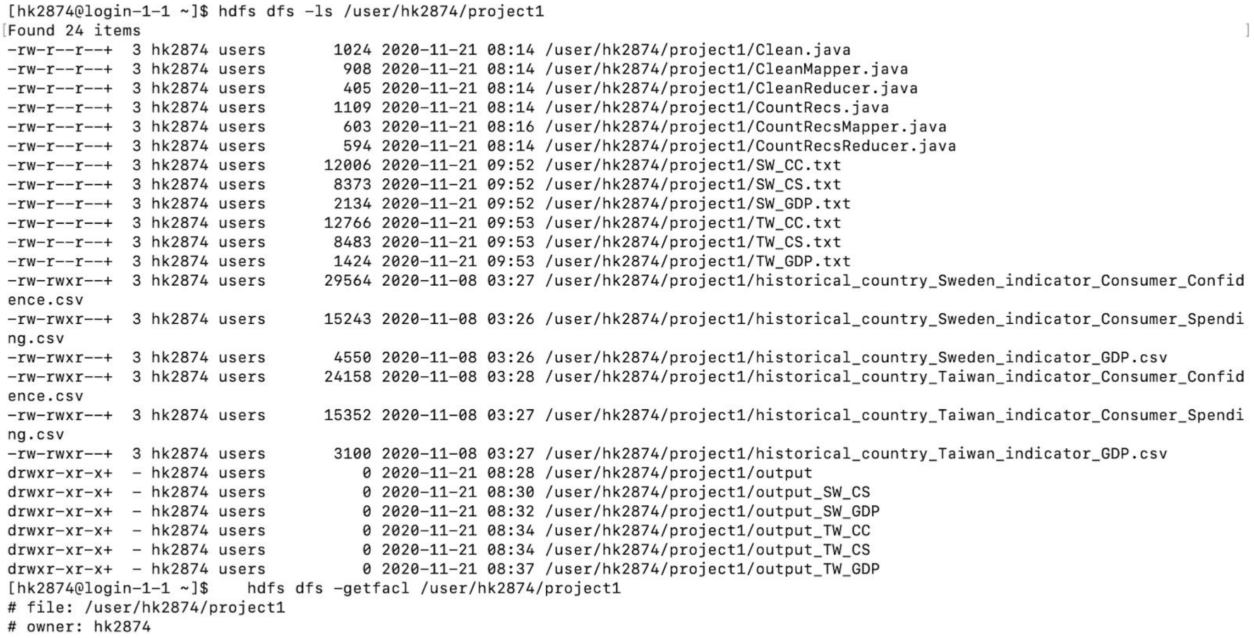
Hw 11: Analytics Code

Part 1:

