Project Set up:
Install python 3.7.6 (pyspark 2.4.5 have some issues with 3.8, so recommended to use 3.7.x)
https://www.python.org/downloads/
Install anaconda latest version (uses python 3.7.6)
https://www.anaconda.com/distribution/

Download code from git:
https://github.com/rajasekhar-learn/ExplorePython

Start HDFS :

rajasekharv@MyPC-RAJ ~/hadoop
$ ./sbin/start-dfs.cmd

Start mysql server:
rajasekharv@MyPC-RAJ /cygdrive/c/Program Files/mysql/bin
$ ./mysqld.exe --console
mysqld: Could not create or access the registry key needed for the MySQL application
to log to the Windows EventLog. Run the application with sufficient
privileges once to create the key, add the key manually, or turn off
logging for that application.
2020-04-10T17:46:17.058092Z 0 [Warning] TIMESTAMP with implicit DEFAULT value is deprecated. Please use --explicit_defaults_for_timestamp server option (see documentation for more details).
2020-04-10T17:46:17.175487Z 0 [Note] --secure-file-priv is set to NULL. Operations related to importing and exporting data are disabled
2020-04-10T17:46:17.276109Z 0 [ERROR] Cannot open Windows EventLog; check privileges, or start server with --log_syslog=0
2020-04-10T17:46:17.276241Z 0 [Note] C:\Program Files\mysql\bin\mysqld.exe (mysqld 5.7.29) starting as process 15472 ...
2020-04-10T17:46:18.082131Z 0 [Note] InnoDB: Mutexes and rw_locks use Windows interlocked functions
2020-04-10T17:46:18.082695Z 0 [Note] InnoDB: Uses event mutexes
2020-04-10T17:46:18.083357Z 0 [Note] InnoDB: Memory barrier is not used
2020-04-10T17:46:18.083922Z 0 [Note] InnoDB: Compressed tables use zlib 1.2.11
2020-04-10T17:46:18.107671Z 0 [Note] InnoDB: Number of pools: 1
2020-04-10T17:46:18.146091Z 0 [Note] InnoDB: Not using CPU crc32 instructions
2020-04-10T17:46:18.160187Z 0 [Note] InnoDB: Initializing buffer pool, total size = 128M, instances = 1, chunk size = 128M
2020-04-10T17:46:18.182057Z 0 [Note] InnoDB: Completed initialization of buffer pool
2020-04-10T17:46:18.788020Z 0 [Note] InnoDB: Highest supported file format is Barracuda.

2020-04-10T17:46:24.856924Z 0 [Note] InnoDB: Creating shared tablespace for temporary tables
2020-04-10T17:46:24.858005Z 0 [Note] InnoDB: Setting file '.\ibtmp1' size to 12 MB. Physically writing the file full; Please wait ...
2020-04-10T17:46:25.239827Z 0 [Note] InnoDB: File '.\ibtmp1' size is now 12 MB.
2020-04-10T17:46:25.350165Z 0 [Note] InnoDB: 96 redo rollback segment(s) found. 96 redo rollback segment(s) are active.
2020-04-10T17:46:25.350568Z 0 [Note] InnoDB: 32 non-redo rollback segment(s) are active.
2020-04-10T17:46:25.352827Z 0 [Note] InnoDB: Waiting for purge to start
2020-04-10T17:46:25.403439Z 0 [Note] InnoDB: page_cleaner: 1000ms intended loop took 7217ms. The settings might not be optimal. (flushed=0 and evicted=0, during the time.)
2020-04-10T17:46:25.413285Z 0 [Note] InnoDB: 5.7.29 started; log sequence number 6224550
2020-04-10T17:46:25.420845Z 0 [Note] InnoDB: Loading buffer pool(s) from C:\Program Files\mysql\data\ib_buffer_pool
2020-04-10T17:46:25.490095Z 0 [Note] Plugin 'FEDERATED' is disabled.
2020-04-10T17:46:26.645786Z 0 [Note] InnoDB: Buffer pool(s) load completed at 200410 23:16:26
2020-04-10T17:46:27.043685Z 0 [Note] Found ca.pem, server-cert.pem and server-key.pem in data directory. Trying to enable SSL support using them.
2020-04-10T17:46:27.044516Z 0 [Note] Skipping generation of SSL certificates as certificate files are present in data directory.
2020-04-10T17:46:27.214434Z 0 [Warning] CA certificate ca.pem is self signed.
2020-04-10T17:46:27.217049Z 0 [Note] Skipping generation of RSA key pair as key files are present in data directory.
2020-04-10T17:46:27.448762Z 0 [Note] Server hostname (bind-address): '*'; port: 3306
2020-04-10T17:46:27.453526Z 0 [Note] IPv6 is available.
2020-04-10T17:46:27.506573Z 0 [Note]   - '::' resolves to '::';
2020-04-10T17:46:27.507335Z 0 [Note] Server socket created on IP: '::'.
2020-04-10T17:46:31.121062Z 0 [Note] Event Scheduler: Loaded 0 events
2020-04-10T17:46:31.121781Z 0 [Note] C:\Program Files\mysql\bin\mysqld.exe: ready for connections.
Version: '5.7.29' socket: '' port: 3306 MySQL Community Server (GPL)
2020-04-10T17:47:31.874853Z 2 [Note] Access denied for user 'root'@'localhost' (using password: NO)

Connect mysql :

