ETL FLIGHT DATA Practice

(이삭엔지니어링 인턴 김형근, fnfn9087@gmail.com)

(CDH 5.15, Oracle DB를 사용하여 실습하였음.)

1. Download the flight Data

링크 : Research and Innovative Technology Administration, Bureau of Transportation Statistics.

Download the flight data

1. Browse to Research and Innovative Technology Administration, Bureau of Transportation Statistics.
2. On the page, select the following values:

Name	Value
Filter Year	2013
Filter Period	January
Fields	Year, FlightDate, UniqueCarrier, Carrier, FlightNum, OriginAirportID, Origin, OriginCityName, OriginState, DestAirportID, Dest, DestCityName, DestState, DepDelayMinutes, ArrDelay, ArrDelayMinutes, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay.

Clear all other fields

3. Select **Download**. You get a .zip file with the data fields you selected.

On-Time : Reporting Carrier On-Time Performance (1987-present)

Download Instructions

Latest Available Data: September 2018

Filter Geography	Filter Year	Filter Period
All	2013	January

☐ Prezipped File ☐ % Missing ☐ Documentation ☐ Terms

Field Name	Description	Support Table
Time Period		
☑ Year	Year	
☐ Quarter	Quarter (1-4)	Get Lookup Table
☐ Month	Month	Get Lookup Table
☐ DayofMonth	Day of Month	
☐ DayOfWeek	Day of Week	Get Lookup Table
☑ FlightDate	Flight Date (yyyymmdd)	
Airline		
☑ Reporting_Airline	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.	Get Lookup Table
☑ DOT_ID_Reporting_Airline	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.	Get Lookup Table
☐ IATA_CODE_Reporting_Airline	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.	Get Lookup Table
☐ Tail_Number	Tail Number	
☑ Flight_Number_Reporting_Airline	Flight Number	

Origin		
☑ OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.	Get Lookup Table
☐ OriginAirportSeqID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.	Get Lookup Table
☐ OriginCityMarketID	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.	Get Lookup Table
☑ Origin	Origin Airport	Get Lookup Table
☑ OriginCityName	Origin Airport, City Name	
☑ OriginState	Origin Airport, State Code	Get Lookup Table
☐ OriginStateFips	Origin Airport, State Fips	Get Lookup Table
☑ OriginStateName	Origin Airport, State Name	
☐ OriginWac	Origin Airport, World Area Code	Get Lookup Table
Destination		
☑ DestAirportID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.	Get Lookup Table
☐ DestAirportSeqID	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.	Get Lookup Table
☐ DestCityMarketID	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate	Get Lookup Table

☑ Dest	Destination Airport	Get Lookup Table
☑ DestCityName	Destination Airport, City Name	
☑ DestState	Destination Airport, State Code	Get Lookup Table
☐ DestStateFips	Destination Airport, State Fips	Get Lookup Table
☐ DestStateName	Destination Airport, State Name	
☐ DestWac	Destination Airport, World Area Code	Get Lookup Table
Departure Performance		
☐ CRSDepTime	CRS Departure Time (local time: hhmm)	
☐ DepTime	Actual Departure Time (local time: hhmm)	
☐ DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.	
☑ DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.	
☐ DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)	Get Lookup Table
☐ DepartureDelayGroups	Departure Delay intervals, every (15 minutes from <-15 to >180)	Get Lookup Table
☐ DepTimeBlk	CRS Departure Time Block, Hourly Intervals	Get Lookup Table
☐ TaxiOut	Taxi Out Time, in Minutes	
☐ WheelsOff	Wheels Off Time (local time: hhmm)	
Arrival Performance		
☐ WheelsOn	Wheels On Time (local time: hhmm)	
☐ TaxiIn	Taxi In Time, in Minutes	
☐ CRSArrTime	CRS Arrival Time (local time: hhmm)	
☐ ArrTime	Actual Arrival Time (local time: hhmm)	
☑ ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.	
☑ ArrDelayMinutes	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.	
☐ ArrDel15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)	Get Lookup Table
☐ ArrivalDelayGroups	Arrival Delay intervals, every (15-minutes from <-15 to >180)	Get Lookup Table
Cause of Delay (Data starts 6/2003)		
☑ CarrierDelay	Carrier Delay, in Minutes	
☑ WeatherDelay	Weather Delay, in Minutes	
☑ NASDelay	National Air System Delay, in Minutes	
☑ SecurityDelay	Security Delay, in Minutes	
☑ LateAircraftDelay	Late Aircraft Delay, in Minutes	