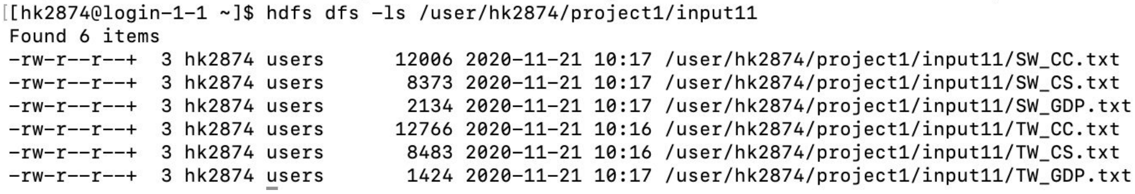
From the last homework, we organized cleaned data files in our project directory.



```
[hk2874@login-1-1 ~]$ hdfs dfs -ls /user/hk2874/project1
Found 24 items
-rw-r--r--+  3 hk2874 users       1024 2020-11-21 08:14 /user/hk2874/project1/Clean.java
-rw-r--r--+  3 hk2874 users        908 2020-11-21 08:14 /user/hk2874/project1/CleanMapper.java
-rw-r--r--+  3 hk2874 users        405 2020-11-21 08:14 /user/hk2874/project1/CleanReducer.java
-rw-r--r--+  3 hk2874 users       1109 2020-11-21 08:14 /user/hk2874/project1/CountRecs.java
-rw-r--r--+  3 hk2874 users        603 2020-11-21 08:16 /user/hk2874/project1/CountRecsMapper.java
-rw-r--r--+  3 hk2874 users        594 2020-11-21 08:14 /user/hk2874/project1/CountRecsReducer.java
-rw-r--r--+  3 hk2874 users      12006 2020-11-21 09:52 /user/hk2874/project1/SW_CC.txt
-rw-r--r--+  3 hk2874 users       8373 2020-11-21 09:52 /user/hk2874/project1/SW_CS.txt
-rw-r--r--+  3 hk2874 users       2134 2020-11-21 09:52 /user/hk2874/project1/SW_GDP.txt
-rw-r--r--+  3 hk2874 users      12766 2020-11-21 09:53 /user/hk2874/project1/TW_CC.txt
-rw-r--r--+  3 hk2874 users       8483 2020-11-21 09:53 /user/hk2874/project1/TW_CS.txt
-rw-r--r--+  3 hk2874 users       1424 2020-11-21 09:53 /user/hk2874/project1/TW_GDP.txt
-rw-rwxr--+  3 hk2874 users      29564 2020-11-08 03:27 /user/hk2874/project1/historical_country_Sweden_indicator_Consumer_Confid
ence.csv
-rw-rwxr--+  3 hk2874 users      15243 2020-11-08 03:26 /user/hk2874/project1/historical_country_Sweden_indicator_Consumer_Spendi
ng.csv
-rw-rwxr--+  3 hk2874 users       4550 2020-11-08 03:26 /user/hk2874/project1/historical_country_Sweden_indicator_GDP.csv
-rw-rwxr--+  3 hk2874 users      24158 2020-11-08 03:28 /user/hk2874/project1/historical_country_Taiwan_indicator_Consumer_Confid
ence.csv
-rw-rwxr--+  3 hk2874 users      15352 2020-11-08 03:27 /user/hk2874/project1/historical_country_Taiwan_indicator_Consumer_Spendi
ng.csv
-rw-rwxr--+  3 hk2874 users       3100 2020-11-08 03:27 /user/hk2874/project1/historical_country_Taiwan_indicator_GDP.csv
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:28 /user/hk2874/project1/output
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:30 /user/hk2874/project1/output_SW_CS
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:32 /user/hk2874/project1/output_SW_GDP
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:34 /user/hk2874/project1/output_TW_CC
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:34 /user/hk2874/project1/output_TW_CS
drwxr-xr-x+  - hk2874 users          0 2020-11-21 08:37 /user/hk2874/project1/output_TW_GDP
[hk2874@login-1-1 ~]$    hdfs dfs -getfacl /user/hk2874/project1
# file: /user/hk2874/project1
# owner: hk2874
```

- Created directory to store hive input
  hdfs dfs -ls user/hk2874/project1/hive_hw11 //for output

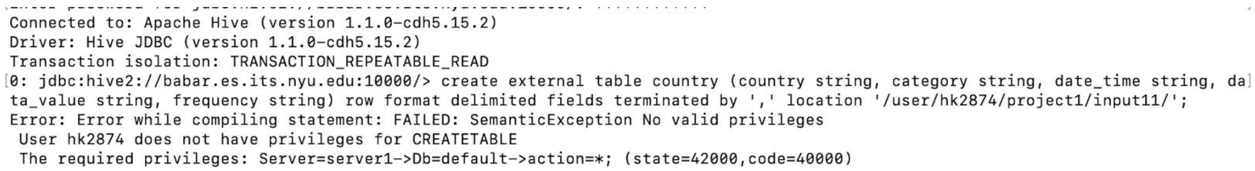
- Created input directory and moved all files to the input directory for HIVE
  hdfs dfs -ls /user/hk2874/project1/input11 // for hive input



```
[hk2874@login-1-1 ~]$ hdfs dfs -ls /user/hk2874/project1/input11
Found 6 items
-rw-r--r--+  3 hk2874 users      12006 2020-11-21 10:17 /user/hk2874/project1/input11/SW_CC.txt
-rw-r--r--+  3 hk2874 users       8373 2020-11-21 10:17 /user/hk2874/project1/input11/SW_CS.txt
-rw-r--r--+  3 hk2874 users       2134 2020-11-21 10:17 /user/hk2874/project1/input11/SW_GDP.txt
-rw-r--r--+  3 hk2874 users      12766 2020-11-21 10:16 /user/hk2874/project1/input11/TW_CC.txt
-rw-r--r--+  3 hk2874 users       8483 2020-11-21 10:16 /user/hk2874/project1/input11/TW_CS.txt
-rw-r--r--+  3 hk2874 users       1424 2020-11-21 10:17 /user/hk2874/project1/input11/TW_GDP.txt
```

