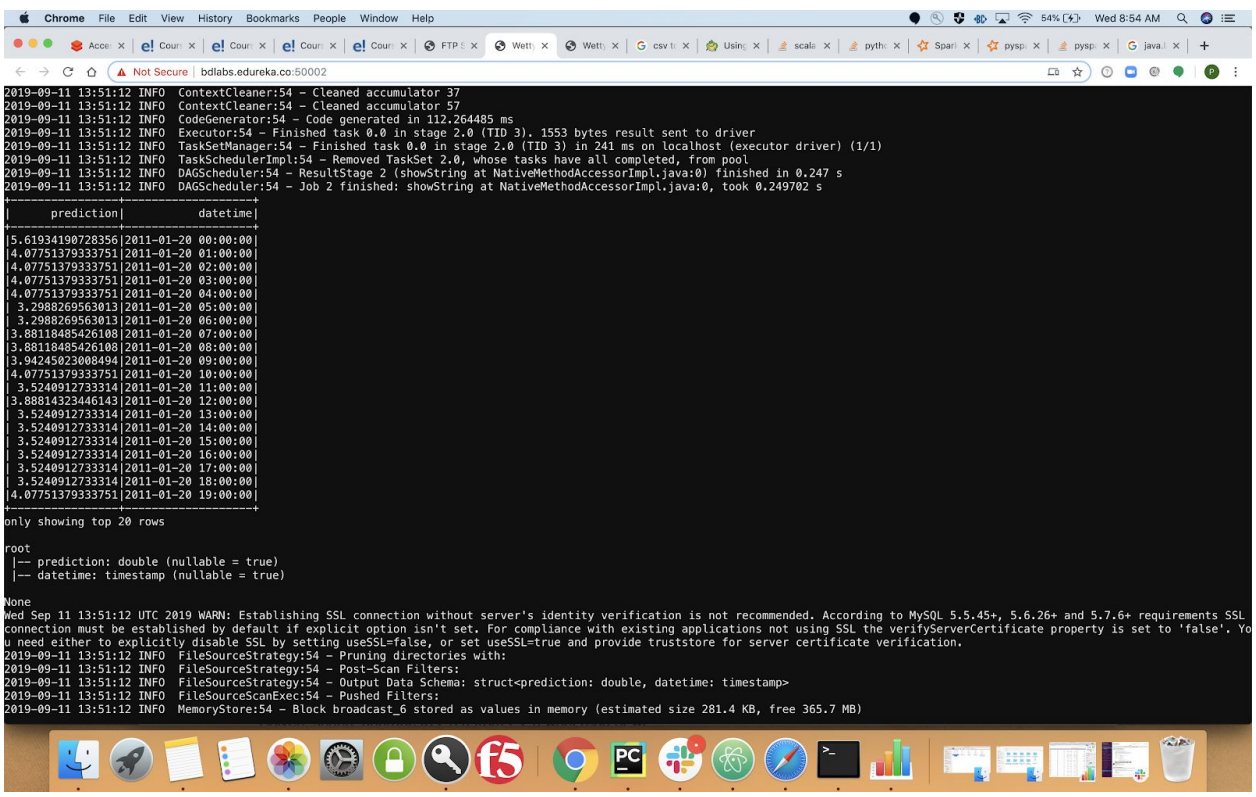
Bicycle Sharing Demand

Bike_sharing_project_file.ipynb
Initial Data Exploration and Transformation is done in data bricks platform and have attached the codebase as Ipython Notebook which explicitly displays each intermediate result and model performances.

bike_sharing_model_generation.py
This file is used to give a trained model taking training files as input by cleaning them and using one of the algorithms which gave best results GBTRegressor which gave least root mean square error. This pyspark code gives trained model files as **bike_sharing_gbt_file.model**

bike_sharing_prediction.py
This file is uses the model generated out of previous file and predicts the bike sharing demand on the test files given. Then finally outputs the predictions as csv file with name **predictions.csv** and also saves them in RDBMS as shown below.

