BDA ASSIGNMENT-2

USING POSTGRES AND APACHE SPARK

Command used - spark-submit --driver-class-path C:\postgresql-42.2.19.jar pythonscript.py

Challenges faced - The data is already stored from previous assignment, difficulty is in to connect with right APIs postgres Server with Pyspark.

1a.

```
21/03/11 18:02:09 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 2.709
21/03/11 18:02:09 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this jo
21/03/11 18:02:09 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
21/03/11 18:02:09 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAccessorImpl.java:0, took 6.058977 s
21/03/11 18:02:09 INFO CodeGenerator: Code generated in 28.1936 ms
21/03/11 18:02:09 INFO CodeGenerator: Code generated in 28.2614 ms
+-------------------+--------------------+---------------------+
|               date|               event|count(pull_requestid)|
+-------------------+--------------------+---------------------+
|2010-09-02 00:00:00|opened              ...|                    2|
|2010-09-06 00:00:00|opened              ...|                    1|
|2010-09-08 00:00:00|opened              ...|                    1|
|2010-09-09 00:00:00|opened              ...|                    4|
|2010-09-10 00:00:00|opened              ...|                    3|
|2010-09-11 00:00:00|opened              ...|                    3|
|2010-09-12 00:00:00|opened              ...|                    3|
|2010-09-13 00:00:00|opened              ...|                    3|
|2010-09-15 00:00:00|opened              ...|                    2|
|2010-09-16 00:00:00|opened              ...|                    2|
|2010-09-18 00:00:00|opened              ...|                    6|
|2010-09-19 00:00:00|opened              ...|                    4|
|2010-09-20 00:00:00|opened              ...|                    2|
|2010-09-22 00:00:00|opened              ...|                    1|
|2010-09-23 00:00:00|opened              ...|                    4|
|2010-09-24 00:00:00|opened              ...|                    5|
|2010-09-25 00:00:00|opened              ...|                    5|
|2010-09-27 00:00:00|opened              ...|                    4|
|2010-09-28 00:00:00|opened              ...|                    2|
|2010-09-29 00:00:00|opened              ...|                    2|
+-------------------+--------------------+---------------------+
only showing top 20 rows

Total runtime of the program is 30.89186120033264
21/03/11 18:02:10 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:02:10 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:4040
21/03/11 18:02:10 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/03/11 18:02:11 INFO MemoryStore: MemoryStore cleared
21/03/11 18:02:11 INFO BlockManager: BlockManager stopped
21/03/11 18:02:11 INFO BlockManagerMaster: BlockManagerMaster stopped
21/03/11 18:02:11 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
```

1b.

```
21/03/11 18:11:31 INFO TaskSetManager: Finished task 198.0 in stage 1.0 (TID 199) in 83 ms on TANISHQ (executor dri
21/03/11 18:11:31 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
21/03/11 18:11:31 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 3.83
21/03/11 18:11:31 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this j
21/03/11 18:11:31 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
21/03/11 18:11:31 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAccessorImpl.java:0, took 7.823005 s
21/03/11 18:11:31 INFO CodeGenerator: Code generated in 31.2799 ms
21/03/11 18:11:31 INFO CodeGenerator: Code generated in 50.3608 ms
+-------------------+--------------------+---------------------+
|               date|               event|count(pull_requestid)|
+-------------------+--------------------+---------------------+
|2010-09-09 00:00:00|discussed        ...|                    6|
|2010-09-10 00:00:00|discussed        ...|                   10|
|2010-09-11 00:00:00|discussed        ...|                   13|
|2010-09-12 00:00:00|discussed        ...|                    5|
|2010-09-13 00:00:00|discussed        ...|                    7|
|2010-09-14 00:00:00|discussed        ...|                    1|
|2010-09-15 00:00:00|discussed        ...|                    3|
|2010-09-16 00:00:00|discussed        ...|                    2|
|2010-09-17 00:00:00|discussed        ...|                    1|
|2010-09-21 00:00:00|discussed        ...|                    6|
|2010-09-22 00:00:00|discussed        ...|                    3|
|2010-09-23 00:00:00|discussed        ...|                    5|
|2010-09-24 00:00:00|discussed        ...|                    3|
|2010-09-25 00:00:00|discussed        ...|                    5|
|2010-09-27 00:00:00|discussed        ...|                    3|
|2010-09-29 00:00:00|discussed        ...|                    2|
|2010-09-30 00:00:00|discussed        ...|                    3|
|2010-10-01 00:00:00|discussed        ...|                    2|
|2010-10-04 00:00:00|discussed        ...|                    8|
|2010-10-06 00:00:00|discussed        ...|                   15|
+-------------------+--------------------+---------------------+
only showing top 20 rows

Total runtime of the program is 31.272918224334717
21/03/11 18:11:32 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:11:32 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:4040
21/03/11 18:11:32 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

2.