- We kept getting privilege error messages when we tried to create table.
- We even tried the hive queries from previous homework again (worked fine)

  create external table country (country string, category string, date_time string, data_value string, frequency string) row format delimited fields terminated by ',' location '/user/hk2874/project1/input11/';



```
Connected to: Apache Hive (version 1.1.0-cdh5.15.2)
Driver: Hive JDBC (version 1.1.0-cdh5.15.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
[0: jdbc:hive2://babar.es.its.nyu.edu:10000/> create external table country (country string, category string, date_time string, da]
ta_value string, frequency string) row format delimited fields terminated by ',' location '/user/hk2874/project1/input11/';
Error: Error while compiling statement: FAILED: SemanticException No valid privileges
 User hk2874 does not have privileges for CREATETABLE
 The required privileges: Server=server1->Db=default->action=*; (state=42000,code=40000)
```

- After trying all different kinds of queries and creating different directory files, we realized it was because of a database connection problem. With these queries, we created tables successfully.

  use_*eachofournetid;*

Part 2: Code

Below are the code we used to create table

sw_cs = “Sweden Consumer Spending”

```
create external table sw_cs (country string, category string, date_time string, data_value int, frequency string) row format delimited fields terminated by ',' location '/user/hk2874/hiveInput/';
```

tw_cs = “Taiwan Consumer Spending”

```
create external table tw_cs (country string, category string, date_time string, data_value int, frequency string) row format delimited fields terminated by ',' location '/user/hk2874/hiveInputTW/';
```

sw_cc = “Sweden Consumer Consumption”

```
create external table sw_cc (country string, category string, date_time string, data_value int, frequency string) row format delimited fields terminated by ',' location '/user/sh4023/hiveInput2/';
```

tw_cc = “Taiwan Consumer Consumption”

```
create external table tw_cc (country string, category string, date_time string, data_value int, frequency string) row format delimited fields terminated by ',' location '/user/sh4023/hiveInput/';
```

sweden_parks = “Sweden Parks”

```
create external table sweden_parks(entity string, code string, date string, parks float) row format delimited fields terminated by ',' location '/user/sh4023/sweden_parks_input/‘;
```

sweden_res = “Sweden Residence”

```
create external table sweden_res(entity string, code string, date string, parks float) row format delimited fields terminated by ',' location '/user/sh4023/sweden_res_input/‘;
```

sweden_groc = “Sweden Grocery”

```
create external table sweden_groc(entity string, code string, date string, parks float) row format delimited fields terminated by ',’ location '/user/sh4023/sweden_groc_input/‘;
```

taiwan_parks = “Taiwan Parks”

```
create external table taiwan_parks(entity string, code string, date string, parks float) row format delimited fields terminated by ',' location '/user/sh4023/taiwan_parks_input/‘;
```

taiwan_res = “Taiwan Residence”

```
create external table taiwan_res(entity string, code string, date string, parks float) row format delimited fields terminated by ','
location '/user/sh4023/taiwan_res_input/‘;
```

taiwan_groc = “Taiwan Grocery”

```
create external table taiwan_groc(entity string, code string, date string, parks float) row format delimited fields terminated by
',' location '/user/sh4023/taiwain_groc_input/‘;
```

QUERIES WE TRIED TO ANALYZE DATA

- Query to compare Taiwan and Sweden original data_values

```
select sw.date_time, sw.data_value  as Sweden_val, tw.date_value as Taiwan_value
from sw_cs sw
inner join tw_cs tw on tw.date_time =sw.date_time;
```