Predictions are printed below which are in data frame

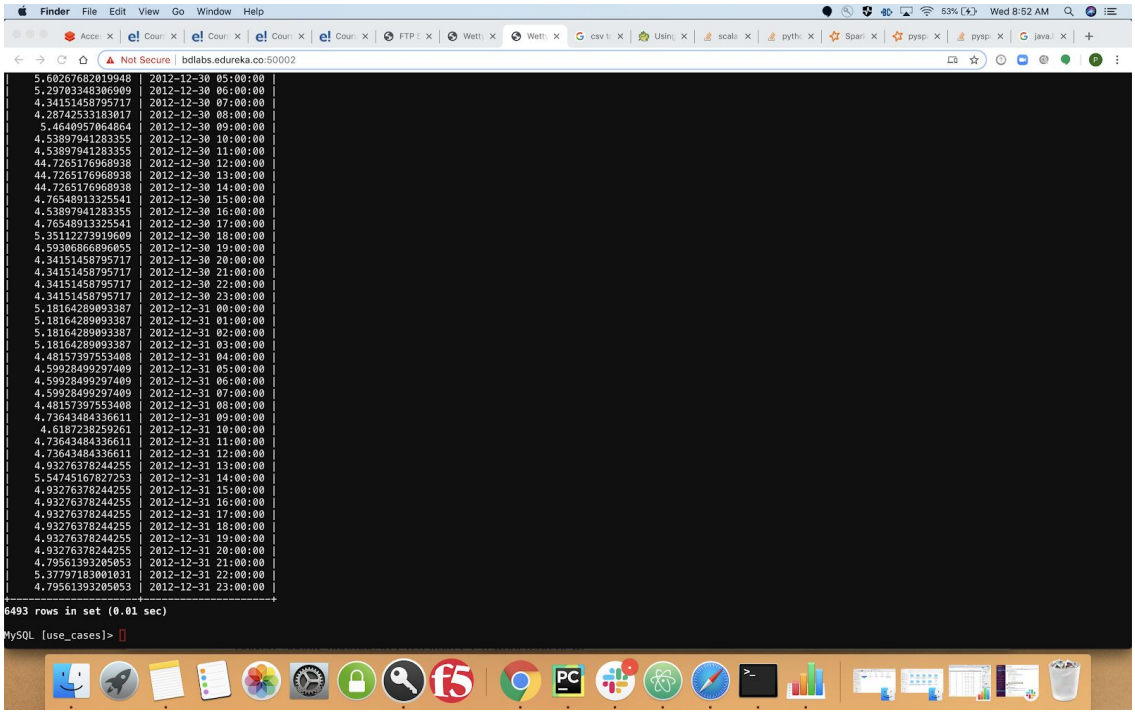


Predictions are saved in RDBMs tables to run this pyspark code required driver packages are also passed as parameters.

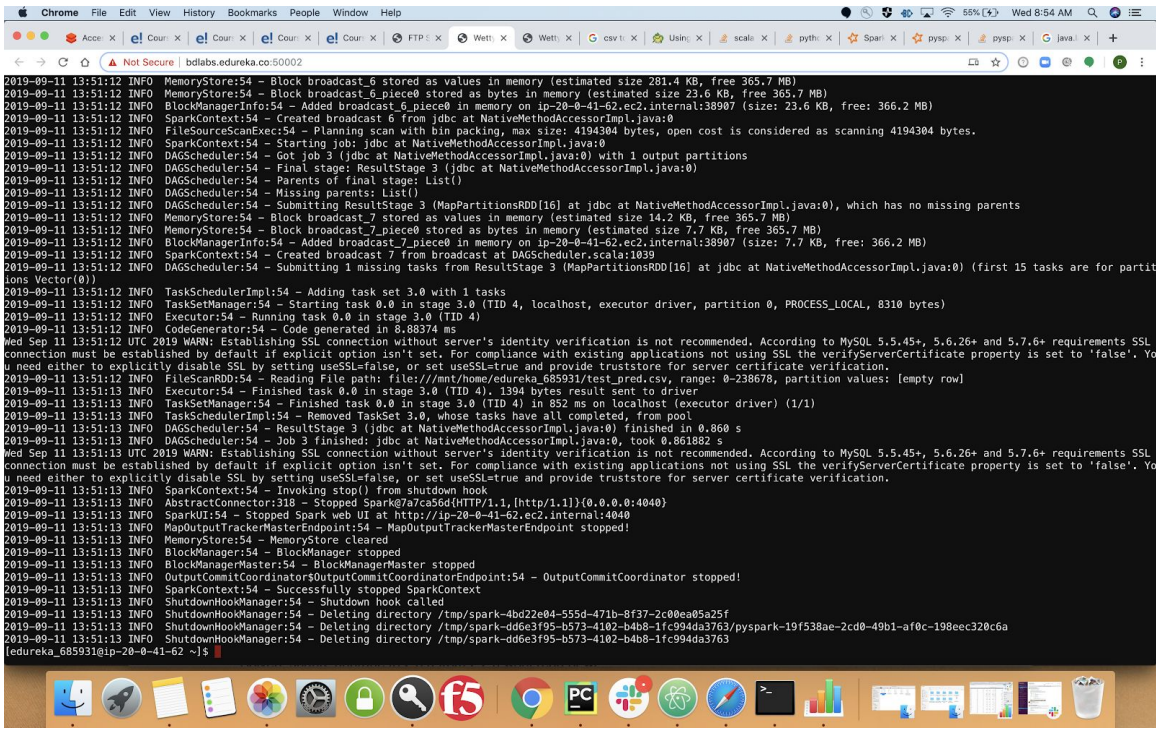
```
2019-09-11 13:49:43 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-4f1fee07-ab08-4b70-be84-71cf5b4bd5b0
2019-09-11 13:49:43 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-0d3e36c8-6eea-4789-b67f-5f1c3ea2373a
[edureka_685931@ip-20-0-41-62 ~]$ vim df_db_check.py
[edureka_685931@ip-20-0-41-62 ~]$ spark-submit --packages mysql:mysql-connector-java:5.1.39 df_db_check.py
Ivy Default Cache set to: /mnt/home/edureka_685931/.ivy2/cache
The jars for the packages stored in: /mnt/home/edureka_685931/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/python2.7/site-packages/pyspark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
mysql#mysql-connector-java added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-cebea908-9c03-44d1-907a-0f600e9a6014;1.0
	confs: [default]
	found mysql#mysql-connector-java;5.1.39 in central
:: resolution report :: resolve 275ms :: artifacts dl 13ms
	:: modules in use:
	mysql#mysql-connector-java;5.1.39 from central in [default]
	---------------------------------------------------------------------
