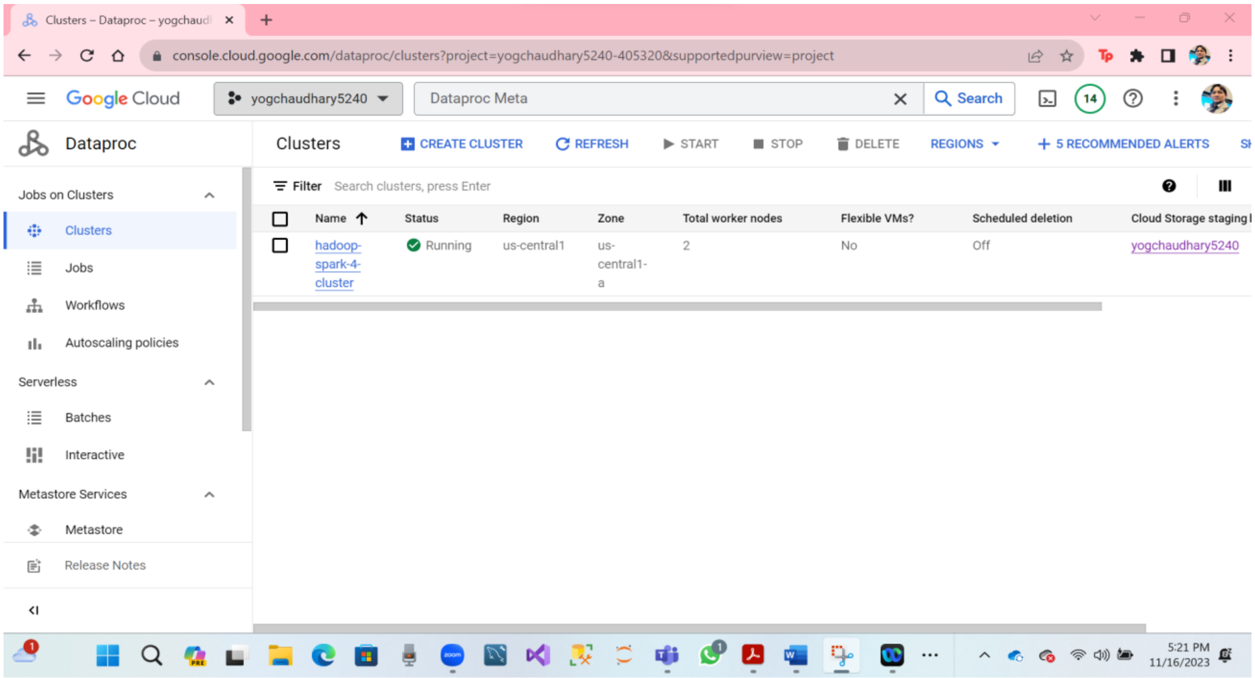
Yog Chaudhary
11727095
ADTA 5240 Week 13’th (harvesting, Storing, And Retrieving Data)
Professor: Dr. Zeynep Orhan
University Of North Texas

Nov 16, 2023

Homework Assignment: Spark Queries

Step 1. For a Google Console.

- I click on three horizontal lines.
- In the search bar I typed Cluster and clicked on Clusters Dataproc.
- Hare screenshot.