-sample output

-

- //See average for taiwan and sweden values

```
select substring(date_time, 0,4) as YEAR, avg(data_value)
from sw_cs
group by substring(date_time, 0,4);
```

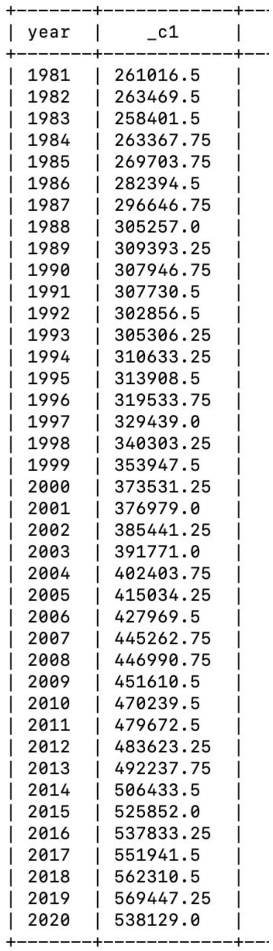
- Calculated standard deviation

```
select STDDEV(data_value) as standard_deviation from sw_cc;
select STDDEV(data_value) as standard_deviation from tw_cc;
select STDDEV(data_value) as standard_deviation from sw_cs;
select STDDEV(data_value) as standard_deviation from tw_cs;
```