	|                  |            modules            ||   artifacts   |
	|       conf       | number| search|dwnlded|evicted|| number|dwnlded|
	---------------------------------------------------------------------
	|      default     |   1   |   0   |   0   |   0   ||   1   |   0   |
	---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-cebea908-9c03-44d1-907a-0f600e9a6014
	confs: [default]
	0 artifacts copied, 1 already retrieved (0kB/12ms)
2019-09-11 13:51:01 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2019-09-11 13:51:04 INFO  SparkContext:54 - Running Spark version 2.3.1
2019-09-11 13:51:04 INFO  SparkContext:54 - Submitted application: df_db_check.py
2019-09-11 13:51:04 INFO  SecurityManager:54 - Changing view acls to: edureka_685931
2019-09-11 13:51:04 INFO  SecurityManager:54 - Changing modify acls to: edureka_685931
2019-09-11 13:51:04 INFO  SecurityManager:54 - Changing view acls groups to:
2019-09-11 13:51:04 INFO  SecurityManager:54 - Changing modify acls groups to:
2019-09-11 13:51:04 INFO  SecurityManager:54 - SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(edureka_685931); groups with view permissions
: Set(); users  with modify permissions: Set(edureka_685931); groups with modify permissions: Set()
2019-09-11 13:51:04 INFO  Utils:54 - Successfully started service 'sparkDriver' on port 35782.