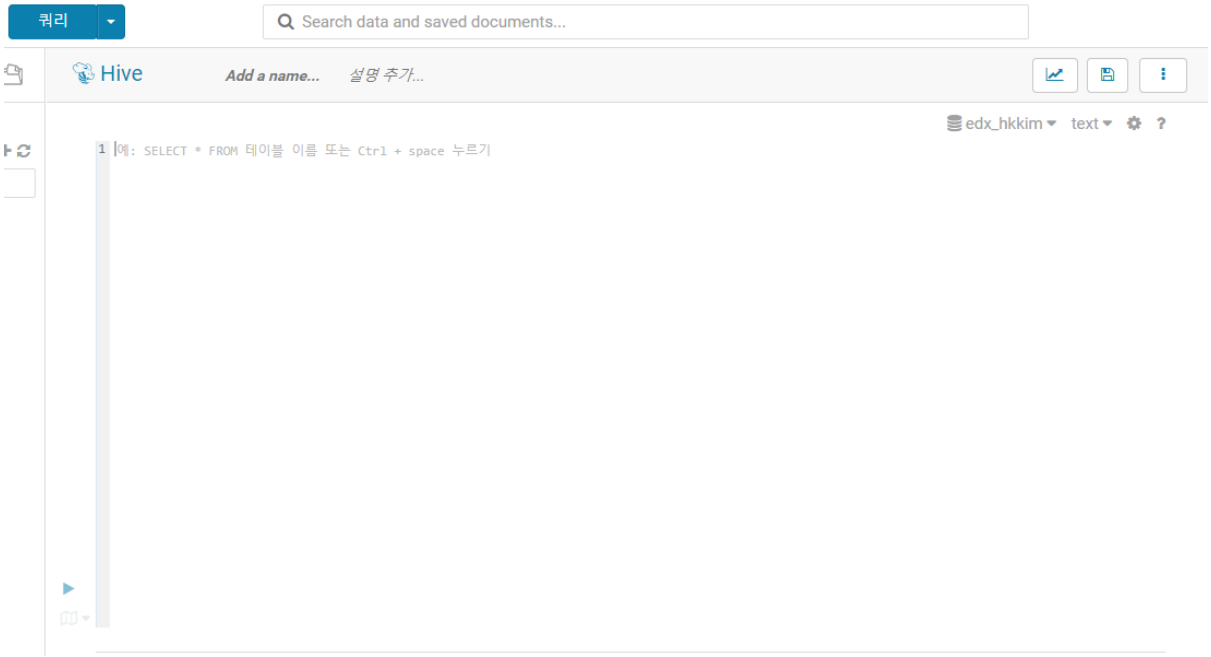
2. Use the following commands to create a directory, and copy the *.csv* file to the directory

hdfs dfs -mkdir -p hdfs://iseHA1/user/hkkim/tutorials/flightdelays/data

hdfs dfs -put <FILE_NAME>.csv hdfs://iseHA1/user/hkkim/tutorials/flightdelays/data/

3. In the Hue interface('하둡 클러스터 마스터 노드 address':8888), enter the following queries in Hive Query Interface

(You must proceed "edx_hkkim" database.)




```
DROP TABLE delays_raw;

-- Creates an external table over the csv file

CREATE EXTERNAL TABLE delays_raw (
YEAR string,
FL_DATE string,
UNIQUE_CARRIER string,
CARRIER string,
FL_NUM string,
ORIGIN_AIRPORT_ID string,
ORIGIN string,
ORIGIN_CITY_NAME string,
ORIGIN_CITY_NAME_TEMP string,
ORIGIN_STATE_ABR string,
DEST_AIRPORT_ID string,
DEST string,
DEST_CITY_NAME string,
DEST_CITY_NAME_TEMP string,
DEST_STATE_ABR string,
```

```
DEP_DELAY_NEW float,
ARR_DELAY_NEW float,
CARRIER_DELAY float,
WEATHER_DELAY float,
NAS_DELAY float,
SECURITY_DELAY float,
LATE_AIRCRAFT_DELAY float)

-- The following lines describe the format and location of the file
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION ‘user/hkkim/tutorials/flightdelays/data’;


-- Drop the delays table if it exists
DROP TABLE delays;


-- Create the delays table and populate it with data
-- pulled in from the CSV file (via the external table defined previously)
CREATE TABLE delays

LOCATION ‘user/hkkim/tutorials/flightdelays/processed’
AS
SELECT YEAR AS year, FL_DATE AS flight_date,
substring(UNIQUE_CARRIER, 2, length(UNIQUE_CARRIER) -1) AS unique_carrier,
substring(CARRIER, 2, length(CARRIER) -1) AS carrier,
substring(FL_NUM, 2, length(FL_NUM) -1) AS flight_num,
ORIGIN_AIRPORT_ID AS origin_airport_id,
substring(ORIGIN, 2, length(ORIGIN) -1) AS origin_airport_code,
substring(ORIGIN_CITY_NAME, 2) AS origin_city_name,
substring(ORIGIN_STATE_ABR, 2, length(ORIGIN_STATE_ABR) -
1)  AS origin_state_abr,
DEST_AIRPORT_ID AS dest_airport_id,
substring(DEST, 2, length(DEST)-1) AS dest_airport_code,
substring(DEST_CITY_NAME,2) AS dest_city_name,
substring(DEST_STATE_ABR, 2, length(DEST_STATE_ABR) -1) AS dest_state_abr,
DEP_DELAY_NEW AS dep_delay_new,
ARR_DELAY_NEW AS arr_delay_new,
CARRIER_DELAY AS carrier_delay,
WEATHER_DELAY AS weather_delay,
NAS_DELAY AS nas_delay,
SECURITY_DELAY AS security_delay,
LATE_AIRCRAFT_DELAY AS late_aircraft_delay
FROM delays_raw;
```

