IT - Developer

JAVA

Task for application

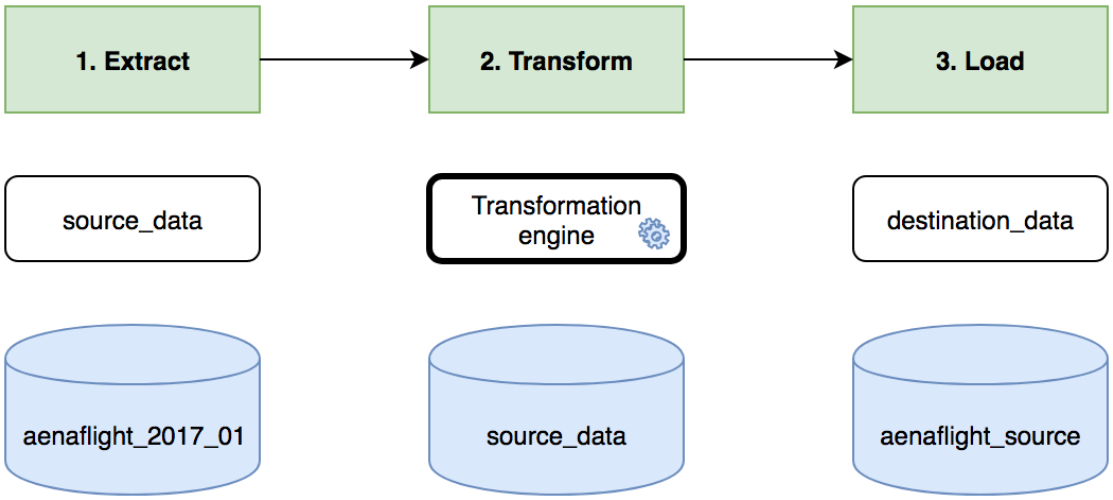
Technology

- JAVA 8
- Framework: Spring / Hibernate
- Maven
- PostgreSQL (9.x -10.x)
- Unix (Centos)

Description

ETL (Extract Transform Load).

Diagram



Task consist of 3 steps.

1. extract - during this step data loaded from source table.

2. transform - during this step data is processed (removed duplicates, added/merged missing information from existing records, removed unused records, parse/correct existing data). After transform source table should contain one cleaned record which contains all information of previous records.

3. load - load processed information into destination table.

Application should store actual state of processing, in case of stop or death should be able to restart process from stopped location and state. Additional resources/tables for processing can be created. Application should use multi-threading where possible.

Transformation process description

A scraper is gathering information (on a flight) over a period of time. Therefore we have multiple records for a single flight. This is just a quick note, why we have multiple records for one flight.

We keep the newest (based on the latest unix timestamp when record was created = **created_at**) information of the flight. Sometimes we keep all previous information.

column	processing/transformation description
act_arr_date_time_lt	final record should contain latest possible value
arr_apt_code_iata	final record should contain latest possible value
aircraft_name_scheduled	final record should contain latest possible value
arr_apt_name_es	final record should contain latest possible value
baggage_info	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
carrier_airline_name_en	final record should contain latest possible value
carrier_icao_code	final record should contain latest possible value
carrier_number	final record should contain latest possible value
counter	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
dep_apt_name_es	final record should contain latest possible value
dep_apt_code_iata	final record should contain latest possible value
est_arr_date_time_lt	final record should contain latest possible value
est_dep_date_time_lt	final record should contain latest possible value
flight_airline_name_en	final record should contain latest possible value
flight_airline_name	final record should contain latest possible value
flight_icao_code	final record should contain latest possible value
flight_number	final record should contain latest possible value
flt_leg_seq_no	final record should contain latest possible value
gate_info	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
lounge_info	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
schd_arr_only_date_lt	final record should contain latest possible value
schd_arr_only_time_lt	final record should contain latest possible value
source_data	final record should contain latest possible value
status_info	final record should contain latest possible value
terminal_info	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
arr_terminal_info	final record should contain comma-separated aggregation of previous values without duplicates (LIFO)
act_dep_date_time_lt	final record should contain latest possible value
schd_dep_only_date_lt	final record should contain latest possible value
schd_dep_only_time_lt	final record should contain latest possible value

LIFO (last in first out) - latest record will be first in list.

Example: A1,A1,B2,B2,B2,C3,D1 should be be transformed to: D1,C3,B2,A1.