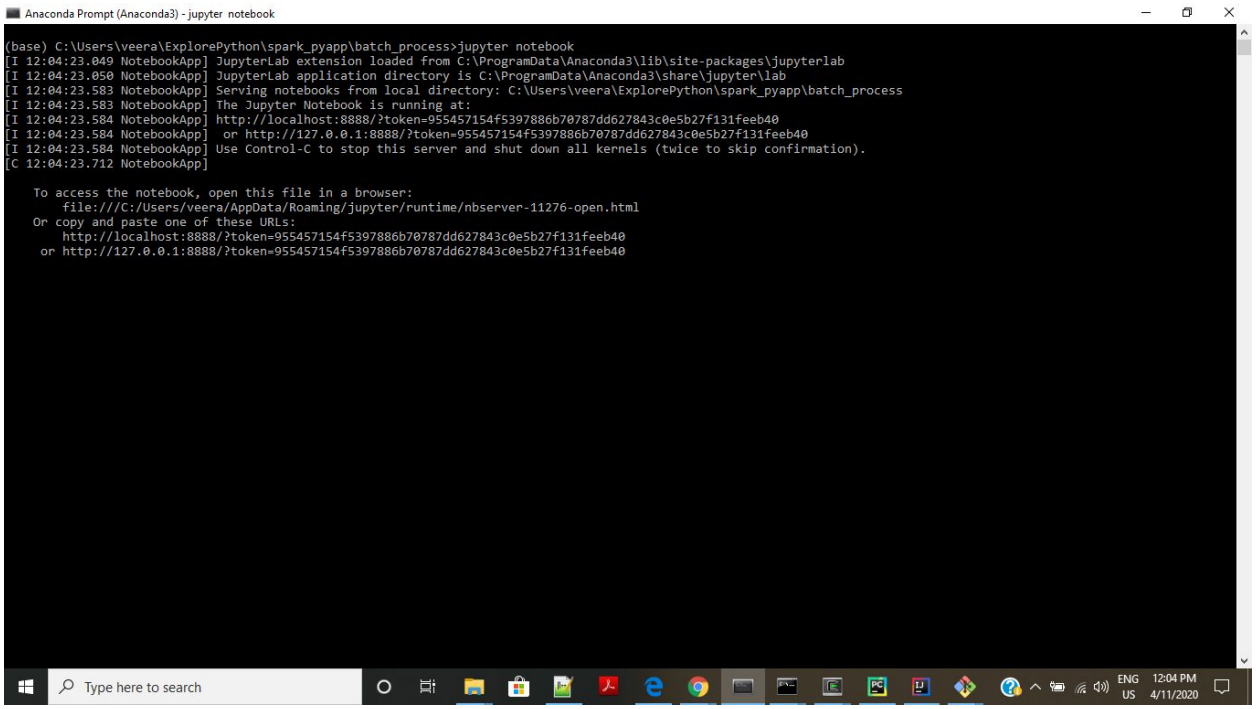
rajasekharv@MyPC-RAJ /cygdrive/c/Program Files/mysql/bin
$ ./mysql -u root -p
Enter password: ****
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 4

Server version: 5.7.29 MySQL Community Server (GPL)

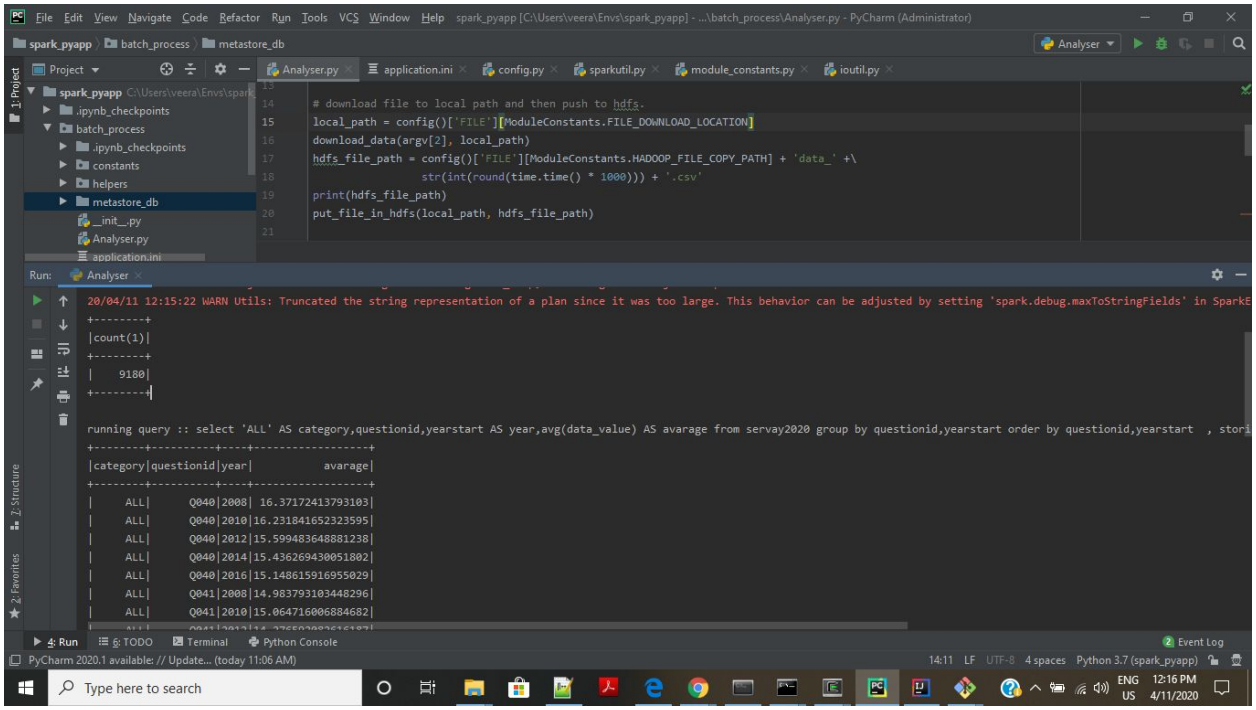
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

```
mysql> use hivedb;
Database changed
```