2019-09-11 13:51:04 INFO  SparkEnv:54 - Registering MapOutputTracker
2019-09-11 13:51:04 INFO  SparkEnv:54 - Registering BlockManagerMaster
2019-09-11 13:51:04 INFO  BlockManagerMasterEndpoint:54 - Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
2019-09-11 13:51:04 INFO  BlockManagerMasterEndpoint:54 - BlockManagerMasterEndpoint up
2019-09-11 13:51:04 INFO  DiskBlockManager:54 - Created local directory at /tmp/blockmgr-6c759e49-30c5-4d95-acbd-ca986b870d2b
2019-09-11 13:51:04 INFO  MemoryStore:54 - MemoryStore started with capacity 366.3 MB
2019-09-11 13:51:04 INFO  SparkEnv:54 - Registering OutputCommitCoordinator
2019-09-11 13:51:04 INFO  log:192 - Logging initialized @4741ms
2019-09-11 13:51:04 INFO  Server:346 - jetty-9.3.z-SNAPSHOT
2019-09-11 13:51:04 INFO  Server:414 - Started @4829ms
2019-09-11 13:51:04 INFO  AbstractConnector:278 - Started ServerConnector@7a7ca56d{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2019-09-11 13:51:04 INFO  Utils:54 - Successfully started service 'SparkUI' on port 4040.
2019-09-11 13:51:04 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@54641a14{/jobs,null,AVAILABLE,@Spark}
2019-09-11 13:51:04 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@4c00f3d5{/jobs/json,null,AVAILABLE,@Spark}
```

The predictions along with the datatime are pushed to RDBMS from the dataframe

```
| bike_sharing_predictions |
| bike_sharing_preditions  |
| departments              |
| financial_regulation     |
| order_products__prior    |
| order_products__train    |
| orders                   |
| paysim                   |
| products                 |
| spam_message             |
+--------------------------+
12 rows in set (0.00 sec)

MySQL [use_cases]> select * from bike_sharing_predictions
    -> ;
Empty set (0.00 sec)

MySQL [use_cases]> select * from bike_sharing_predictions;
Empty set (0.00 sec)

MySQL [use_cases]> select * from bike_sharing_predictions;
Empty set (0.00 sec)

MySQL [use_cases]> select * from bike_sharing_predictions;
+--------------------+---------------------+
| prediction         | datetime            |
+--------------------+---------------------+
|   5.61934190728356 | 2011-01-20 00:00:00 |
|   4.07751379333751 | 2011-01-20 01:00:00 |
|   4.07751379333751 | 2011-01-20 02:00:00 |
|   4.07751379333751 | 2011-01-20 03:00:00 |
|   4.07751379333751 | 2011-01-20 04:00:00 |
|    3.2988269563013 | 2011-01-20 05:00:00 |
|    3.2988269563013 | 2011-01-20 06:00:00 |
|   3.88118485426108 | 2011-01-20 07:00:00 |
|   3.88118485426108 | 2011-01-20 08:00:00 |
|   3.94245023008494 | 2011-01-20 09:00:00 |
|   4.07751379333751 | 2011-01-20 10:00:00 |
|    3.5240912733314 | 2011-01-20 11:00:00 |
|   3.88814323446143 | 2011-01-20 12:00:00 |
|    3.5240912733314 | 2011-01-20 13:00:00 |
|    3.5240912733314 | 2011-01-20 14:00:00 |
|    3.5240912733314 | 2011-01-20 15:00:00 |
