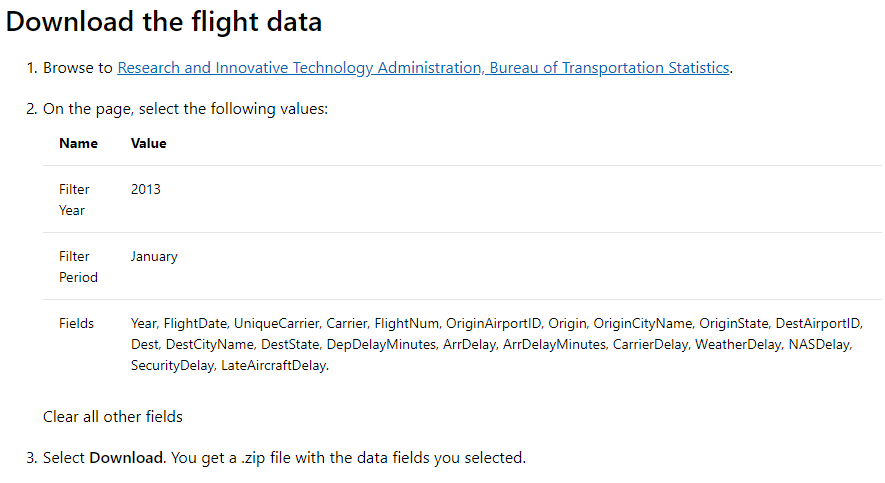
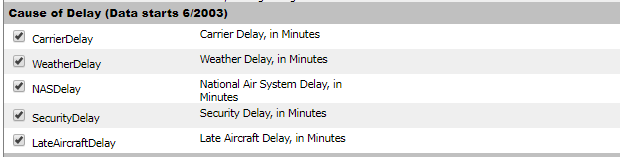
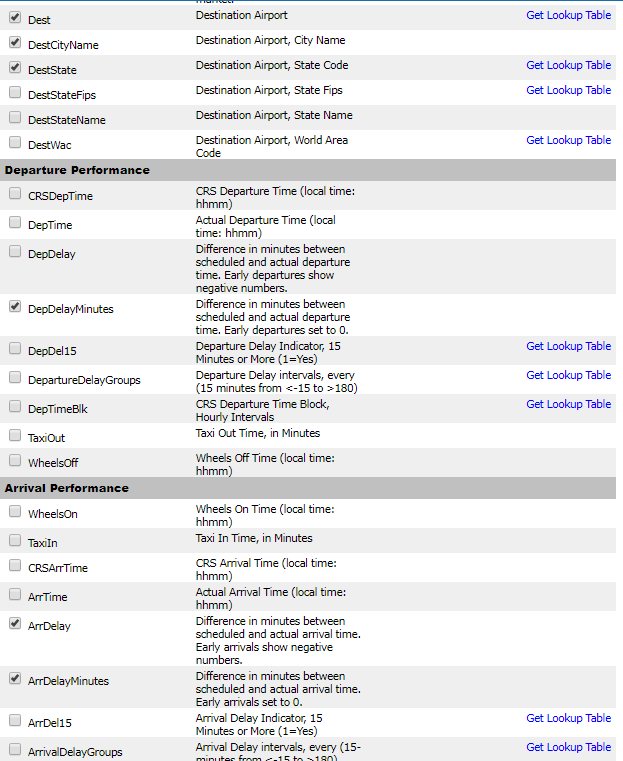
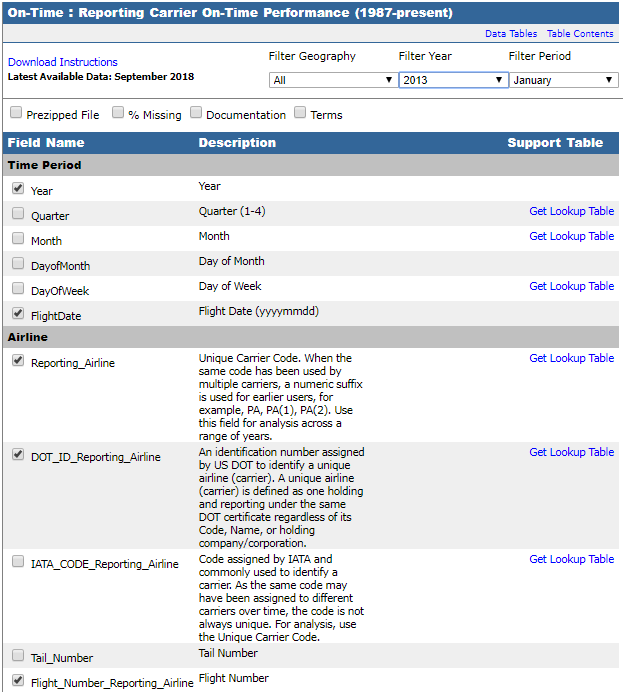
ETL FLIGHT DATA Practice  
(이삭엔지니어링 인턴 김형근, fnfn9087@gmail.com)  
(CDH 5.15, Oracle DB를 사용하여 실습하였음.)