4. After finish "3." , enter the following queries.

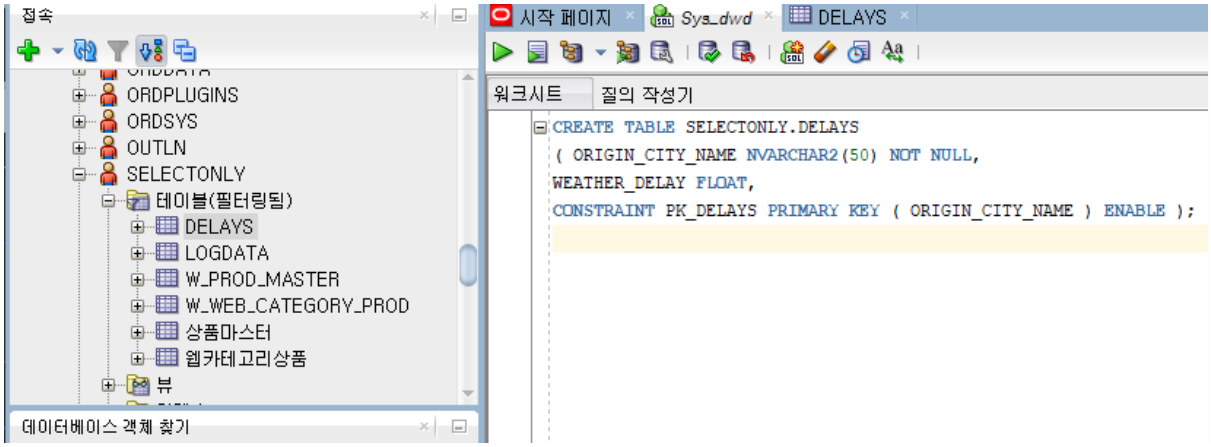
```
INSERT OVERWRITE DIRECTORY '/user/hkkim/tutorials/flightdelays/output'
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
SELECT regexp_replace(origin_city_name, '''', ''),avg(weather_delay)
FROM delays
WHERE weather_delay IS NOT NULL
GROUP BY origin_city_name;
```