Open jupytor notebook using anaconda prompt:
Jyputor notebook



Intellij execution:



Intellij execution logs:

C:\Users\veera\Envs\spark_pyapp\Scripts\python.exe
C:/Users/veera/Envs/spark_pyapp/batch_process/Analyser.py 1
https://chronicdata.cdc.gov/views/735e-byxc/rows.csv?accessType=DOWNLOAD
hdfs://localhost:9000/
hdfs://localhost:9000/data_1586587303346.csv
Running shell command: hdfs dfs -put C:\\Users\\veera\\Desktop\\csv\\data.csv
hdfs://localhost:9000/data_1586587303346.csv
0 b'' b''
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/04/11 12:12:30 WARN ObjectStore: Version information not found in metastore.
hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0
20/04/11 12:12:30 WARN ObjectStore: Failed to get database default, returning
NoSuchObjectException
20/04/11 12:12:33 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
20/04/11 12:12:36 WARN ObjectStore: Failed to get database global_temp, returning
NoSuchObjectException
+--------+

|count(1)|
+--------+
| 9180|
+--------+

running query :: select 'ALL' AS category,questionid,yearstart AS year,avg(data_value) AS avarage from servay2020 group by questionid,yearstart order by questionid,yearstart , storing results in :: servay_analysis_all_results

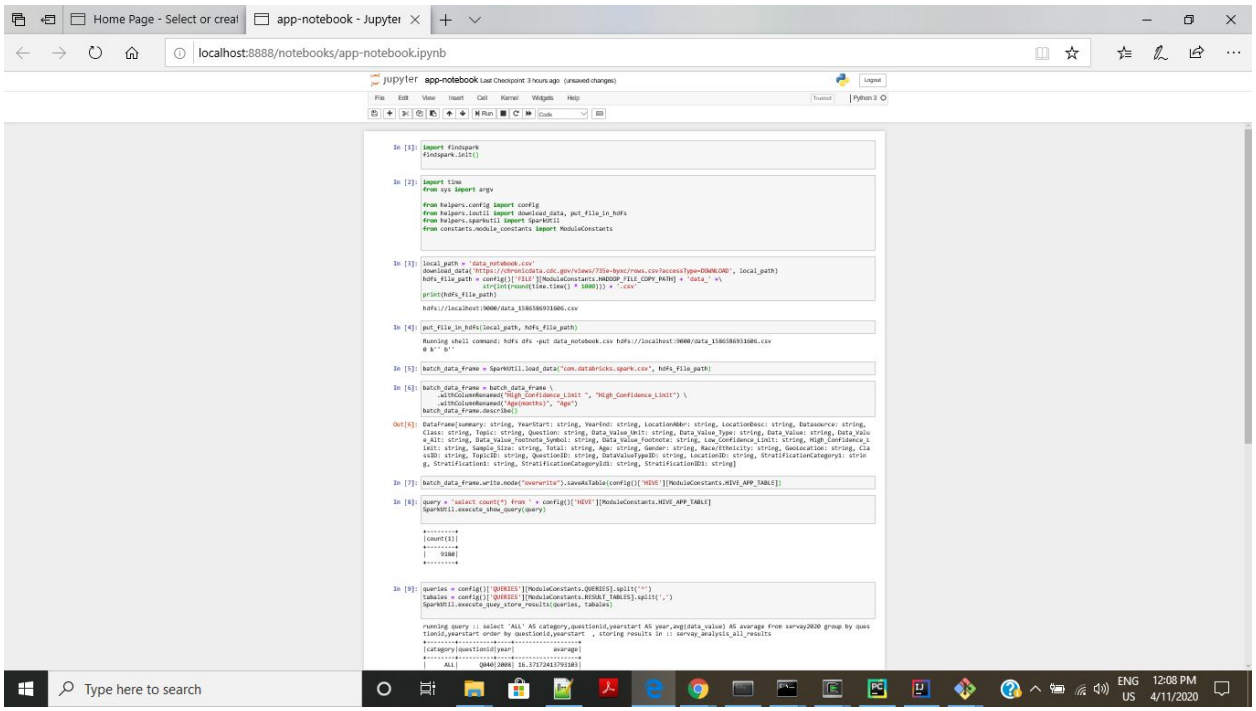
+--------+----------+----+-----------------+
|category|questionid|year| avarage|
+--------+----------+----+-----------------+
ALL	Q040	2008	16.37172413793103
ALL	Q040	2010	16.231841652323595
ALL	Q040	2012	15.599483648881238
ALL	Q040	2014	15.436269430051802
ALL	Q040	2016	15.148615916955029
ALL	Q041	2008	14.983793103448296
ALL	Q041	2010	15.064716006884682
ALL	Q041	2012	14.276592082616187
ALL	Q041	2014	13.943350604490504
ALL	Q041	2016	13.53235294117647
ALL	Q060	2008	13.218181818181813
ALL	Q060	2010	13.046984126984121
ALL	Q060	2012	11.963057324840753
ALL	Q060	2014	11.450638977635784
ALL	Q060	2016	11.577635782747599
+--------+----------+----+-----------------+