-sample outputs

```
+--------+--+
|  _c0   |
+--------+--+
| 15.0   |
| 17.0   |
| 18.0   |
| 20.0   |
| 22.0   |
| 24.0   |
| 26.0   |
| 29.0   |
| 31.0   |
| 33.0   |
| 38.0   |
| 41.0   |
| 48.0   |
| 59.0   |
| 66.0   |
| 82.0   |
| 89.0   |
| 94.0   |
| 104.0  |
| 123.0  |
| 142.0  |
| 129.0  |
| 114.0  |
| 105.0  |
| 109.0  |
| 114.0  |
| 150.0  |
| 183.0  |
| 206.0  |
| 217.0  |
| 261.0  |
| 274.0  |
| 284.0  |
| 212.0  |
| 229.0  |
| 267.0  |
| 291.0  |
| 268.0  |
| 270.0  |
| 274.0  |
| 262.0  |
| 242.0  |
| 266.0  |
| 334.0  |
| 385.0  |
| 392.0  |
| 423.0  |
| 491.0  |
| 517.0  |
| 436.0  |
| 495.0  |
| 574.0  |
| 552.0  |
| 586.0  |
| 581.0  |
| 505.0  |
| 515.0  |
| 541.0  |
| 555.0  |
| 530.0  |
+--------+--+
```

```
+--------+--+
|  _c0   |
+--------+--+
| 42.0   |
| 48.0   |
| 49.0   |
| 54.0   |
| 61.0   |
| 63.0   |
| 78.0   |
| 105.0  |
| 126.0  |
| 152.0  |
| 166.0  |
| 187.0  |
| 223.0  |
| 236.0  |
| 256.0  |
| 279.0  |
| 292.0  |
| 303.0  |
| 280.0  |
| 304.0  |
| 331.0  |
| 300.0  |
| 308.0  |
| 318.0  |
| 348.0  |
| 375.0  |
| 388.0  |
| 408.0  |
| 417.0  |
| 392.0  |
| 446.0  |
| 485.0  |
| 495.0  |
| 511.0  |
| 530.0  |
| 525.0  |
| 531.0  |
| 574.0  |
| 589.0  |
| 605.0  |
+--------+--+
```



```
+-------+------------+--
| year  |    _c1     |
+-------+------------+--
| 1981  | 261016.5   |
| 1982  | 263469.5   |
| 1983  | 258401.5   |
| 1984  | 263367.75  |
| 1985  | 269703.75  |
| 1986  | 282394.5   |
| 1987  | 296646.75  |
| 1988  | 305257.0   |
| 1989  | 309393.25  |
| 1990  | 307946.75  |
| 1991  | 307730.5   |
| 1992  | 302856.5   |
| 1993  | 305306.25  |
| 1994  | 310633.25  |
| 1995  | 313908.5   |
| 1996  | 319533.75  |
| 1997  | 329439.0   |
| 1998  | 340303.25  |
| 1999  | 353947.5   |
| 2000  | 373531.25  |
| 2001  | 376979.0   |
| 2002  | 385441.25  |
| 2003  | 391771.0   |
| 2004  | 402403.75  |
| 2005  | 415034.25  |
| 2006  | 427969.5   |
| 2007  | 445262.75  |
| 2008  | 446990.75  |
| 2009  | 451610.5   |
| 2010  | 470239.5   |
| 2011  | 479672.5   |
| 2012  | 483623.25  |
| 2013  | 492237.75  |
| 2014  | 506433.5   |
| 2015  | 525852.0   |
| 2016  | 537833.25  |
| 2017  | 551941.5   |
| 2018  | 562310.5   |
| 2019  | 569447.25  |
| 2020  | 538129.0   |
+-------+------------+--
```

```
alter table sw_cc change date_time date_time date

select substring(date_time, 0,4) as year, avg(data_value) from sw_cc group by substring(date_time, 0, 4);
+-------+---------------------+--+
| year  |         _c1         |
+-------+---------------------+--+
| 1993  | 60.833333333333336  |
| 1994  | 82.25               |
| 1995  | 70.66666666666667   |
| 1996  | 78.5                |
| 1997  | 91.16666666666667   |
| 1998  | 102.0               |
| 1999  | 109.75              |
| 2000  | 118.33333333333333  |
| 2001  | 99.41666666666667   |
| 2002  | 102.5               |
| 2003  | 96.0                |
| 2004  | 102.91666666666667  |
| 2005  | 106.0               |
| 2006  | 110.33333333333333  |
| 2007  | 110.41666666666667  |
| 2008  | 84.41666666666667   |
| 2009  | 90.16666666666667   |
| 2010  | 108.33333333333333  |
| 2011  | 99.58333333333333   |
| 2012  | 94.5                |
| 2013  | 98.16666666666667   |
| 2014  | 99.83333333333333   |
| 2015  | 98.33333333333333   |
| 2016  | 99.08333333333333   |
| 2017  | 104.33333333333333  |
| 2018  | 100.66666666666667  |
| 2019  | 94.0                |
| 2020  | 86.4                |
+-------+---------------------+--+
```

```
select substring(date_time, 0,4) as year, avg(data_value) from tw_cc group by substring(date_time, 0, 4);
```

```
+-------+---------------------+--+
| year  |         _c1         |
+-------+---------------------+--+
| 1999  | 87.0                |
| 2000  | 84.0                |
| 2001  | 65.16666666666667   |
| 2002  | 74.91666666666667   |
| 2003  | 78.0                |
| 2004  | 77.16666666666667   |
| 2005  | 73.41666666666667   |
| 2006  | 68.25               |
| 2007  | 65.83333333333333   |
| 2008  | 58.333333333333336  |
| 2009  | 54.083333333333336  |
| 2010  | 77.0                |
| 2011  | 83.83333333333333   |
| 2012  | 75.66666666666667   |
| 2013  | 76.16666666666667   |
| 2014  | 83.5                |
| 2015  | 87.5                |
| 2016  | 79.08333333333333   |
| 2017  | 79.66666666666667   |
| 2018  | 83.5                |
| 2019  | 81.5                |
| 2020  | 73.3                |
+-------+---------------------+--+
```

```
select sw.date_time, sw.data_value  as Sweden_val, tw.data_value as Taiwan_value