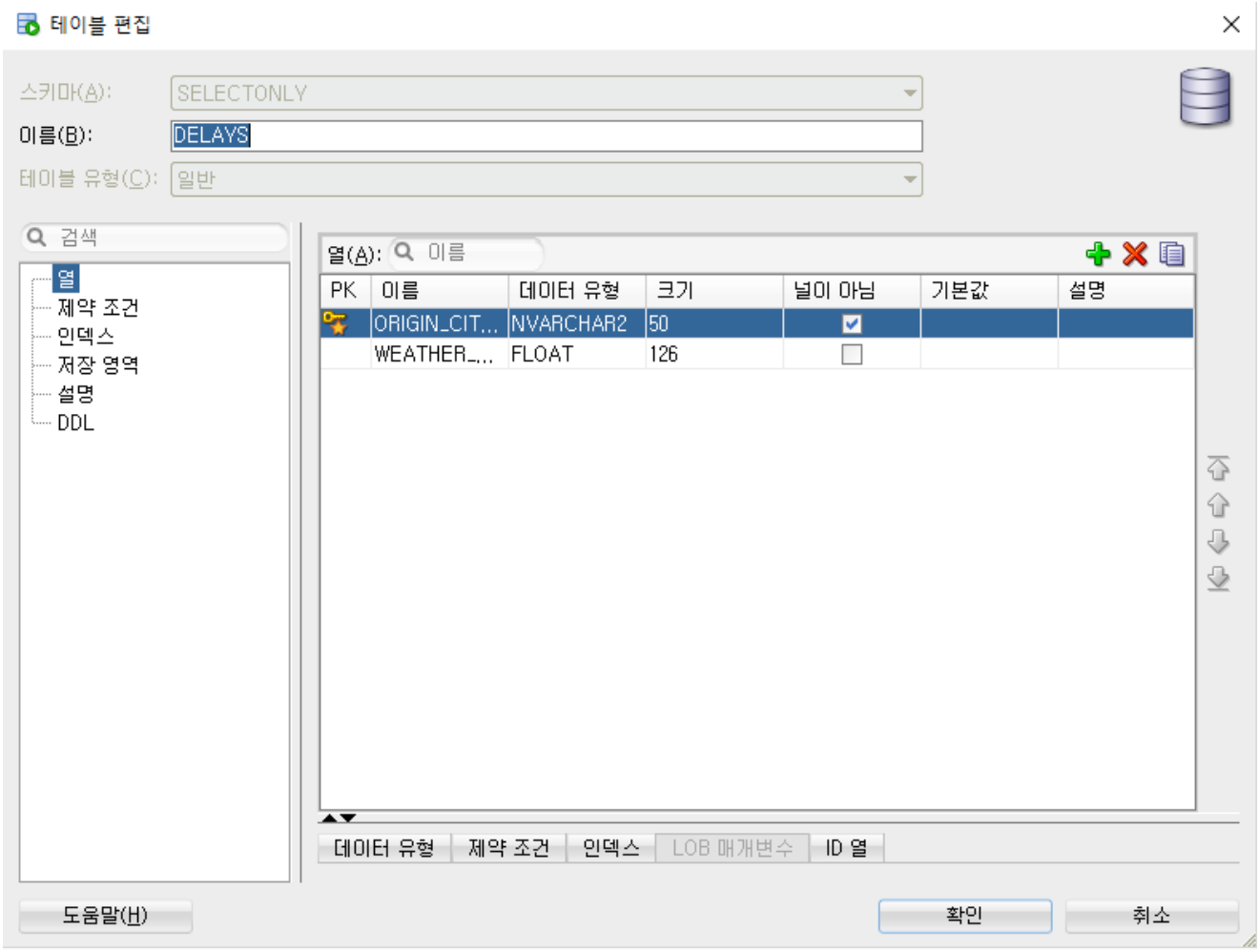
5. In Oracle DB, login to "username = sys as sysdba, password = sys". DB's name is "dwd" And, Go to user named "SELECTONLY", and Create Tables.

```
CREATE TABLE SELECTONLY.DELAY
( ORIGIN_CITY_NAME NVARCHAR2(50) NOT NULL,
  WEATHER_DELAY FLOAT,
  CONSTRAINT DELAY_PK PRIMARY KEY ( ORIGIN_CITY_NAME ) ENABLE
);
```

원문은 아래이나, CONSTRAINT [PK_delays] PRIMARY KEY CLUSTERED 구문이 Oracle 에서 문법 오류가 계속 발생, 그래서 Oracle 에서 위 구문으로만 진행.

```
([origin_city_name] ASC)
"CREATE TABLE [dbo].[delays](
[origin_city_name] [nvarchar](50) NOT NULL,
[weather_delay] float,
CONSTRAINT [PK_delays] PRIMARY KEY CLUSTERED
([origin_city_name] ASC))"
```





6. Use Sqoop export, and insert data to “DELAYS” tables.

```
sqoop export
--connect jdbc:oracle:thin:@10.100.3.152:1521:dwd
--username selectonly --password ise1212
--direct
--export-dir /user/hkkim/tutorials/flightdelays/output
--table DELAYS --fields-terminated-by '\t' --m 1
```

	ORIGIN_CITY_NAME	WEATHER_DELAY
1	Fairbanks	12.28888888888889
2	Marquette	22.77777777777778
3	Modesto	36.16
4	Orlando	20.24937447873228
5	Key West	13.658536585365853
6	Texarkana	80.58823529411765
7	Seattle	15.989651928504234
8	Des Moines	11.46774193548387
9	Akron	7.198198198198198
10	San Antonio	15.043165467625899
11	Green Bay	19.16867469879518
12	Jackson	30.136986301369863
13	New York	19.193140794223826
14	Portland	15.444073455759598
15	Staunton	9
16	Manhattan/Ft. Riley	28.88
17	Deadhorse	6
18	Del Rio	2.7777777777777777
19	Aguadilla	15.5625
20	La Crosse	14.166666666666666
21	Moline	23.96629213483146
22	Melbourne	31.333333333333332
23	Dayton	25.819148936170212
24	Pasco/Kennewick/Richland	9.136363636363637
25	San Francisco	16.615592853275583
26	Grand Island	2.6363636363636362
27	Columbia	28.846153846153847

Reference:
https://docs.microsoft.com/en-us/azure/storage/data-lake-storage/tutorial-extract-transform-load-hive