellation_ID") MATCH SIMPLE
        ON UPDATE NO ACTION
```

```
        ON DELETE NO ACTION
        NOT VALID,
    CONSTRAINT fact_table_destination_airport_fkey FOREIGN KEY ("Destination Airport
Surrogate Key")
        REFERENCES "Staging".dim_destination_airport ("Destination Airport Surrogate Key")
MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID,
    CONSTRAINT fact_table_flight_date_fkey FOREIGN KEY ("Flight Date Surrogate Key")
        REFERENCES "Staging".dim_flight_date ("Flight Date ID") MATCH SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID,
    CONSTRAINT fact_table_origin_airport_fkey FOREIGN KEY ("Origin Airport Surrogate Key")
        REFERENCES "Staging".dim_origin_airport ("Origin Airport Surrogate Key") MATCH
SIMPLE
        ON UPDATE NO ACTION
        ON DELETE NO ACTION
        NOT VALID
)
```

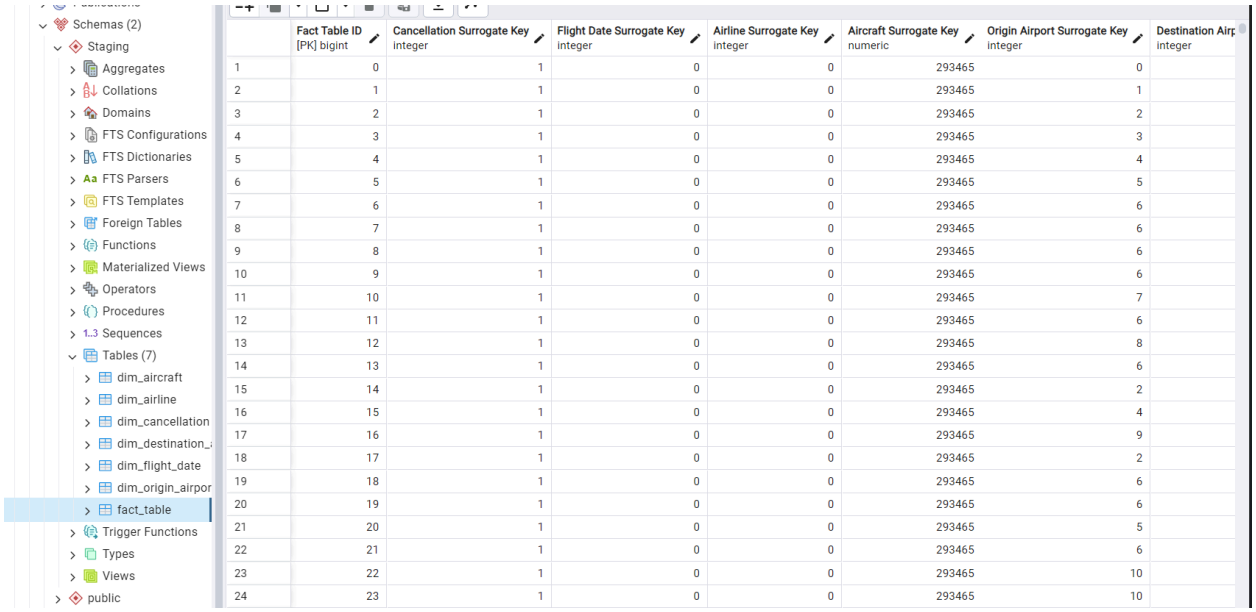