|    3.5240912733314 | 2011-01-20 16:00:00 |
|    3.5240912733314 | 2011-01-20 17:00:00 |
|    3.5240912733314 | 2011-01-20 18:00:00 |
|   4.07751379333751 | 2011-01-20 19:00:00 |
```



```
|    5.60267682019948 | 2012-12-30 05:00:00 |
|    5.29703348306909 | 2012-12-30 06:00:00 |
|    4.34151458795717 | 2012-12-30 07:00:00 |
|    4.28742533183017 | 2012-12-30 08:00:00 |
|     5.4640957064864 | 2012-12-30 09:00:00 |
|    4.53897941283355 | 2012-12-30 10:00:00 |
|    4.53897941283355 | 2012-12-30 11:00:00 |
|    44.7265176968938 | 2012-12-30 12:00:00 |
|    44.7265176968938 | 2012-12-30 13:00:00 |
|    44.7265176968938 | 2012-12-30 14:00:00 |
|    4.76548913325541 | 2012-12-30 15:00:00 |
|    4.53897941283355 | 2012-12-30 16:00:00 |
|    4.76548913325541 | 2012-12-30 17:00:00 |
|    5.35112273919609 | 2012-12-30 18:00:00 |
|    4.59306866896055 | 2012-12-30 19:00:00 |
|    4.34151458795717 | 2012-12-30 20:00:00 |
|    4.34151458795717 | 2012-12-30 21:00:00 |
|    4.34151458795717 | 2012-12-30 22:00:00 |
|    4.34151458795717 | 2012-12-30 23:00:00 |
|    5.18164289093387 | 2012-12-31 00:00:00 |
|    5.18164289093387 | 2012-12-31 01:00:00 |
|    5.18164289093387 | 2012-12-31 02:00:00 |
|    5.18164289093387 | 2012-12-31 03:00:00 |
|    4.48157397553408 | 2012-12-31 04:00:00 |
|    4.59928499297409 | 2012-12-31 05:00:00 |
|    4.59928499297409 | 2012-12-31 06:00:00 |
|    4.59928499297409 | 2012-12-31 07:00:00 |
|    4.48157397553408 | 2012-12-31 08:00:00 |
|    4.73643484336611 | 2012-12-31 09:00:00 |
|     4.6187238259261 | 2012-12-31 10:00:00 |
|    4.73643484336611 | 2012-12-31 11:00:00 |
|    4.73643484336611 | 2012-12-31 12:00:00 |
|    4.93276378244255 | 2012-12-31 13:00:00 |
|    5.54745167827253 | 2012-12-31 14:00:00 |
|    4.93276378244255 | 2012-12-31 15:00:00 |
|    4.93276378244255 | 2012-12-31 16:00:00 |
|    4.93276378244255 | 2012-12-31 17:00:00 |
|    4.93276378244255 | 2012-12-31 18:00:00 |
|    4.93276378244255 | 2012-12-31 19:00:00 |
|    4.93276378244255 | 2012-12-31 20:00:00 |
|    4.79561393205053 | 2012-12-31 21:00:00 |
|    5.37797183001031 | 2012-12-31 22:00:00 |
|    4.79561393205053 | 2012-12-31 23:00:00 |
+---------------------+---------------------+
6493 rows in set (0.01 sec)

MySQL [use_cases]>
```



```
2019-09-11 13:51:12 INFO  MemoryStore:54 - Block broadcast_6 stored as values in memory (estimated size 281.4 KB, free 365.7 MB)
2019-09-11 13:51:12 INFO  MemoryStore:54 - Block broadcast_6_piece0 stored as bytes in memory (estimated size 23.6 KB, free 365.7 MB)
2019-09-11 13:51:12 INFO  BlockManagerInfo:54 - Added broadcast_6_piece0 in memory on ip-20-0-41-62.ec2.internal:38907 (size: 23.6 KB, free: 366.2 MB)
2019-09-11 13:51:12 INFO  SparkContext:54 - Created broadcast 6 from jdbc at NativeMethodAccessorImpl.java:0
2019-09-11 13:51:12 INFO  FileSourceScanExec:54 - Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.