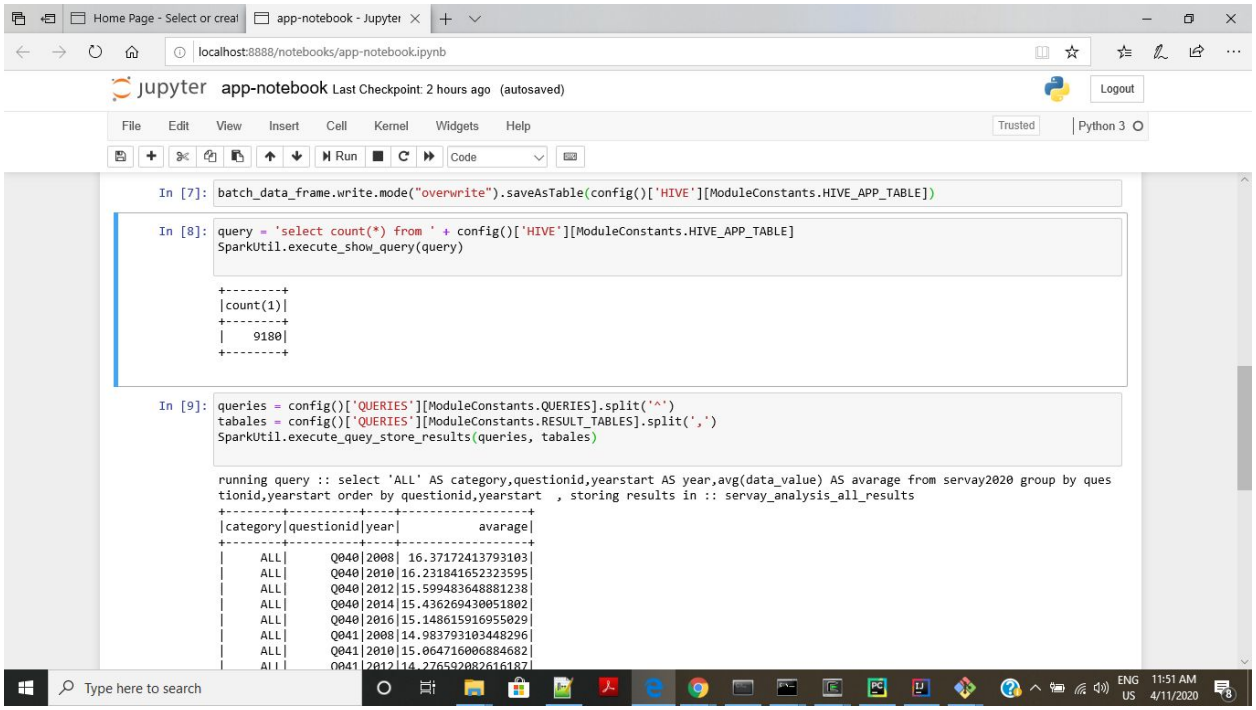
running query :: select 'FEMALE' AS category,questionid,yearstart AS year,avg(data_value) AS avarage from servay2020 where gender ='Female' group by questionid,yearstart order by questionid,yearstart , storing results in :: servay_analysis_female_results
+--------+----------+----+-----------------+
|category|questionid|year| avarage|
+--------+----------+----+-----------------+
FEMALE	Q040	2008	16.2962962962963
FEMALE	Q040	2010	16.23518518518518
FEMALE	Q040	2012	15.648148148148145
FEMALE	Q040	2014	15.411111111111113
FEMALE	Q040	2016	15.39259259259259
FEMALE	Q041	2008	13.91666666666667
FEMALE	Q041	2010	14.161111111111111
FEMALE	Q041	2012	13.594444444444445

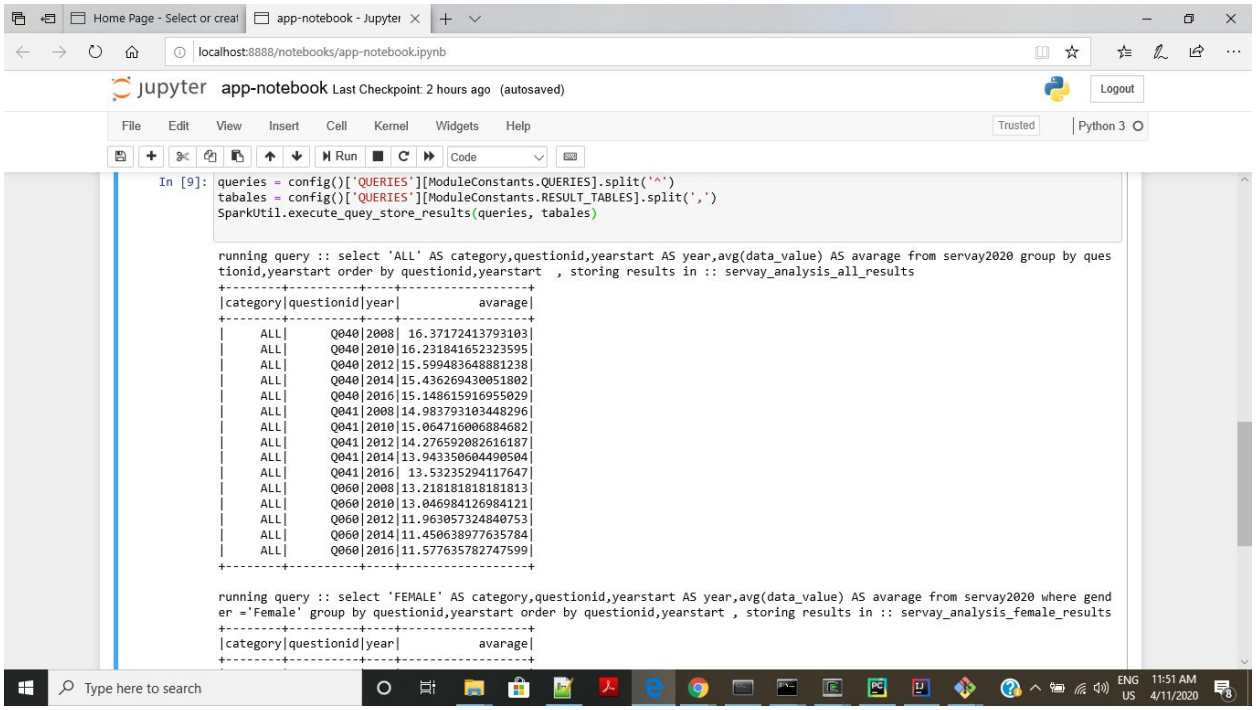
```
|  FEMALE|      Q041|2014|13.198148148148144|
|  FEMALE|      Q041|2016|12.957407407407413|
|  FEMALE|      Q060|2008| 12.72962962962963|
|  FEMALE|      Q060|2010| 12.38148148148148|
|  FEMALE|      Q060|2012|11.264814814814816|
|  FEMALE|      Q060|2014|10.725925925925926|
|  FEMALE|      Q060|2016|10.996296296296292|
+--------+----------+----+------------------+
```