Added a Column is_cancelled in our fact_table:

```
-- Add the new column is_cancelled to the fact_table
ALTER TABLE "Staging".fact_table
ADD COLUMN is_cancelled INTEGER;

-- Update the values of the is_cancelled column based on the condition
UPDATE "Staging".fact_table
SET is_cancelled = CASE
                    WHEN "Cancellation Surrogate Key" = 4 THEN 0
                    ELSE 1
                END;
```

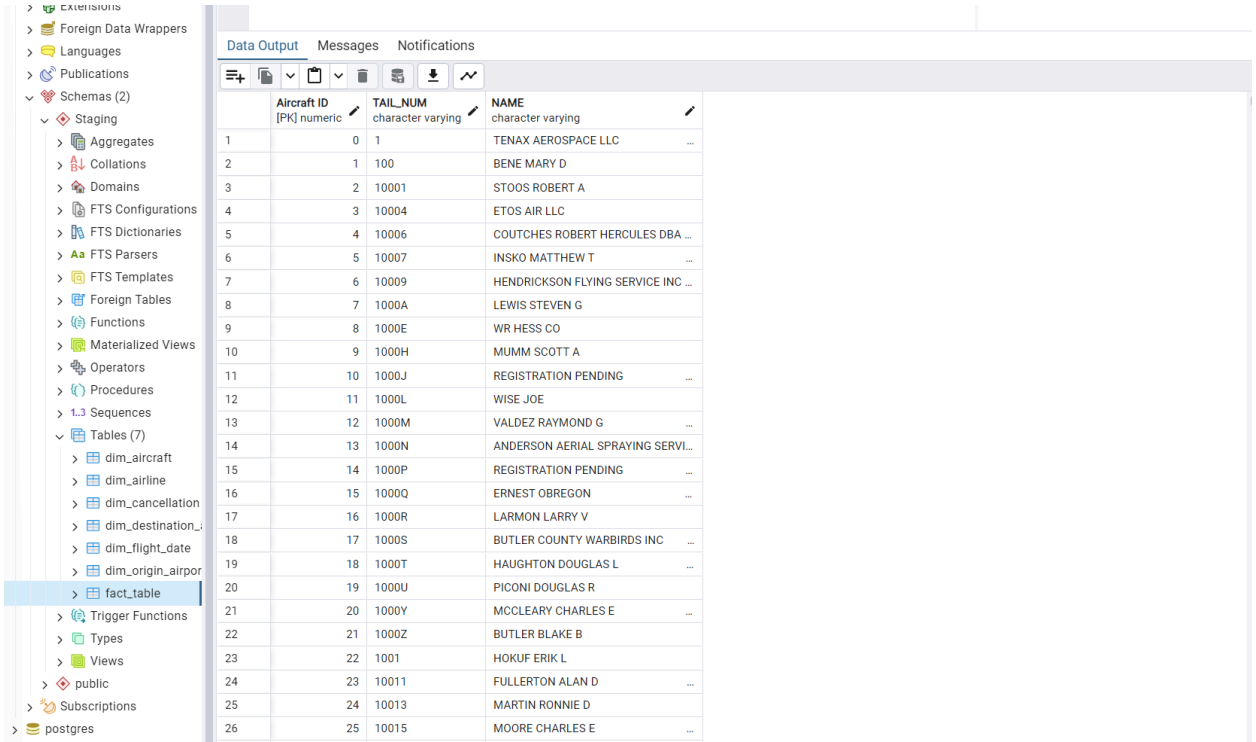
Screenshot for Data Warehouse in DBMS:

Fact Table



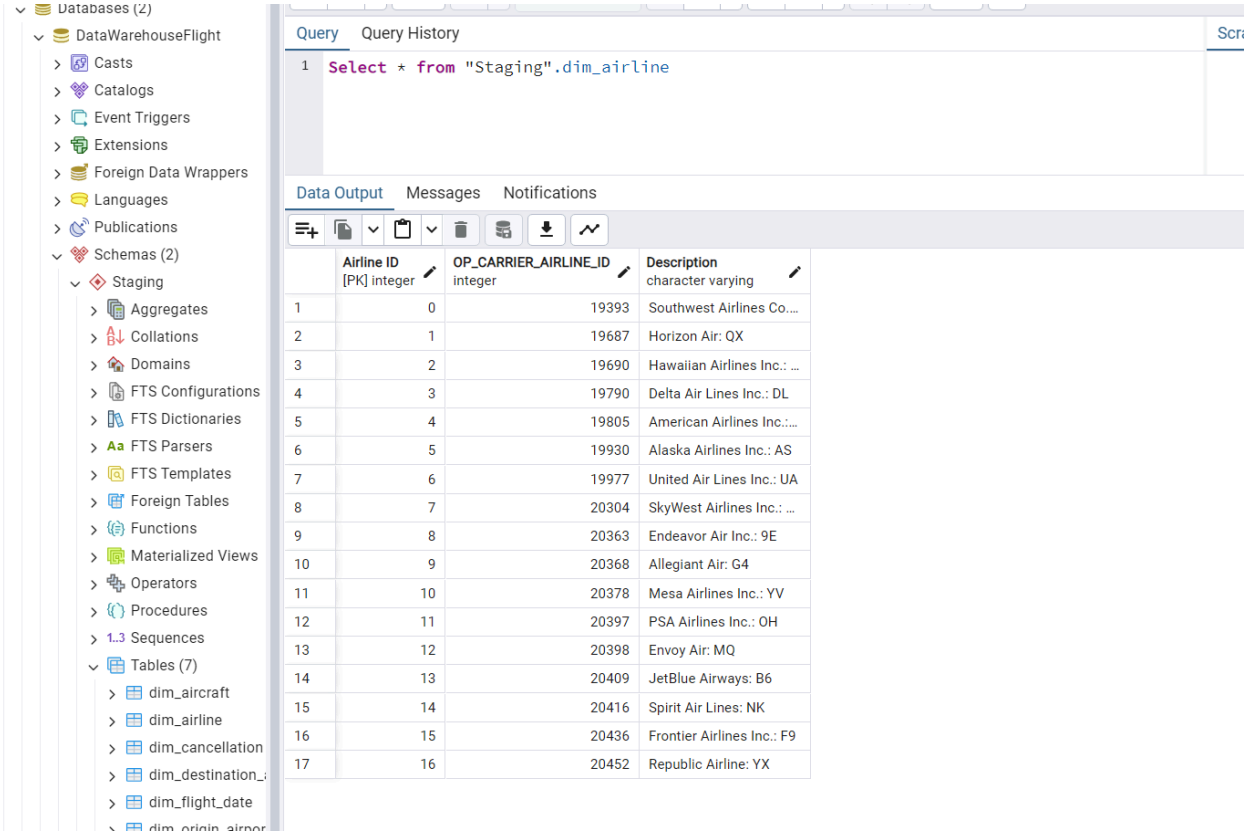
	Fact Table ID [PK] bigint	Cancellation Surrogate Key integer	Flight Date Surrogate Key integer	Airline Surrogate Key integer	Aircraft Surrogate Key numeric	Origin Airport Surrogate Key integer	Destination Airp integer
1	0	1	0	0	293465	0	
2	1	1	0	0	293465	1	
3	2	1	0	0	293465	2	
4	3	1	0	0	293465	3	
5	4	1	0	0	293465	4	
6	5	1	0	0	293465	5	
