Hive Assignment #3

Data Set: The NBER U.S. Patent Data
Data Link: http://www.nber.org/patents/

We will use two of the data sets:

1. Patent data, including constructed variables:
Data link: http://www.nber.org/patents/pat63_99.zip
Summary of this data is given here - http://www.nber.org/patents/pat63_99.txt

2. Class codes with corresponding class names:
Data Link: http://www.nber.org/patents/list_of_classes.txt
Summary of data: This data need to be cleaned in a way so that it only contains class Id and title in tab separated form. Just remove first 9 lines.

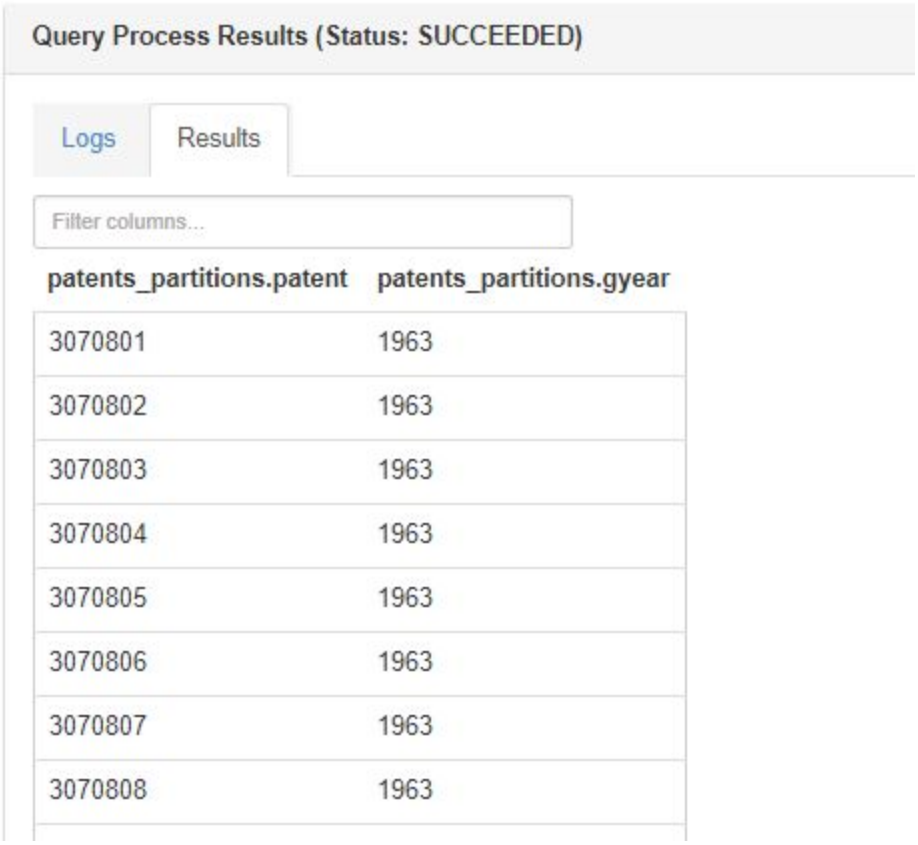
Problems
For this problem we have divided data for patents into multiple files, one file containing patents granted
in one particular year.

1. Create an external table for Patent data set so that it can be used efficiently for queries which require looking into
patents granted for given year.

```
DROP TABLE IF EXISTS temp;
CREATE EXTERNAL TABLE temp
(
  PATENT INT, GYEAR INT,  GDATE INT,
  APPYEAR INT, COUNTRY STRING, POSTATE STRING,
  ASSIGNEE INT, ASSCODE INT, CLAIMS INT,
  NCLASS INT, CAT INT, SUBCAT INT, CMADE INT,
  CRECEIVE INT, RATIOCIT DECIMAL(10,5),
  GENERAL DECIMAL(10,5), ORIGINAL DECIMAL(10,5),
  FWDAPLAG DECIMAL(10,5), BCKGTLAG DECIMAL(10,5),
  SELFCTUB DECIMAL(10,5), SELFCTLB DECIMAL(10,5),
  SECDUPBD DECIMAL(10,5), SECDLWBD DECIMAL(10,5)
)
```

```
      ROW FORMAT DELIMITED
      FIELDS TERMINATED BY ','
      LINES TERMINATED BY '\n'
      STORED AS TEXTFILE
LOCATION '/user/maria_dev/hive_assignment/patent_parts';

--test
--SELECT * FROM temp LIMIT 20;

DROP TABLE IF EXISTS patents_partitions;
CREATE  TABLE patents_partitions
(
  PATENT INT, GDATE INT, APPYEAR INT, COUNTRY STRING,
  POSTATE STRING, ASSIGNEE INT, ASSCODE INT, CLAIMS INT,
  NCLASS INT, CAT INT, SUBCAT INT, CMADE INT, CRECEIVE INT,
  RATIOCIT DECIMAL(10,5), GENERAL DECIMAL(10,5),
  ORIGINAL DECIMAL(10,5), FWDAPLAG DECIMAL(10,5),
  BCKGTLAG DECIMAL(10,5), SELFCTUB DECIMAL(10,5),
  SELFCTLB DECIMAL(10,5), SECDUPBD DECIMAL(10,5),
  SECDLWBD DECIMAL(10,5)
)
PARTITIONED BY (GYEAR INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS SEQUENCEFILE
LOCATION '/maria_dev/hive_assignments/';

--test
--DESCRIBE FORMATTED patents_partitions;
--SHOW PARTITIONS patents_partitions;

set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
INSERT OVERWRITE TABLE patents_partitions
      PARTITION (GYEAR)
      SELECT
            PATENT, GDATE, APPYEAR, COUNTRY, POSTATE,
            ASSIGNEE, ASSCODE, CLAIMS, NCLASS, CAT,
            SUBCAT, CMADE, CRECEIVE, RATIOCIT, GENERAL,
            ORIGINAL, FWDAPLAG, BCKGTLAG, SELFCTUB,
            SELFCTLB, SECDUPBD, SECDLWBD, GYEAR
```

```
FROM temp;

--test
SELECT * FROM patents_pa
```