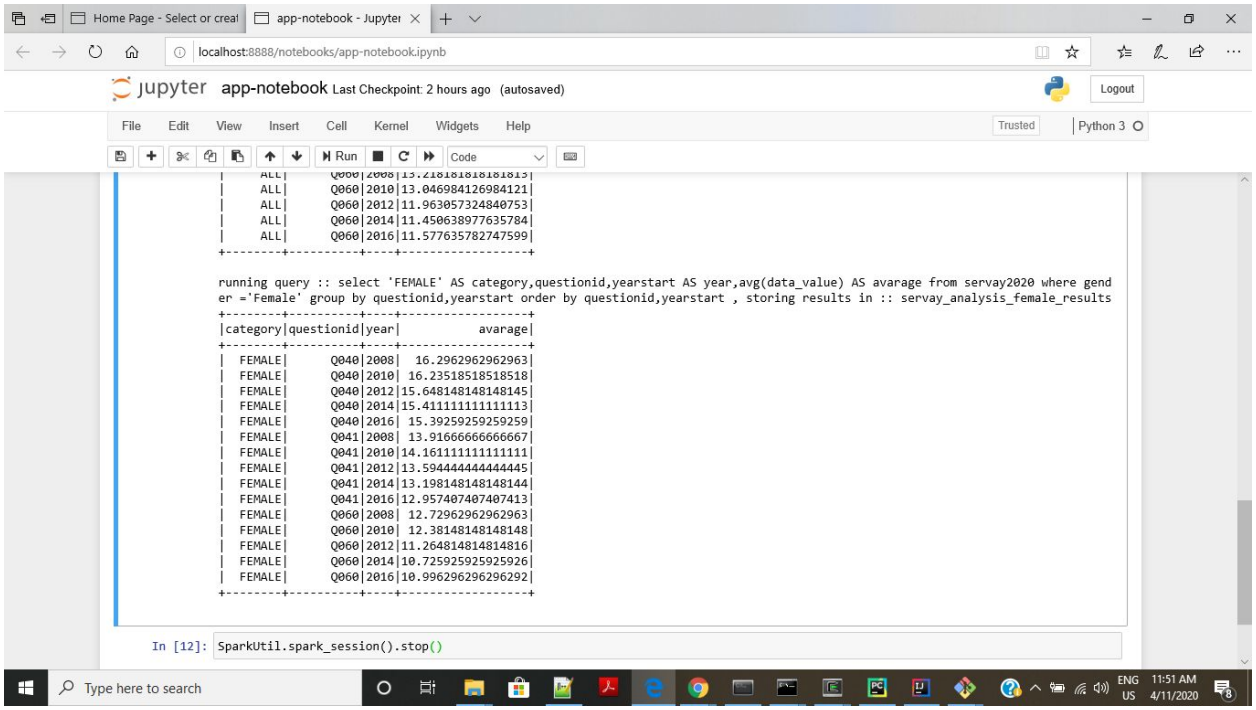
Process finished with exit code 0

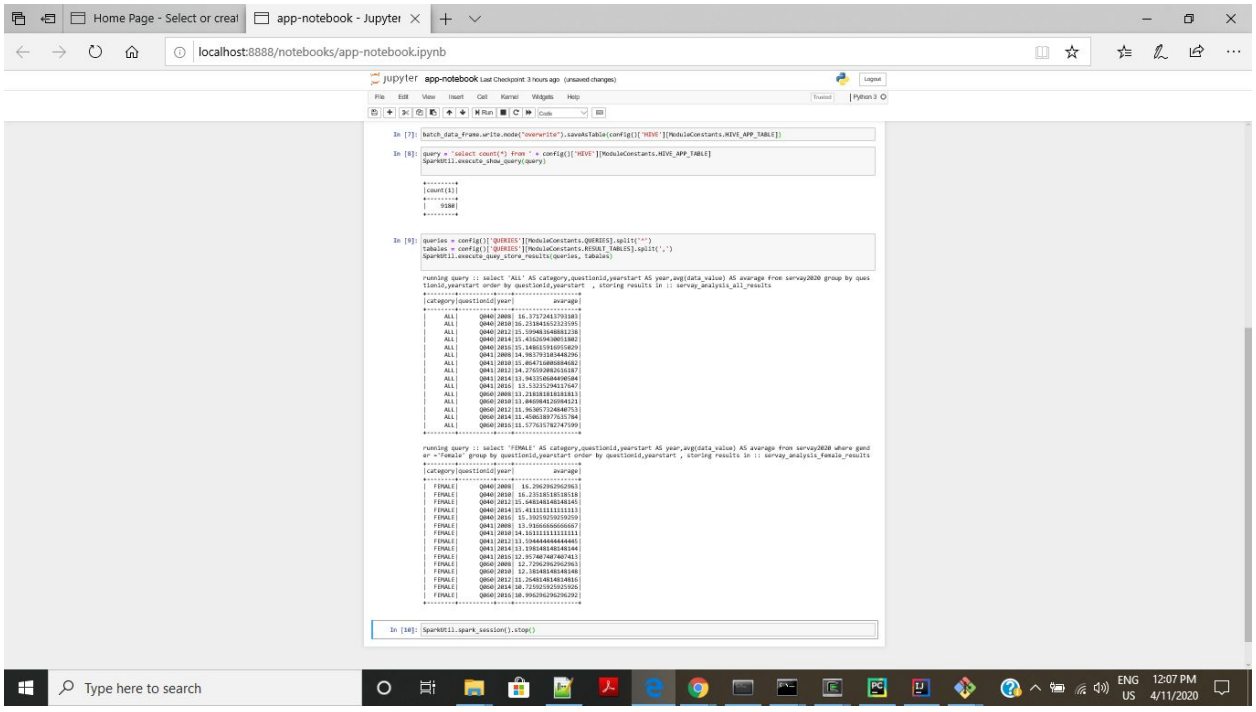
Jupytor notebook execution screens:



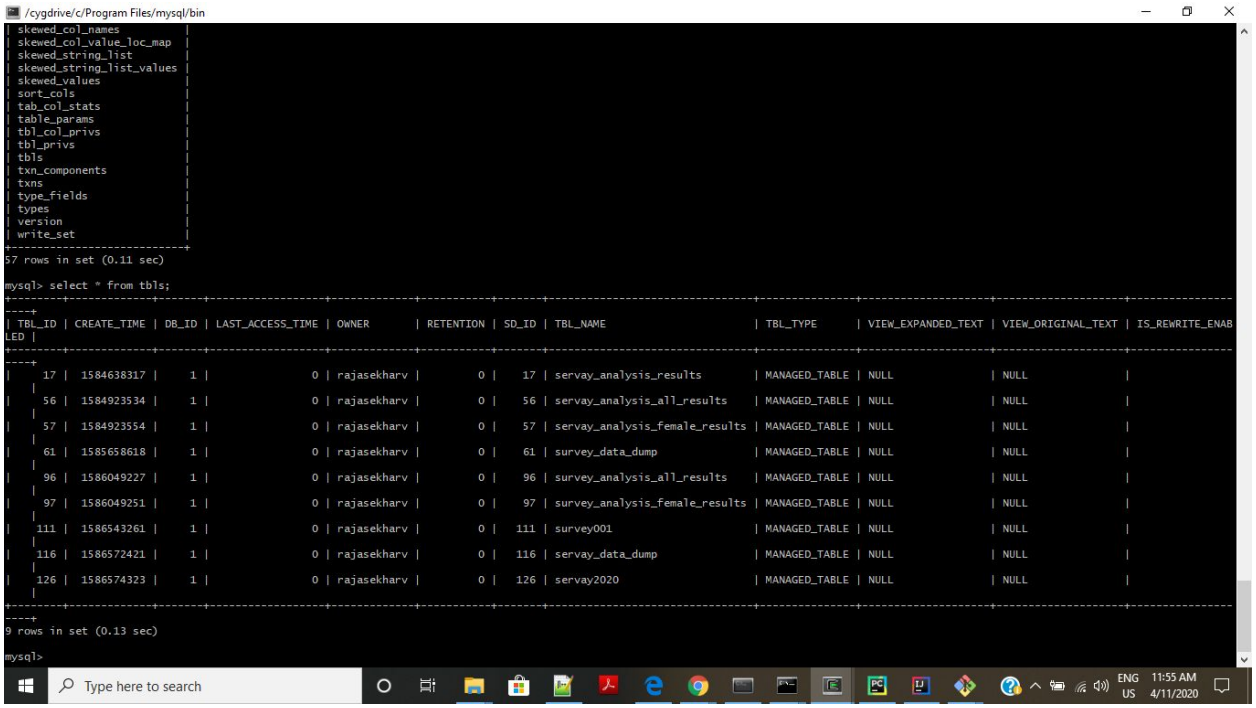

Home Page - Select or creat
app-notebook - Jupyter
localhost:8888/notebooks/app-notebook.ipynb
jupyter app-notebook Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [7]: batch_data_frame.write.mode("overwrite").saveAsTable(config()['HIVE'][ModuleConstants.HIVE_APP_TABLE])
In [8]: query = 'select count(*) from ' + config()['HIVE'][ModuleConstants.HIVE_APP_TABLE]
SparkUtil.execute_show_query(query)
+--------+
|count(1)|
+--------+
| 9180|
+--------+
In [9]: queries = config()['QUERIES'][ModuleConstants.QUERIES].split('^')
tabales = config()['QUERIES'][ModuleConstants.RESULT_TABLES].split(',')
SparkUtil.execute_quey_store_results(queries, tabales)
running query :: select 'ALL' AS category,questionid,yearstart AS year,avg(data_value) AS avarage from servay2020 group by ques
tionid,yearstart order by questionid,yearstart , storing results in :: servay_analysis_all_results
+--------+----------+----+------------------+
|category|questionid|year| avarage|
+--------+----------+----+------------------+
ALL	Q040	2008	16.37172413793103
ALL	Q040	2010	16.231841652323595
ALL	Q040	2012	15.599483648881238
ALL	Q040	2014	15.436269430051802
ALL	Q040	2016	15.148615916955029
ALL	Q041	2008	14.983793103448296
ALL	Q041	2010	15.064716006884682
Type here to search
ENG US
11:51 AM
4/11/2020