7	6	1	0	0	293465	6	
8	7	1	0	0	293465	6	
9	8	1	0	0	293465	6	
10	9	1	0	0	293465	6	
11	10	1	0	0	293465	7	
12	11	1	0	0	293465	6	
13	12	1	0	0	293465	8	
14	13	1	0	0	293465	6	
15	14	1	0	0	293465	2	
16	15	1	0	0	293465	4	
17	16	1	0	0	293465	9	
18	17	1	0	0	293465	2	
19	18	1	0	0	293465	6	
20	19	1	0	0	293465	6	
21	20	1	0	0	293465	5	
22	21	1	0	0	293465	6	
23	22	1	0	0	293465	10	
24	23	1	0	0	293465	10	

Aircraft Dimension:



	Aircraft ID [PK] numeric	TAIL_NUM character varying	NAME character varying
1	0	1	TENAX AEROSPACE LLC
2	1	100	BENE MARY D
3	2	10001	STOOS ROBERT A
4	3	10004	ETOS AIR LLC
5	4	10006	COUTCHES ROBERT HERCULES DBA ...
6	5	10007	INSKO MATTHEW T
7	6	10009	HENDRICKSON FLYING SERVICE INC ...
8	7	1000A	LEWIS STEVEN G
9	8	1000E	WR HESS CO
10	9	1000H	MUMM SCOTT A
11	10	1000J	REGISTRATION PENDING
12	11	1000L	WISE JOE
13	12	1000M	VALDEZ RAYMOND G
14	13	1000N	ANDERSON AERIAL SPRAYING SERVI...