```
21/03/11 18:19:55 INFO DAGScheduler: ResultStage 2 (showString at Nativ
21/03/11 18:19:55 INFO DAGScheduler: Job 0 is finished. Cancelling pote
21/03/11 18:19:55 INFO TaskSchedulerImpl: Killing all running tasks in
21/03/11 18:19:55 INFO DAGScheduler: Job 0 finished: showString at Nati
21/03/11 18:19:55 INFO CodeGenerator: Code generated in 29.3556 ms
21/03/11 18:19:55 INFO CodeGenerator: Code generated in 34.1077 ms
+---------+--------------------+
|date_time|max(number_of_times)|
+---------+--------------------+
|        1|                 828|
|        2|                 555|
|        3|                 580|
|        4|                 758|
|        5|                 915|
|        6|                 582|
|        7|                 579|
|        8|                 648|
|        9|                 585|
|       10|                 667|
|       11|                 590|
|       12|                 546|
+---------+--------------------+

Total runtime of the program is 43.52223587036133
21/03/11 18:19:56 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:19:56 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:
21/03/11 18:19:56 INFO BlockManagerInfo: Removed broadcast_2_piece0 on
21/03/11 18:19:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTracker
```

3.

```
21/03/11 18:23:21 INFO CodeGenerator: Code generated in 24.7689 ms
21/03/11 18:23:21 INFO CodeGenerator: Code generated in 20.9016 ms
+---------+--------------------+
|date_time|max(number_of_times)|
+---------+--------------------+
|        1|                 298|
|        2|                 134|
|        3|                 119|
|        4|                 210|
|        5|                 158|
|        6|                 124|
|        7|                 141|
|        8|                 150|
|        9|                 151|
|       10|                  93|
|       11|                  96|
|       12|                 136|
|       13|                 210|
|       14|                 168|
|       15|                 191|
|       16|                 188|
|       17|                 169|
|       18|                 206|
|       19|                 190|
|       20|                 207|
+---------+--------------------+
only showing top 20 rows

Total runtime of the program is 49.235933065414 43
21/03/11 18:23:22 INFO SparkContext: Invoking stop() from shutdown hoo
21/03/11 18:23:22 INFO SparkUI: Stopped Spark web UI at http://TANISHQ
21/03/11 18:23:22 INFO MapOutputTrackerMasterEndpoint: MapOutputTracke
```

4.

```
21/03/11 18:27:46 INFO TaskSchedulerImpl: Killing all running tasks in sta
21/03/11 18:27:46 INFO DAGScheduler: Job 0 finished: showString at NativeM
21/03/11 18:27:46 INFO CodeGenerator: Code generated in 27.148 ms
21/03/11 18:27:46 INFO CodeGenerator: Code generated in 25.8119 ms
+-------------------+---------------------+
|              weeks|count(pull_requestid)|
+-------------------+---------------------+
|2010-08-30 00:00:00|                    2|
|2010-09-06 00:00:00|                   15|
|2010-09-13 00:00:00|                   17|
|2010-09-20 00:00:00|                   17|
|2010-09-27 00:00:00|                   13|
|2010-10-04 00:00:00|                   10|
|2010-10-11 00:00:00|                    5|
|2010-10-18 00:00:00|                    5|
|2010-10-25 00:00:00|                    3|
|2010-11-01 00:00:00|                    4|
|2010-11-08 00:00:00|                    9|
|2010-11-15 00:00:00|                    8|
|2010-11-22 00:00:00|                    9|
|2010-11-29 00:00:00|                    6|
|2010-12-06 00:00:00|                    5|
|2010-12-13 00:00:00|                    6|
|2010-12-20 00:00:00|                    7|
|2010-12-27 00:00:00|                    4|
|2011-01-03 00:00:00|                    9|
|2011-01-10 00:00:00|                    7|
+-------------------+---------------------+
only showing top 20 rows

Total runtime of the program is 31.708921194076538
21/03/11 18:27:47 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:27:47 INFO BlockManagerInfo: Removed broadcast_1_piece0 on IAN
```

5.