Data definitions

source_data table definition

column	type	description
id	bigint	primary key, unique identifier
act_arr_date_time_lt	character varying(64)	actual arrival timestamp
aircraft_name_scheduled	text	scheduled aircraft name
arr_apt_name_es	character varying(128)	Arrival airport name in Spanish
arr_apt_code_iata	character varying(8)	IATA/ICAO code of arrival airport
baggage_info	character varying(128)	Baggage information of flight
carrier_airline_name_en	character varying(128)	Carrier airline name English
carrier_icao_code	character varying(8)	Carrier IATA/ICAO code
carrier_number	character varying(8)	Carrier number
counter	character varying(64)	Registration counter
dep_apt_name_es	character varying(128)	Departure airport name Spanish
dep_apt_code_iata	character varying(8)	IATA/ICAO code of departure airport
est_arr_date_time_lt	character varying(64)	estimated arrival timestamp
est_dep_date_time_lt	character varying(64)	estimated departure timestamp
flight_airline_name_en	character varying(128)	Flight airline name English
flight_airline_name	character varying(128)	Flight Airline name
flight_icao_code	character varying(8)	IATA/ICAO flight airline code
flight_number	character varying(8)	flight number
flt_leg_seq_no	character varying(8)	flight leg sequence id
gate_info	character varying(128)	gate information
lounge_info	character varying(128)	lounge information
schd_arr_only_date_lt	character varying(32)	scheduled arrival date
schd_arr_only_time_lt	character varying(32)	scheduled arrival time
source_data	text	source of data
status_info	character varying(128)	flight status
terminal_info	character varying(128)	terminal information
arr_terminal_info	character varying(128)	arrival terminal information
act_dep_date_time_lt	character varying(64)	actual departure timestamp
schd_dep_only_date_lt	character varying(32)	scheduled departure date
schd_dep_only_time_lt	character varying(32)	scheduled departure time
created_at	bigint	unix timestamp when record was created

destination_data table definition

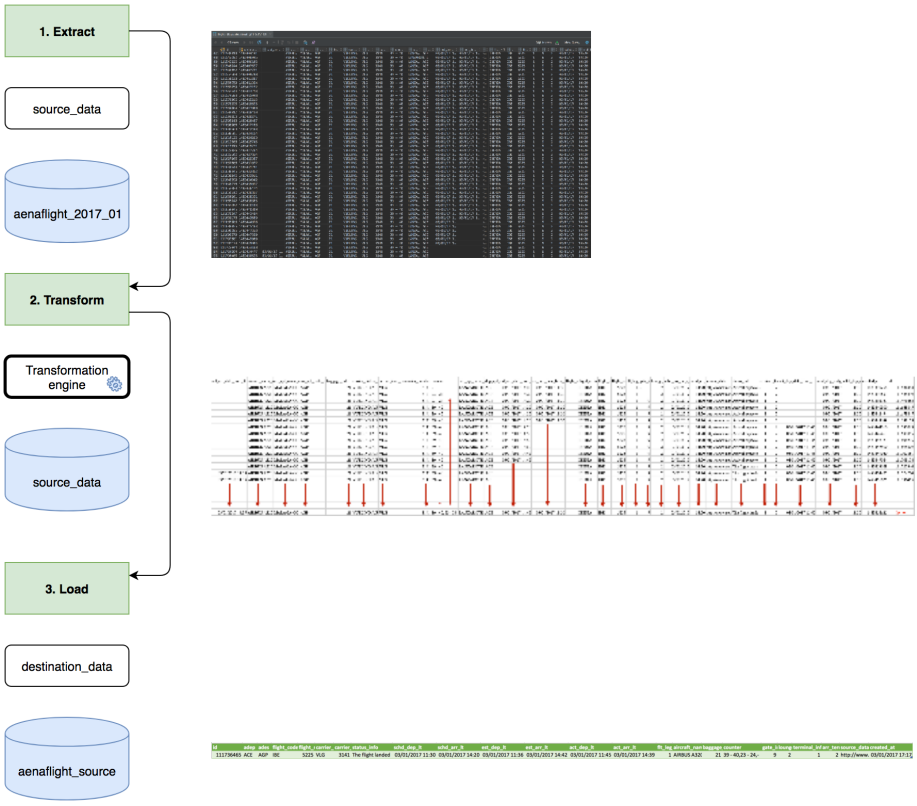
column	type	description
id	bigint	primary key, unique identifier
adep	character varying(8)	IATA/ICAO code of departure airport
ades	character varying(8)	IATA/ICAO code of destination airport
flight_code	character varying(8)	IATA/ICAO airline code of flight
flight_number	character varying(8)	flight number
carrier_code	character varying(8)	IATA/ICAO airline code of carrier
carrier_number	character varying(8)	carrier number
status_info	character varying(256)	flight status
schd_dep_lt	timestamp without time zone	scheduled departure timestamp
schd_arr_lt	timestamp without time zone	scheduled arrival timestamp
est_dep_lt	timestamp without time zone	estimated departure timestamp
est_arr_lt	timestamp without time zone	estimated arrival timestamp
act_dep_lt	timestamp without time zone	actual departure timestamp
act_arr_lt	timestamp without time zone	actual arrival timestamp
flt_leg_seq_no	integer	flight leg sequence number
aircraft_name_scheduled	text	scheduled aircraft name
baggage_info	character varying(128)	baggage information
counter	character varying(128)	counter information
gate_info	character varying(128)	gate information
lounge_info	character varying(128)	lounge information
terminal_info	character varying(128)	terminal information
arr_terminal_info	character varying(128)	arrival terminal information
source_data	text	source of data
created_at	timestamp without time zone	record creation timestamp