15	14	1000P	REGISTRATION PENDING
16	15	1000Q	ERNEST OBREGON
17	16	1000R	LARMON LARRY V
18	17	1000S	BUTLER COUNTY WARBIRDS INC
19	18	1000T	HAUGHTON DOUGLAS L
20	19	1000U	PICONI DOUGLAS R
21	20	1000Y	MCCLEARY CHARLES E
22	21	1000Z	BUTLER BLAKE B
23	22	1001	HOKUF ERIK L
24	23	10011	FULLERTON ALAN D
25	24	10013	MARTIN RONNIE D
26	25	10015	MOORE CHARLES E

Airline Dimension:



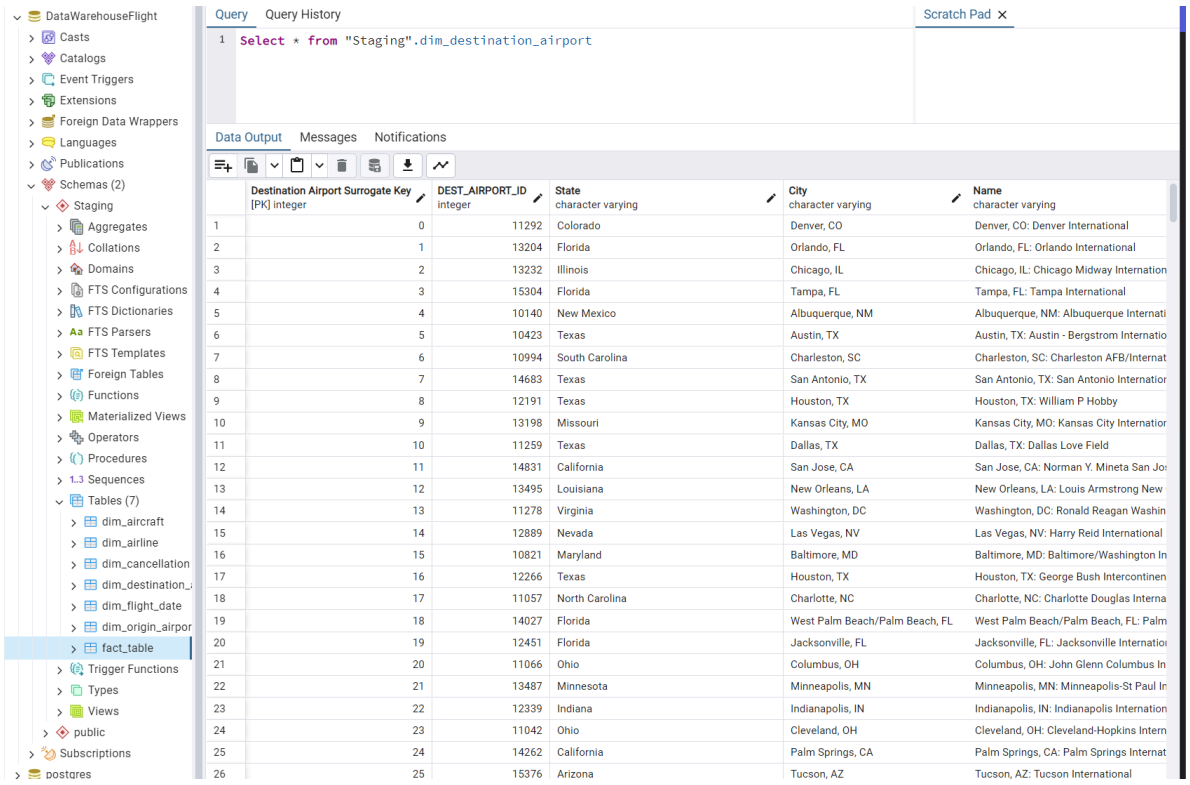
Cancellation Dimension:

```
1  Select * from "Staging".dim_cancellation
```

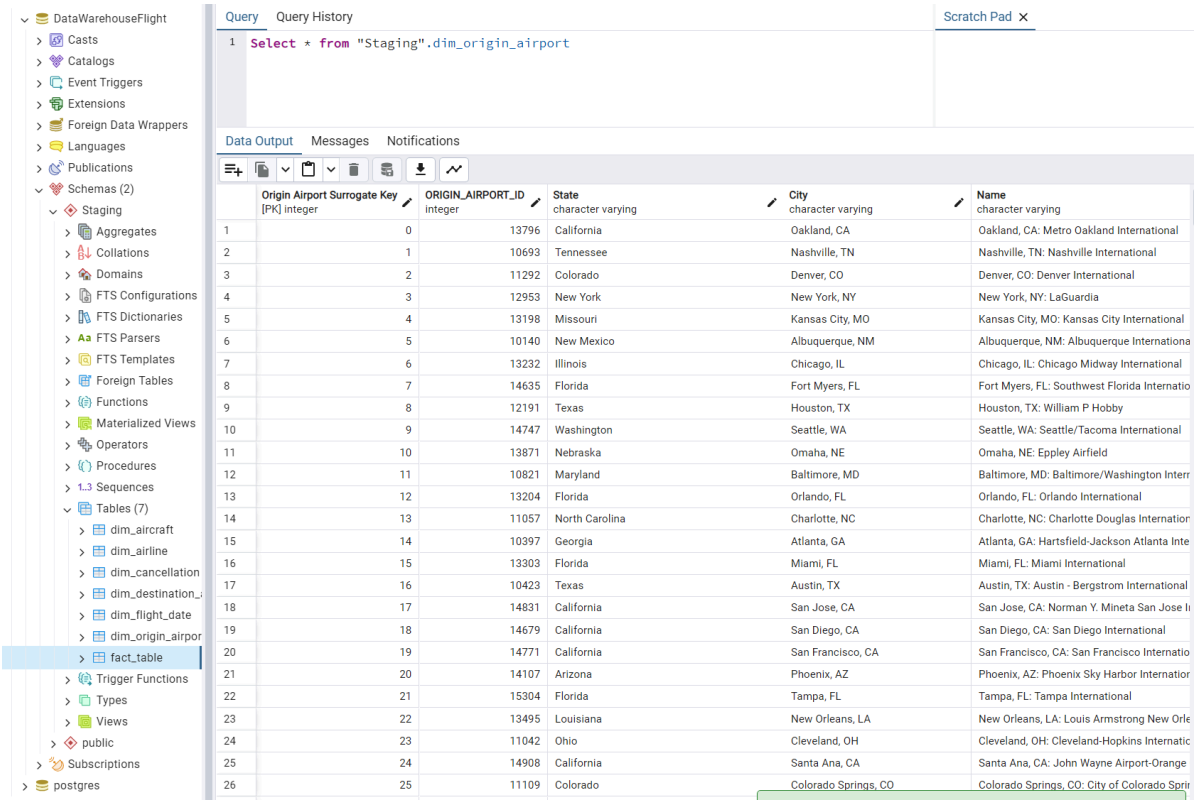
Data Output | Messages | Notifications

	Cancellation_ID [PK] integer	Code "char"	Description character varying