```
21/03/11 18:34:13 INFO Executor: Finished task 197.0 in stage 1.0 (TID 198). 5598 by
21/03/11 18:34:13 INFO TaskSetManager: Finished task 199.0 in stage 1.0 (TID 200) in
21/03/11 18:34:13 INFO TaskSetManager: Finished task 196.0 in stage 1.0 (TID 197) in
21/03/11 18:34:13 INFO TaskSetManager: Finished task 197.0 in stage 1.0 (TID 198) in
21/03/11 18:34:13 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all
21/03/11 18:34:13 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccess
21/03/11 18:34:13 INFO DAGScheduler: Job 0 is finished. Cancelling potential specula
21/03/11 18:34:13 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stag
21/03/11 18:34:13 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAcces
21/03/11 18:34:13 INFO CodeGenerator: Code generated in 21.7365 ms
+------+--------------------+
|months|count(pull_requestid)|
+------+--------------------+
+------+--------------------+

Total runtime of the program is 28.406574249267578
21/03/11 18:34:14 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:34:14 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:4040
21/03/11 18:34:14 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoin
21/03/11 18:34:14 INFO MemoryStore: MemoryStore cleared
21/03/11 18:34:14 INFO BlockManager: BlockManager stopped
21/03/11 18:34:14 INFO BlockManagerMaster: BlockManagerMaster stopped
21/03/11 18:34:14 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: Outp
```

6.

```
21/03/11 18:37:45 INFO TaskSchedulerImpl: Killing all running tasks in st
21/03/11 18:37:45 INFO DAGScheduler: Job 0 finished: showString at Native
21/03/11 18:37:45 INFO CodeGenerator: Code generated in 22.664 ms
21/03/11 18:37:45 INFO CodeGenerator: Code generated in 23.5522 ms
+-------------------+------------+
|               date|count(event)|
+-------------------+------------+
|2010-09-02 00:00:00|           2|
|2010-09-06 00:00:00|           1|
|2010-09-08 00:00:00|           1|
|2010-09-09 00:00:00|          10|
|2010-09-10 00:00:00|          13|
|2010-09-11 00:00:00|          16|
|2010-09-12 00:00:00|           8|
|2010-09-13 00:00:00|          10|
|2010-09-14 00:00:00|           1|
|2010-09-15 00:00:00|           5|
|2010-09-16 00:00:00|           4|
|2010-09-17 00:00:00|           1|
|2010-09-18 00:00:00|           6|
|2010-09-19 00:00:00|           4|
|2010-09-20 00:00:00|           2|
|2010-09-21 00:00:00|           6|
|2010-09-22 00:00:00|           4|
|2010-09-23 00:00:00|           9|
|2010-09-24 00:00:00|           8|
|2010-09-25 00:00:00|          10|
+-------------------+------------+
only showing top 20 rows

Total runtime of the program is 27.561600923538208
21/03/11 18:37:46 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 18:37:46 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:40
21/03/11 18:37:46 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMa
21/03/11 18:37:46 INFO MemoryStore: MemoryStore cleared
```

7.