Target/destination table

destination table sql

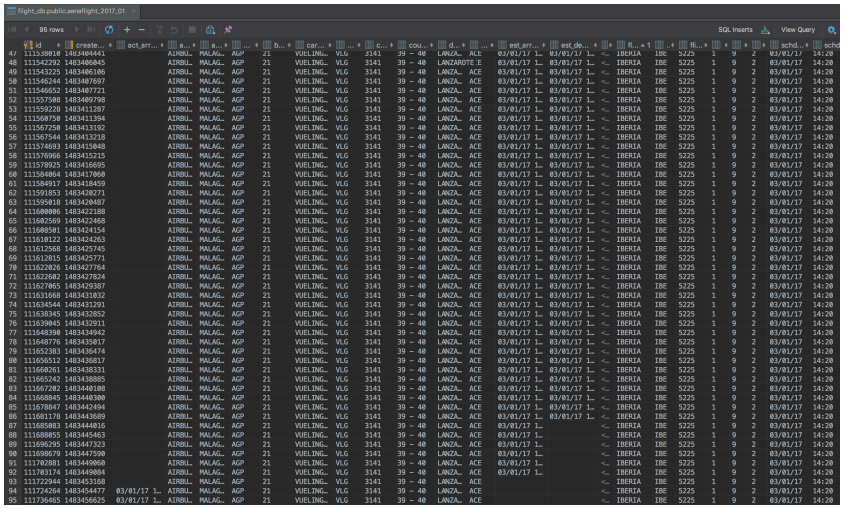
```
CREATE TABLE aenaflight_source
(
    id bigserial NOT NULL PRIMARY KEY,
    adep character varying(8) NOT NULL,
    ades character varying(8) NOT NULL,
    flight_code character varying(8) NOT NULL,
    flight_number character varying(8) NOT NULL,
    carrier_code character varying(8),
    carrier_number character varying(8),
    status_info character varying(256) NOT NULL,
    schd_dep_lt timestamp without time zone NOT NULL,
    schd_arr_lt timestamp without time zone NOT NULL,
    est_dep_lt timestamp without time zone,
    est_arr_lt timestamp without time zone,
    act_dep_lt timestamp without time zone,
    act_arr_lt timestamp without time zone,
    flt_leg_seq_no integer NOT NULL,
    aircraft_name_scheduled text,
    baggage_info character varying(128),
    counter character varying(128),
    gate_info character varying(128),
    lounge_info character varying(128),
    terminal_info character varying(128),
    arr_terminal_info character varying(128),
    source_data text,
    created_at timestamp without time zone NOT NULL,
);
```

Diagram

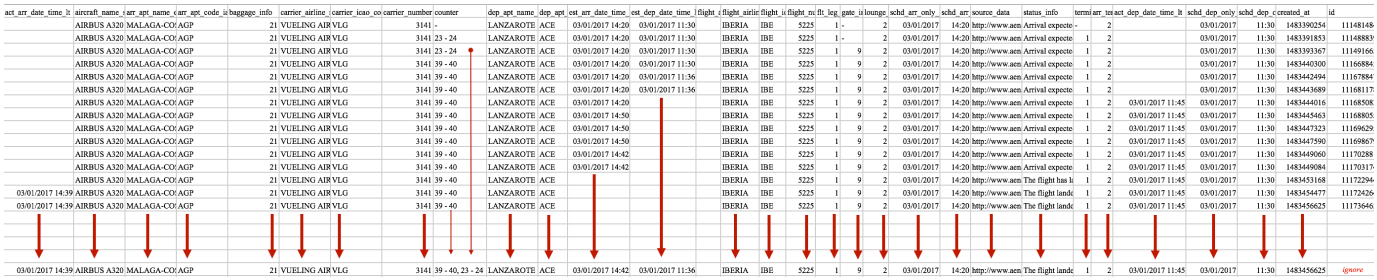


1.Extract

```
select * from aenaflight_2017_01 where flight_icao_code ='IBE' and
flight_number='5225' and schd_dep_only_date_lt='03/01/17' order by id asc;
```



2.Transform



3.Load

```
select * from destination_data where flight_code ='IBE' and
flight_number='5225' and schd_dep_lt::date='2017-01-03';
```

id	adep	ades	flight_code	flight_n	carrier_	carrier_	status_info	schd_dep_lt	schd_arr_lt	est_dep_lt	est_arr_lt	act_dep_lt	act_arr_lt	flt_leg	aircraft_nan	baggage	counter	gate_li	loung	terminal_inf	arr_ten	source_data	created_at
111736465	ACE	AGP	IBE	5225	VLG	3141	The flight landed	03/01/2017 11:30	03/01/2017 14:20	03/01/2017 11:36	03/01/2017 14:42	03/01/2017 11:45	03/01/2017 14:39	1	AIRBUS A32(	21	39 - 40,23 - 24,-	9	2	1	2	http://www.	03/01/2017 17:1