Home Page - Select or creat
app-notebook - Jupyter
localhost:8888/notebooks/app-notebook.ipynb
jupyter app-notebook Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [9]: queries = config()['QUERIES'][ModuleConstants.QUERIES].split('^')
tabales = config()['QUERIES'][ModuleConstants.RESULT_TABLES].split(',')
SparkUtil.execute_quey_store_results(queries, tabales)
running query :: select 'ALL' AS category,questionid,yearstart AS year,avg(data_value) AS avarage from servay2020 group by ques
tionid,yearstart order by questionid,yearstart , storing results in :: servay_analysis_all_results
+--------+----------+----+------------------+
|category|questionid|year| avarage|
+--------+----------+----+------------------+
ALL	Q040	2008	16.37172413793103
ALL	Q040	2010	16.231841652323595
ALL	Q040	2012	15.599483648881238
ALL	Q040	2014	15.436269430051802
ALL	Q040	2016	15.148615916955029
ALL	Q041	2008	14.983793103448296
ALL	Q041	2010	15.064716006884682
ALL	Q041	2012	14.276592082616187
ALL	Q041	2014	13.943350604490504
ALL	Q041	2016	13.53235294117647
ALL	Q060	2008	13.218181818181813
ALL	Q060	2010	13.046984126984121
ALL	Q060	2012	11.963057324840753
ALL	Q060	2014	11.450638977635784
ALL	Q060	2016	11.577635782747599
+--------+----------+----+------------------+
running query :: select 'FEMALE' AS category,questionid,yearstart AS year,avg(data_value) AS avarage from servay2020 where gend
er ='Female' group by questionid,yearstart order by questionid,yearstart , storing results in :: servay_analysis_female_results
+--------+----------+----+------------------+
|category|questionid|year| avarage|
+--------+----------+----+------------------+
Type here to search
ENG US
11:51 AM
4/11/2020