```
21/03/11 17:55:49 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 2.969
21/03/11 17:55:49 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this jo
21/03/11 17:55:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
21/03/11 17:55:49 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAccessorImpl.java:0, took 7.378705 s
21/03/11 17:55:49 INFO CodeGenerator: Code generated in 26.9968 ms
21/03/11 17:55:49 INFO CodeGenerator: Code generated in 28.9618 ms
+-------------------+-------------------+---------------------+
|               date|              event|count(pull_requestid)|
+-------------------+-------------------+---------------------+
|2010-09-02 00:00:00|opened           ...|                    2|
|2010-09-06 00:00:00|opened           ...|                    1|
|2010-09-08 00:00:00|opened           ...|                    1|
|2010-09-09 00:00:00|opened           ...|                    4|
|2010-09-10 00:00:00|opened           ...|                    3|
|2010-09-11 00:00:00|opened           ...|                    3|
|2010-09-12 00:00:00|opened           ...|                    3|
|2010-09-13 00:00:00|opened           ...|                    3|
|2010-09-15 00:00:00|opened           ...|                    2|
|2010-09-16 00:00:00|opened           ...|                    2|
|2010-09-18 00:00:00|opened           ...|                    6|
|2010-09-19 00:00:00|opened           ...|                    4|
|2010-09-20 00:00:00|opened           ...|                    2|
|2010-09-22 00:00:00|opened           ...|                    1|
|2010-09-23 00:00:00|opened           ...|                    4|
|2010-09-24 00:00:00|opened           ...|                    5|
|2010-09-25 00:00:00|opened           ...|                    5|
|2010-09-27 00:00:00|opened           ...|                    4|
|2010-09-28 00:00:00|opened           ...|                    2|
|2010-09-29 00:00:00|opened           ...|                    2|
+-------------------+-------------------+---------------------+
only showing top 20 rows

Total runtime of the program is 51.09798216819763
21/03/11 17:55:51 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 17:55:51 INFO SparkUI: Stopped Spark web UI at http://TANISHQ:4040
21/03/11 17:55:51 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/03/11 17:55:51 INFO MemoryStore: MemoryStore cleared
21/03/11 17:55:51 INFO BlockManager: BlockManager stopped
21/03/11 17:55:51 INFO BlockManagerMaster: BlockManagerMaster stopped
21/03/11 17:55:51 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
```

Methodology:

First connect to server of postgres at port 5432, then create SparkSession and,

1). In case of reading data, copy data from postgres into dataframe and then running queries on that dataframe.

2). In case of sending query, without importing data from postgres directly send queries into postgres.

4 Python files attached with doc each case with either 1 executor or 2 executor

Timing Analysis Step:

Avg. time with 1 executor to take data from postgres in dataframe and run query on it :

1. Total runtime of the program is 30.89186120033264
2. Total runtime of the program is 43.522223587036133
3. Total runtime of the program is 49.23593306541443
4. Total runtime of the program is 31.708921194076538

5. Total runtime of the program is 28.406574249267578
6. Total runtime of the program is 27.561600923538208
7. Total runtime of the program is 31.09798216819763

Avg. time with 2 executor to take data from postgres in dataframe and run query on it :

1. Total runtime of the program is 33.79321050643921
2. Total runtime of the program is 43.42631006240845
3. Total runtime of the program is 69.37893843650818
4. Total runtime of the program is 51.32006812095642
5. Total runtime of the program is 36.27773976325989
6. Total runtime of the program is 38.92901921272278
7. Total runtime of the program is 37.707122802734375

Avg. time with 1 executor to directly run query on backend postgres :

1.Total runtime of the program is 28.64920473098755
2.Total runtime of the program is 28.333160161972046
3.Total runtime of the program is 26.358773231506348
4.Total runtime of the program is 29.958852291107178
5.Total runtime of the program is 29.1346538066864
6.Total runtime of the program is 29.53973937034607
7.Total runtime of the program is 26.611144304275513

Avg. time with 2 executor to directly run query on backend postgres :

1.Total runtime of the program is 29.964597463607788
2.Total runtime of the program is 26.782146453857422
3.Total runtime of the program is 28.815183401107788
4.Total runtime of the program is 28.66307258605957
5.Total runtime of the program is 26.54670023918152
6.Total runtime of the program is 26.126744270324707
7.Total runtime of the program is 28.201881647109985

Merits and Demerits

As we can see above that there is difference between the run time of the RDD and the Direct queries, so basically time is the main constraint here. In direct queries we are getting less time and in the case of RDD we are getting more time as we need to load the data.

Learning:

We learnt how to install apache spark , postgresql, and how to connect spark with the postgresql using pyspark. We learnt different backend machines like mongodb and hdfs. As we were not able to connect all these we were getting the errors, but now we have learnt how to write code in the pyspark for connecting the apache spark and the postgresql. Also we learnt how to work on the scala and write the queries there. We learnt many new ways to fix the bugs in the databases and their installation system.