**1. Download the flight Data**

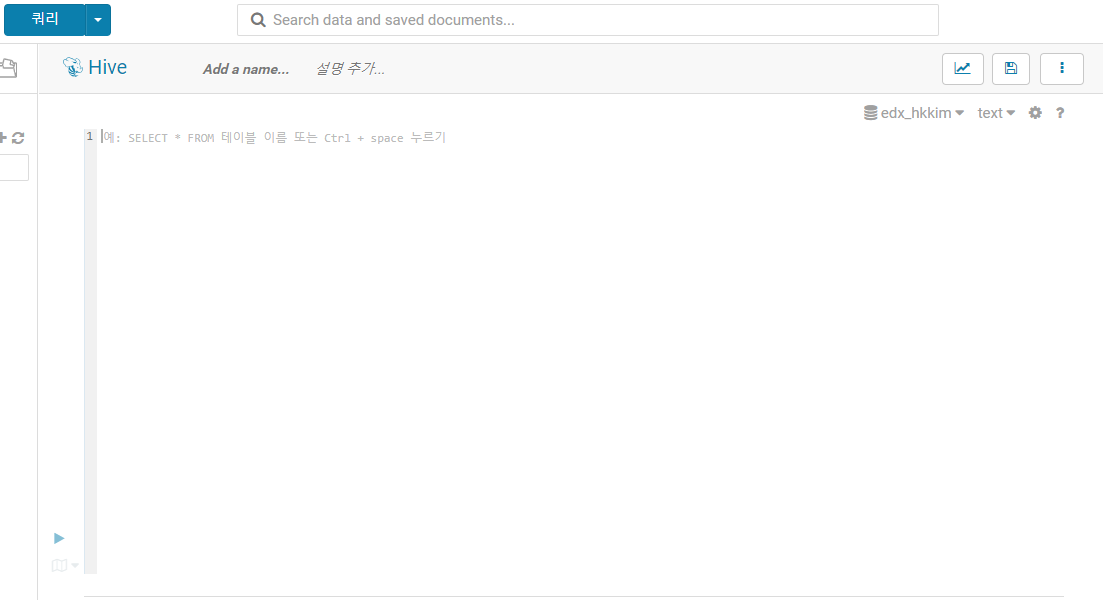
링크 : [Research and Innovative Technology Administration, Bureau of Transportation Statistics](http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time).

**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(10.100.3.211:8888), 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’

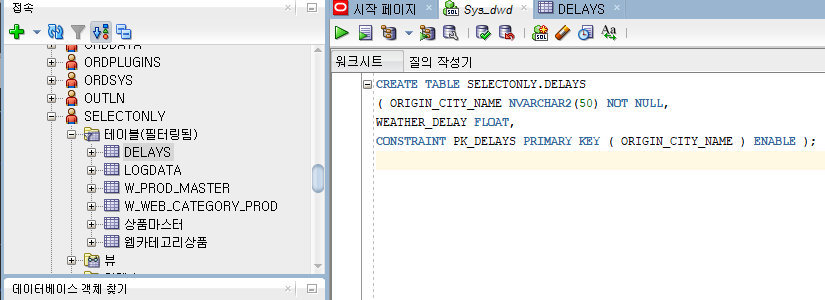
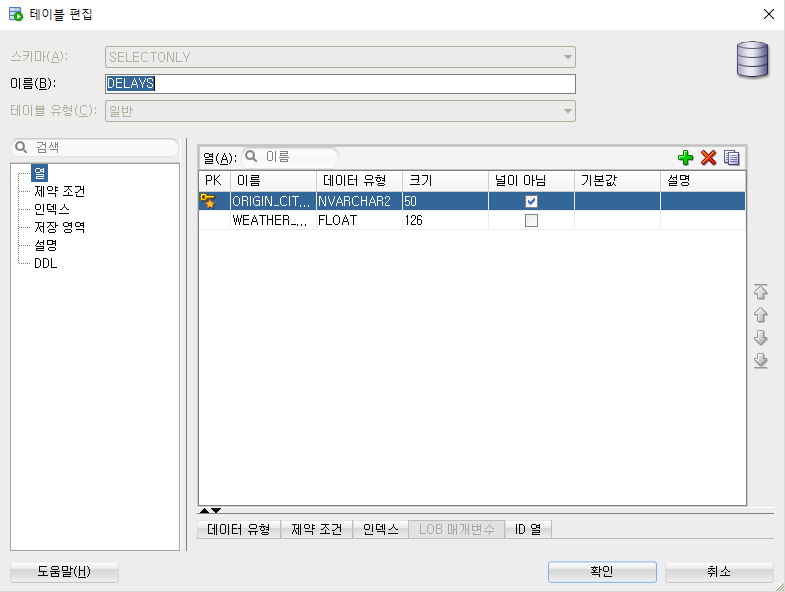
AS  
SELECT YEAR AS year,  
FL\_DATE AS flight\_date,  
substring(UNIQUE\_CARRIER, 2, length(UNIQUE\_CARRIER) -1) AS unique\_carrier,  
substring(CARRIER, 2, length(CARRIER) -1) AS carrier,  
substring(FL\_NUM, 2, length(FL\_NUM) -1) AS flight\_num,  
ORIGIN\_AIRPORT\_ID AS origin\_airport\_id,  
substring(ORIGIN, 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 Oracle에서 문법 오류가 계속 발생. 그래서 Oracle에서 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   
  


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