from sw_cc sw inner join tw_cc tw on tw.date_time =sw.date_time
where cast(substring(sw.date_time, 0,4) AS int) >2009 ;
```

```
+---------------+-------------+---------------+--+
| sw.date_time  | sweden_val  | taiwan_value  |
+---------------+-------------+---------------+--+
| 2016-01-31    | 96          | 80            |
| 2016-02-29    | 97          | 82            |
| 2016-03-31    | 99          | 81            |
| 2016-04-30    | 97          | 80            |
| 2016-05-31    | 97          | 79            |
| 2016-06-30    | 100         | 78            |
| 2016-07-31    | 97          | 80            |
| 2016-08-31    | 97          | 79            |
| 2016-09-30    | 102         | 78            |
| 2016-10-31    | 103         | 78            |
| 2016-11-30    | 103         | 77            |
| 2016-12-31    | 101         | 77            |
| 2017-01-31    | 103         | 74            |
| 2017-02-28    | 104         | 77            |
| 2017-03-31    | 103         | 78            |
| 2017-04-30    | 104         | 78            |
| 2017-05-31    | 108         | 78            |
| 2017-06-30    | 105         | 77            |
| 2017-07-31    | 103         | 78            |
| 2017-08-31    | 102         | 79            |
| 2017-09-30    | 102         | 82            |
| 2017-10-31    | 105         | 83            |
| 2017-11-30    | 107         | 86            |
| 2017-12-31    | 106         | 86            |
| 2018-01-31    | 106         | 87            |
| 2018-02-28    | 104         | 87            |
| 2018-03-31    | 102         | 87            |
| 2018-04-30    | 101         | 86            |
| 2018-05-31    | 101         | 85            |
| 2018-06-30    | 99          | 83            |
| 2018-07-31    | 99          | 82            |
| 2018-08-31    | 103         | 82            |
| 2018-09-30    | 104         | 83            |
| 2018-10-31    | 98          | 81            |
| 2018-11-30    | 96          | 80            |
| 2018-12-31    | 95          | 79            |
| 2019-01-31    | 92          | 83            |
| 2019-02-28    | 94          | 84            |
| 2019-03-31    | 95          | 84            |
| 2019-04-30    | 97          | 85            |
| 2019-05-31    | 94          | 79            |
| 2019-06-30    | 95          | 79            |
| 2019-07-31    | 97          | 81            |
| 2019-08-31    | 94          | 79            |
| 2019-09-30    | 90          | 80            |
| 2019-10-31    | 93          | 80            |
| 2019-11-30    | 92          | 80            |
| 2019-12-31    | 95          | 84            |
| 2020-01-31    | 92          | 85            |
| 2020-02-29    | 99          | 83            |
| 2020-03-31    | 89          | 78            |
| 2020-04-30    | 75          | 73            |
| 2020-05-31    | 78          | 64            |
| 2020-06-30    | 84          | 68            |
| 2020-07-31    | 84          | 69            |
| 2020-08-31    | 85          | 71            |
| 2020-09-30    | 88          | 71            |
| 2020-10-31    | 90          | 71            |
+---------------+-------------+---------------+--+
| 2013-03-31    | 487888      | 2045050       |
| 2013-06-30    | 489236      | 2066160       |
| 2013-09-30    | 494054      | 2075360       |
| 2013-12-31    | 497773      | 2108538       |
| 2014-03-31    | 500426      | 2116662       |
| 2014-06-30    | 506801      | 2143224       |
| 2014-09-30    | 506456      | 2168149       |
| 2014-12-31    | 512051      | 2174364       |
| 2015-03-31    | 519573      | 2190128       |
| 2015-06-30    | 522277      | 2223162       |
| 2015-09-30    | 528485      | 2201526       |
| 2015-12-31    | 533073      | 2233738       |
| 2016-03-31    | 536850      | 2250995       |
| 2016-06-30    | 534000      | 2268540       |
| 2016-09-30    | 539636      | 2273310       |
| 2016-12-31    | 540847      | 2289230       |
| 2017-03-31    | 549264      | 2304428       |
| 2017-06-30    | 548936      | 2315257       |
| 2017-09-30    | 553634      | 2342588       |
| 2017-12-31    | 555932      | 2364963       |
| 2018-03-31    | 561019      | 2368752       |
| 2018-06-30    | 563255      | 2380035       |
| 2018-09-30    | 560730      | 2375539       |
| 2018-12-31    | 564238      | 2393421       |
| 2019-03-31    | 563394      | 2414906       |
| 2019-06-30    | 567664      | 2421307       |
| 2019-09-30    | 570490      | 2430891       |
| 2019-12-31    | 576241      | 2453705       |
| 2020-03-31    | 559734      | 2390658       |
| 2020-06-30    | 516524      | 2323849       |
```