-- шаг 2. ETL процедура инкрементной загрузки для устранения дубликатов

-- шаг 2.1. Создание ETL таблиц

create schema if not exists kdz\_7\_etl;

drop table if exists kdz\_7\_etl.load\_flights;

create table if not exists kdz\_7\_etl.load\_flights (

loaded\_ts timestamp not null

);

​

drop table if exists kdz\_7\_etl.load\_weather;

create table if not exists kdz\_7\_etl.load\_weather (

loaded\_ts timestamp not null

);

-- шаг 2.2. ETL процедура для отбора данных по времени

drop table if exists kdz\_7\_etl.load\_min\_max\_flights;

create table if not exists kdz\_7\_etl.load\_min\_max\_flights as

select

min(loaded\_ts) as ts1

, max(loaded\_ts) as ts2

from kdz\_7\_src.flights

where loaded\_ts > coalesce((select max(loaded\_ts) from kdz\_7\_etl.load\_flights), '1970-01-01')

;

select \*

from kdz\_7\_etl.load\_min\_max\_flights

;

drop table if exists kdz\_7\_etl.load\_min\_max\_weather;

create table if not exists kdz\_7\_etl.load\_min\_max\_weather as

select

min(loaded\_ts) as ts1

, max(loaded\_ts) as ts2

from kdz\_7\_src.weather

where loaded\_ts > coalesce((select max(loaded\_ts) from kdz\_7\_etl.load\_weather), '1970-01-01')

;

select \*

from kdz\_7\_etl.load\_min\_max\_weather

;

drop table if exists kdz\_7\_etl.flights;

create table if not exists kdz\_7\_etl.flights as

select distinct

flight\_year,

flight\_quarter,

flight\_month,

to\_date(flight\_date, 'M/DD/YYYY HH:MI:SS AM') as flight\_date,

reporting\_airline,

tail\_number,

flight\_number,

origin,

destination,

to\_timestamp(crs\_dep\_time, 'HH24MI')::time as crs\_dep\_time,

to\_timestamp(dep\_time, 'HH24MI')::time as dep\_time,

dep\_delay\_min,

cancelled,

cancellation\_code,

air\_time,

distance,

weather\_delay

from kdz\_7\_src.flights, kdz\_7\_etl.load\_min\_max\_flights

where 1=1

and loaded\_ts between ts1 and ts2

;

select \*

from kdz\_7\_etl.flights

;

drop table if exists kdz\_7\_etl.weather;

create table if not exists kdz\_7\_etl.weather as

select distinct

icao\_code,

to\_timestamp(local\_datetime, 'DD.MM.YYYY HH24:MI:SS') as local\_datetime,

air\_temp,

p0\_station\_lvl,

p\_sea\_lvl,

humidity,

wind\_direction,

wind\_speed,

max\_gust\_speed,

phenomena\_observed,

phenomena\_significant,

total\_cloud\_cover,

visibility,

dewpoint\_temp

from kdz\_7\_src.weather, kdz\_7\_etl.load\_min\_max\_weather

where 1=1

and loaded\_ts between ts1 and ts2

;

select \*

from kdz\_7\_etl.weather

;

-- шаг 2.3. Создание таблиц staging

create schema if not exists kdz\_7\_staging;

drop table if exists kdz\_7\_staging.flights;

create table kdz\_7\_staging.flights(

flight\_year varchar(5) not null,

flight\_quarter varchar(5) not null,

flight\_month varchar(5) not null,

flight\_date date not null,

reporting\_airline varchar(5) not null,

tail\_number varchar(10) null,

flight\_number varchar(10) null,

origin varchar(5) not null,

destination varchar(5) not null,

crs\_dep\_time time not null,

dep\_time time null,

dep\_delay\_min varchar(10) null,

cancelled float not null,

cancellation\_code char(1) null,

air\_time varchar(10) null,

distance varchar(10) not null,

weather\_delay varchar(10) null,

loaded\_ts timestamp NOT NULL DEFAULT now(),

primary key (flight\_date, flight\_number, origin, destination, crs\_dep\_time)