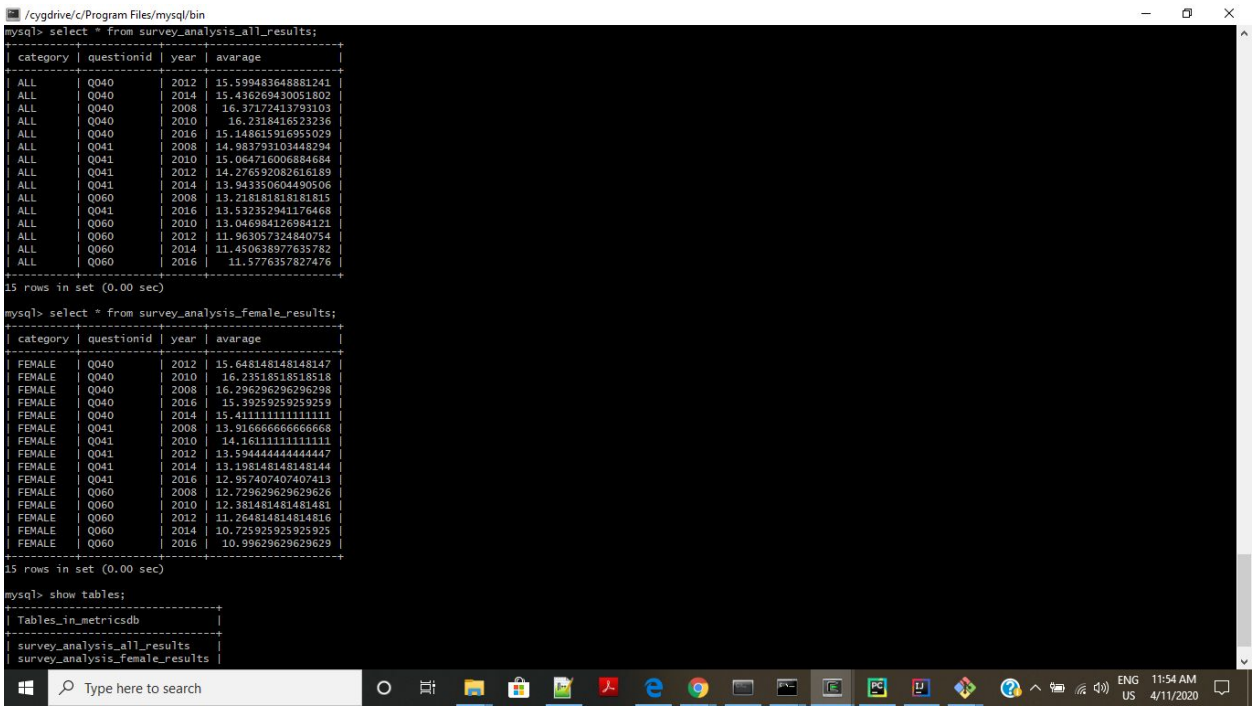




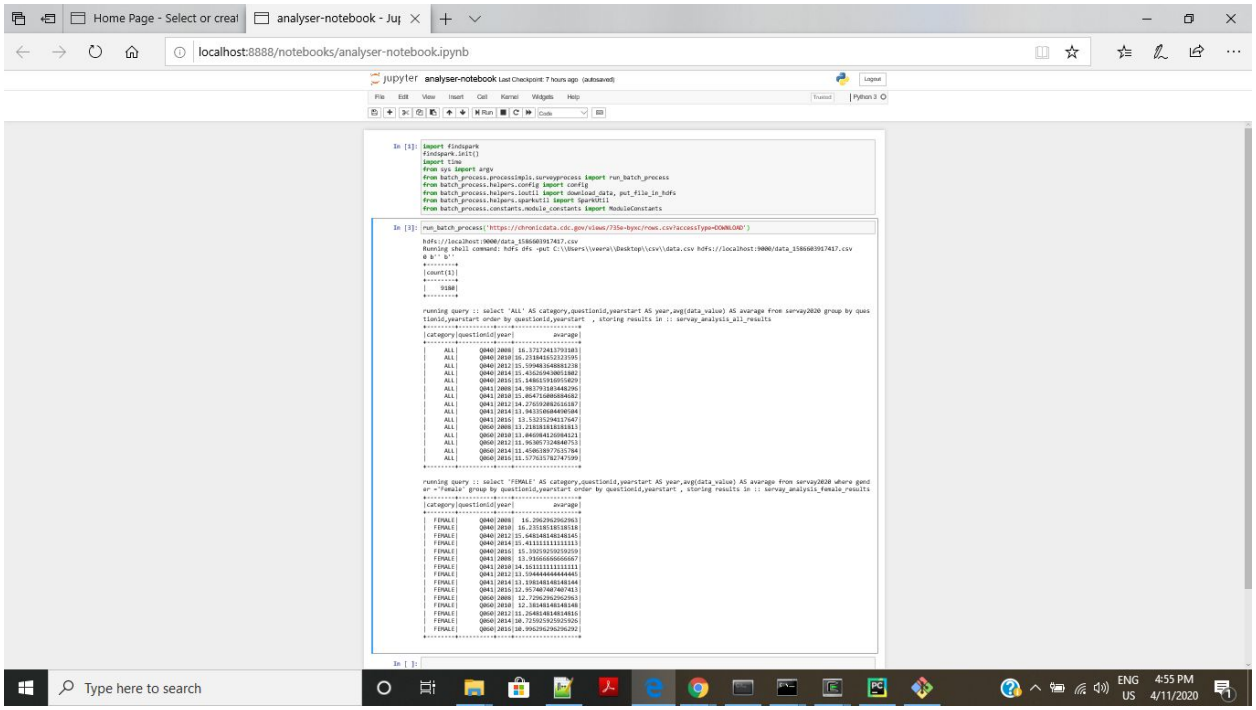
Hive metastore db tables:



Mysql results tables:



After code refactoring final note book: analyser-notebook


Home Page - Select or creat
analyser-notebook - Jup
localhost:8888/notebooks/analyser-notebook.ipynb
jupyter analyser-notebook Last Checkpoint: 7 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [1]:
import findspark
findspark.init()
import time
from sys import argv
from batch_process.processimpls.surveyprocess import run_batch_process
from batch_process.helpers.config import config
from batch_process.helpers.loutil import download_data, put_file_in_hdfs
from batch_process.helpers.sparkutil import SparkUtil
from batch_process.constants.module_constants import ModuleConstants
In [3]:
run_batch_process('https://chronicdata.cdc.gov/views/735e-byxc/rows.csv?accessType=DOWNLOAD')
hdfs://localhost:9000/data_1586603917417.csv
Running shell command: hdfs dfs -put C:\\Users\\veera\\Desktop\\csv\\data.csv hdfs://localhost:9000/data_1586603917417.csv
0 b' b'
+--------+
|count(1)|
+--------+
| 9180|
+--------+
running query :: select 'ALL' AS category,questionid,yearstart AS year,avg(data_value) AS average from survey2020 group by questionid,yearstart order by questionid,yearstart , storing results in :: survey_analysis_all_results
+--------+----------+----+------------------+
|category|questionid|year| average|
+--------+----------+----+------------------+
ALL	QN40	2008	16.37172413793103
ALL	QN40	2010	16.233461523233595
ALL	QN40	2012	15.599481488881238
ALL	QN40	2014	15.416269430051802
ALL	QN40	2016	15.148615910955829
ALL	QN41	2008	14.983791034448296
ALL	QN41	2010	15.054716086884682
ALL	QN41	2012	14.276392892616187
ALL	QN41	2014	13.943350844070584
ALL	QN41	2016	13.53235294117647
ALL	QN50	2008	13.218181818181813
ALL	QN50	2010	13.046984128984121
ALL	QN50	2012	11.963057324044753
ALL	QN50	2014	11.459639775157584
ALL	QN50	2016	11.577635782747599
+--------+----------+----+------------------+
running query :: select 'FEMALE' AS category,questionid,yearstart AS year,avg(data_value) AS average from survey2020 where gender ="female" group by questionid,yearstart order by questionid,yearstart , storing results in :: survey_analysis_female_results
+--------+----------+----+------------------+
|category|questionid|year| average|
+--------+----------+----+------------------+
FEMALE	QN40	2008	16.296296296296293
FEMALE	QN40	2010	16.235185185185185
FEMALE	QN40	2012	15.648148148148145
FEMALE	QN40	2014	15.411111111111113
FEMALE	QN40	2016	15.392592592592593
FEMALE	QN41	2008	13.916666666666667
FEMALE	QN41	2010	14.161111111111111
FEMALE	QN41	2012	13.594444444444445
FEMALE	QN41	2014	13.198148148148144
FEMALE	QN41	2016	12.957407407407415
FEMALE	QN50	2008	12.729629629629633
FEMALE	QN50	2010	12.384814814814816
FEMALE	QN50	2012	11.264814814814816
FEMALE	QN50	2014	10.725925925925926
FEMALE	QN50	2016	10.995296296296292
+--------+----------+----+------------------+
In [ ]:
Type here to search
ENG
US
4:55 PM
4/11/2020