Query Process Results (Status: SUCCEEDED)

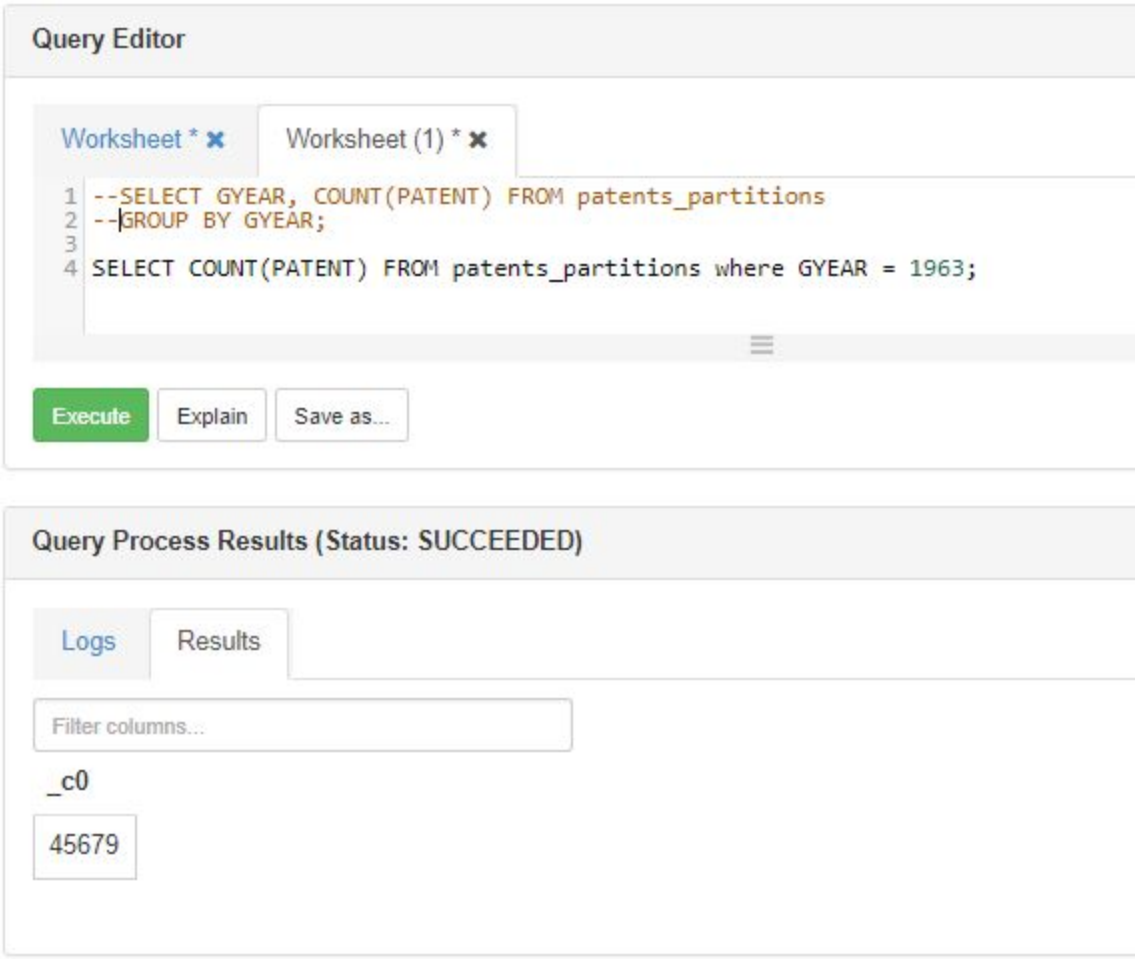
Logs | Results

Filter columns...

patents_partitions.patent	patents_partitions.gyear
3070801	1963
3070802	1963
3070803	1963
3070804	1963
3070805	1963
3070806	1963
3070807	1963
3070808	1963

--2. Find out number of patents granted in year 1963.

```
SELECT  COUNT(PATENT) FROM patents_partitions where GYEAR = 1963;
```



--3. Find out number of patents granted in each country in year 1999.

```
SELECT COUNTRY, COUNT(PATENT) AS patents from patents_partitions
WHERE GYEAR = 1999
GROUP BY COUNTRY
ORDER BY patents;
```

Query Editor

Worksheet * ✖ Worksheet (1) ✖

```
1 SELECT COUNTRY, COUNT(PATENT) AS patents
2 FROM patents_partitions
3 WHERE GYEAR = 1999
4 GROUP BY COUNTRY
5 ORDER BY patents DESC;
```

Execute Explain Save as...

100%

Query Process Results (Status: SUCCEEDED)

Logs Results

Filter columns...

country	patents
"US"	83906
"JP"	31104
"DE"	9337
"FR"	3820
"TW"	3693
"GB"	3572
"KR"	3562
"CA"	3226
"IT"	1492
"SE"	1401
"CH"	1279
"NL"	1247
"IL"	743
"AU"	707