)

;

drop table if exists kdz\_7\_staging.weather;

create table kdz\_7\_staging.weather(

icao\_code varchar(10) default 'PANC',

local\_datetime varchar(25) not null primary key,

air\_temp varchar(10) not null,

p0\_station\_lvl varchar(10) NOT null,

p\_sea\_lvl varchar(10) NOT null,

humidity varchar(5) not null,

wind\_direction varchar(100) null,

wind\_speed varchar(5) null,

max\_gust\_speed varchar(10) null,

phenomena\_observed varchar(50) null,

phenomena\_significant varchar(50) null,

total\_cloud\_cover varchar(1000) not null,

visibility varchar(1000) not null,

dewpoint\_temp varchar(1000) not null,

loaded\_ts timestamp NOT NULL DEFAULT now()

)

;

-- шаг 2.4. Загрузка данных на уровень staging

insert into kdz\_7\_staging.flights (flight\_year, flight\_quarter, flight\_month, flight\_date, reporting\_airline, tail\_number, flight\_number, origin, destination, crs\_dep\_time, dep\_time, dep\_delay\_min, cancelled, cancellation\_code, air\_time, distance, weather\_delay)

select distinct

flight\_year,

flight\_quarter,

flight\_month,

flight\_date,

reporting\_airline,

tail\_number,

flight\_number,

origin,

destination,

crs\_dep\_time,

dep\_time,

dep\_delay\_min,

cancelled,

cancellation\_code,

air\_time,

distance,

weather\_delay

from kdz\_7\_etl.flights

on conflict (flight\_date, flight\_number, origin, destination, crs\_dep\_time) do update

set

flight\_year = excluded.flight\_year,

flight\_quarter = excluded.flight\_quarter,

flight\_month = excluded.flight\_month,

reporting\_airline = excluded.reporting\_airline,

tail\_number = excluded.tail\_number,

dep\_time = excluded.dep\_time,

dep\_delay\_min = excluded.dep\_delay\_min,

cancelled = excluded.cancelled,

cancellation\_code = excluded.cancellation\_code,

air\_time = excluded.air\_time,

distance = excluded.distance,

weather\_delay = excluded.weather\_delay,

loaded\_ts = now()

;

select \*

from kdz\_7\_staging.flights

;

insert into kdz\_7\_staging.weather (icao\_code, local\_datetime, air\_temp, p0\_station\_lvl, p\_sea\_lvl, humidity, wind\_direction, wind\_speed, max\_gust\_speed, phenomena\_observed, phenomena\_significant, total\_cloud\_cover, visibility, dewpoint\_temp )

select

icao\_code,

local\_datetime,

air\_temp,

p0\_station\_lvl,

p\_sea\_lvl,

humidity,

wind\_direction,

wind\_speed,

max\_gust\_speed,

phenomena\_observed,

phenomena\_significant,

total\_cloud\_cover,

visibility,

dewpoint\_temp

from kdz\_7\_etl.weather

on conflict (local\_datetime) do update

set

icao\_code = excluded.icao\_code,

air\_temp = excluded.air\_temp,

p0\_station\_lvl = excluded.p0\_station\_lvl,

p\_sea\_lvl = excluded.p\_sea\_lvl,

humidity = excluded.humidity,

wind\_direction = excluded.wind\_direction,

wind\_speed = excluded.wind\_speed,

max\_gust\_speed = excluded.max\_gust\_speed,

phenomena\_observed = excluded.phenomena\_observed,

phenomena\_significant = excluded.phenomena\_significant,

total\_cloud\_cover = excluded.total\_cloud\_cover,

visibility = excluded.visibility,

dewpoint\_temp = excluded.dewpoint\_temp,

loaded\_ts = now()

;

select \*

from kdz\_7\_staging.weather

;