2019-09-11 13:51:12 INFO  SparkContext:54 - Starting job: jdbc at NativeMethodAccessorImpl.java:0
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Got job 3 (jdbc at NativeMethodAccessorImpl.java:0) with 1 output partitions
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Final stage: ResultStage 3 (jdbc at NativeMethodAccessorImpl.java:0)
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Parents of final stage: List()
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Missing parents: List()
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Submitting ResultStage 3 (MapPartitionsRDD[16] at jdbc at NativeMethodAccessorImpl.java:0), which has no missing parents
2019-09-11 13:51:12 INFO  MemoryStore:54 - Block broadcast_7 stored as values in memory (estimated size 14.2 KB, free 365.7 MB)
2019-09-11 13:51:12 INFO  MemoryStore:54 - Block broadcast_7_piece0 stored as bytes in memory (estimated size 7.7 KB, free 365.7 MB)
2019-09-11 13:51:12 INFO  BlockManagerInfo:54 - Added broadcast_7_piece0 in memory on ip-20-0-41-62.ec2.internal:38907 (size: 7.7 KB, free: 366.2 MB)
2019-09-11 13:51:12 INFO  SparkContext:54 - Created broadcast 7 from broadcast at DAGScheduler.scala:1039
2019-09-11 13:51:12 INFO  DAGScheduler:54 - Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[16] at jdbc at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partit
ions Vector(0))
2019-09-11 13:51:12 INFO  TaskSchedulerImpl:54 - Adding task set 3.0 with 1 tasks
2019-09-11 13:51:12 INFO  TaskSetManager:54 - Starting task 0.0 in stage 3.0 (TID 4, localhost, executor driver, partition 0, PROCESS_LOCAL, 8310 bytes)
2019-09-11 13:51:12 INFO  Executor:54 - Running task 0.0 in stage 3.0 (TID 4)
2019-09-11 13:51:12 INFO  CodeGenerator:54 - Code generated in 8.88374 ms
Wed Sep 11 13:51:12 UTC 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL
connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. Yo
u need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2019-09-11 13:51:12 INFO  FileScanRDD:54 - Reading File path: file:///mnt/home/edureka_685931/test_pred.csv, range: 0-238678, partition values: [empty row]
2019-09-11 13:51:13 INFO  Executor:54 - Finished task 0.0 in stage 3.0 (TID 4). 1394 bytes result sent to driver
2019-09-11 13:51:13 INFO  TaskSetManager:54 - Finished task 0.0 in stage 3.0 (TID 4) in 852 ms on localhost (executor driver) (1/1)
2019-09-11 13:51:13 INFO  TaskSchedulerImpl:54 - Removed TaskSet 3.0, whose tasks have all completed, from pool
2019-09-11 13:51:13 INFO  DAGScheduler:54 - ResultStage 3 (jdbc at NativeMethodAccessorImpl.java:0) finished in 0.860 s
2019-09-11 13:51:13 INFO  DAGScheduler:54 - Job 3 finished: jdbc at NativeMethodAccessorImpl.java:0, took 0.861882 s
Wed Sep 11 13:51:13 UTC 2019 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL
connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. Yo
u need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2019-09-11 13:51:13 INFO  SparkContext:54 - Invoking stop() from shutdown hook
2019-09-11 13:51:13 INFO  AbstractConnector:318 - Stopped Spark@7a7ca56d{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2019-09-11 13:51:13 INFO  SparkUI:54 - Stopped Spark web UI at http://ip-20-0-41-62.ec2.internal:4040
2019-09-11 13:51:13 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-09-11 13:51:13 INFO  MemoryStore:54 - MemoryStore cleared
2019-09-11 13:51:13 INFO  BlockManager:54 - BlockManager stopped
2019-09-11 13:51:13 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-09-11 13:51:13 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-09-11 13:51:13 INFO  SparkContext:54 - Successfully stopped SparkContext
2019-09-11 13:51:13 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-09-11 13:51:13 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-4bd22e04-555d-471b-8f37-2c00ea05a25f
2019-09-11 13:51:13 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-dd6e3f95-b573-4102-b4b8-1fc994da3763/pyspark-19f538ae-2cd0-49b1-af0c-198eec320c6a
2019-09-11 13:51:13 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-dd6e3f95-b573-4102-b4b8-1fc994da3763
[edureka_685931@ip-20-0-41-62 ~]$
```

Streaming_prediction.scala

The streaming prediction is done in scala it uses existing model generated and kafka to stream the features and predict the demand then finally push them to RDBMS connected.