1	0	A	Carrier
2	1	B	Weather
3	2	C	National Air System
4	3	D	Security
5	4	E	No Cancellation

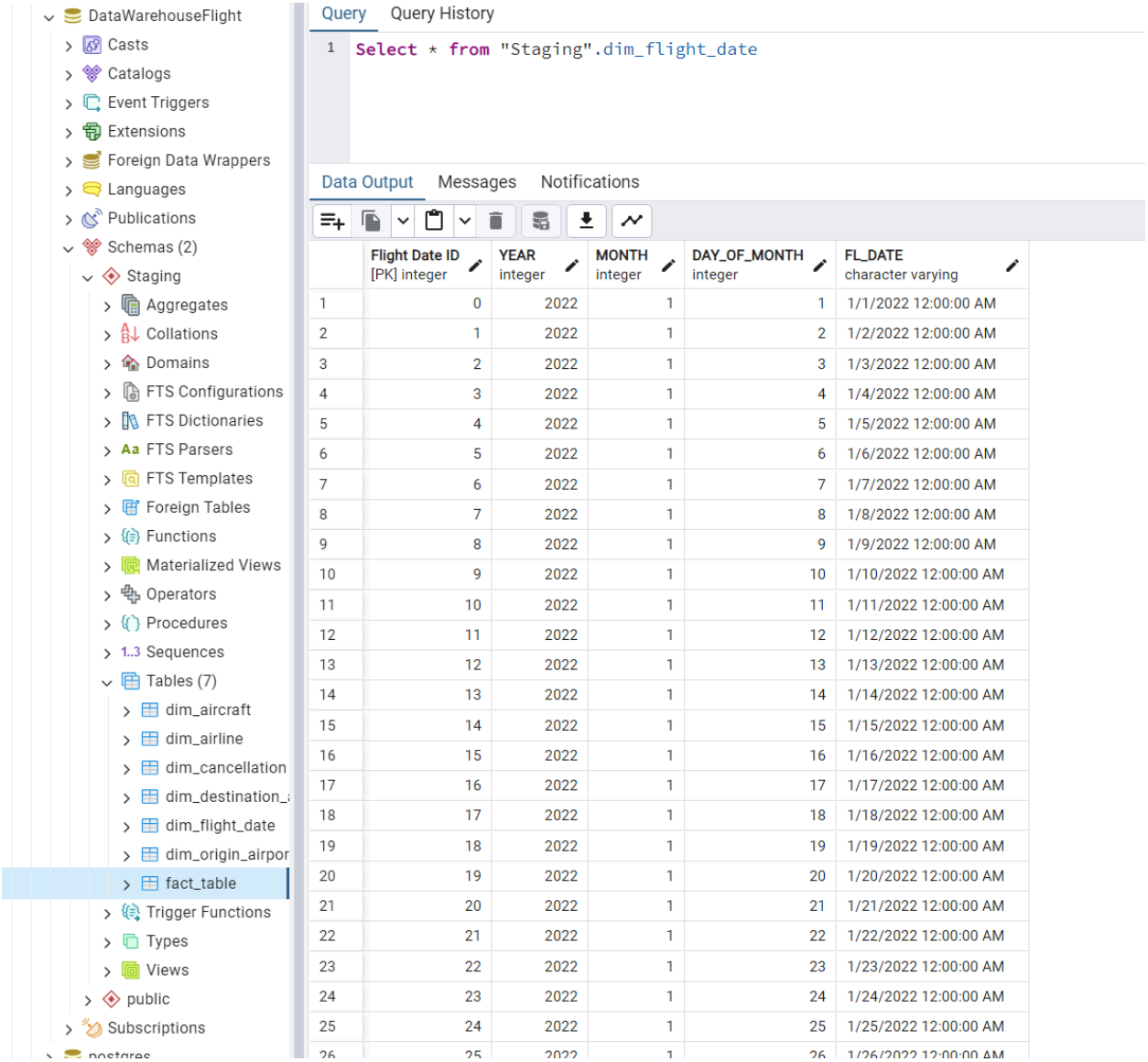
Destination airport Dimension:



Origin airport Dimension:



Flight Date Dimension:



	Flight Date ID [PK] integer	YEAR integer	MONTH integer	DAY_OF_MONTH integer	FL_DATE character varying
1	0	2022	1	1	1/1/2022 12:00:00 AM
2	1	2022	1	2	1/2/2022 12:00:00 AM
3	2	2022	1	3	1/3/2022 12:00:00 AM
4	3	2022	1	4	1/4/2022 12:00:00 AM
5	4	2022	1	5	1/5/2022 12:00:00 AM
6	5	2022	1	6	1/6/2022 12:00:00 AM
7	6	2022	1	7	1/7/2022 12:00:00 AM
8	7	2022	1	8	1/8/2022 12:00:00 AM
9	8	2022	1	9	1/9/2022 12:00:00 AM
10	9	2022	1	10	1/10/2022 12:00:00 AM
11	10	2022	1	11	1/11/2022 12:00:00 AM
12	11	2022	1	12	1/12/2022 12:00:00 AM
13	12	2022	1	13	1/13/2022 12:00:00 AM
14	13	2022	1	14	1/14/2022 12:00:00 AM
15	14	2022	1	15	1/15/2022 12:00:00 AM
16	15	2022	1	16	1/16/2022 12:00:00 AM
17	16	2022	1	17	1/17/2022 12:00:00 AM
18	17	2022	1	18	1/18/2022 12:00:00 AM
19	18	2022	1	19	1/19/2022 12:00:00 AM
20	19	2022	1	20	1/20/2022 12:00:00 AM
21	20	2022	1	21	1/21/2022 12:00:00 AM
22	21	2022	1	22	1/22/2022 12:00:00 AM
23	22	2022	1	23	1/23/2022 12:00:00 AM
24	23	2022	1	24	1/24/2022 12:00:00 AM
25	24	2022	1	25	1/25/2022 12:00:00 AM
26	25	2022	1	26	1/26/2022 12:00:00 AM