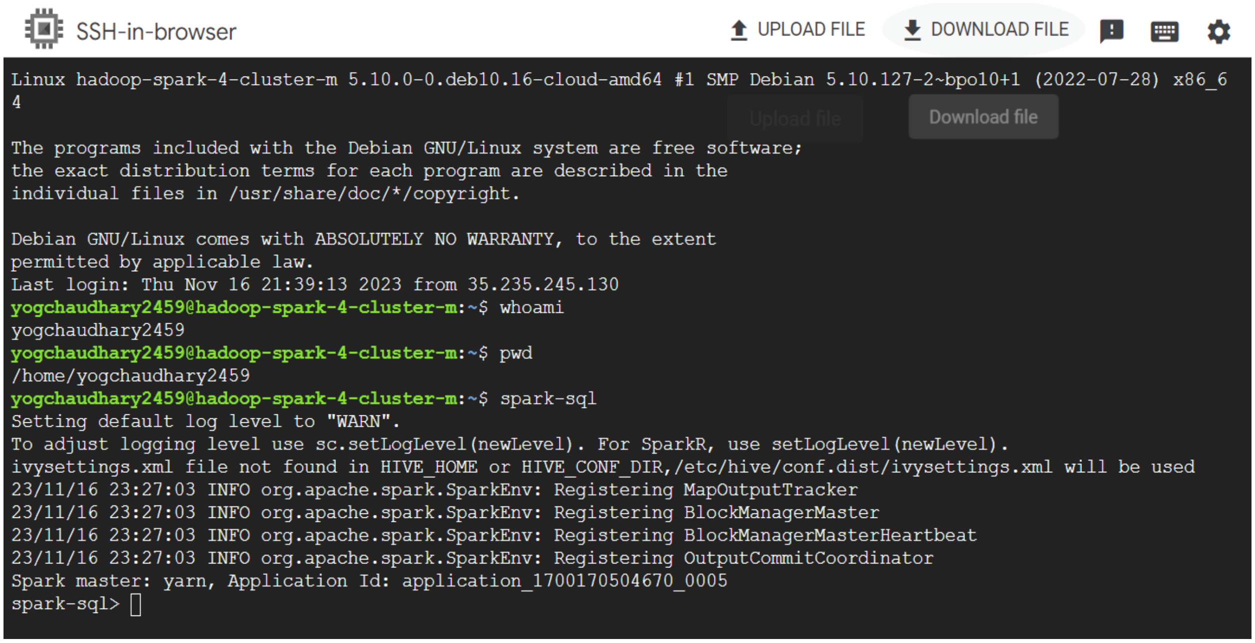
console.cloud.google.com/dataproc/clusters?project=yogchaudhary5240-405320&supportedpurview=project

Filter Search clusters, press Enter

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging
hadoop-spark-4-cluster	Running	us-central1	us-central1-a	2	No	Off	yogchaudhary5240

Step 2. We used the command to start spark SQL: spark-sql

- We are running a spark and connected to the Metastore.
- Hare screenshot.

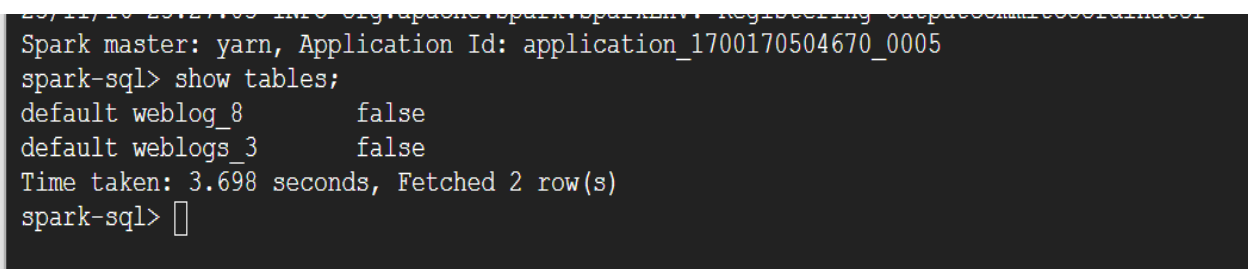


```
Linux hadoop-spark-4-cluster-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_6
4

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Nov 16 21:39:13 2023 from 35.235.245.130
yogchaudhary2459@hadoop-spark-4-cluster-m:~$ whoami
yogchaudhary2459
yogchaudhary2459@hadoop-spark-4-cluster-m:~$ pwd
/home/yogchaudhary2459
yogchaudhary2459@hadoop-spark-4-cluster-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
23/11/16 23:27:03 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/11/16 23:27:03 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/11/16 23:27:03 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/11/16 23:27:03 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1700170504670_0005
spark-sql> 
```

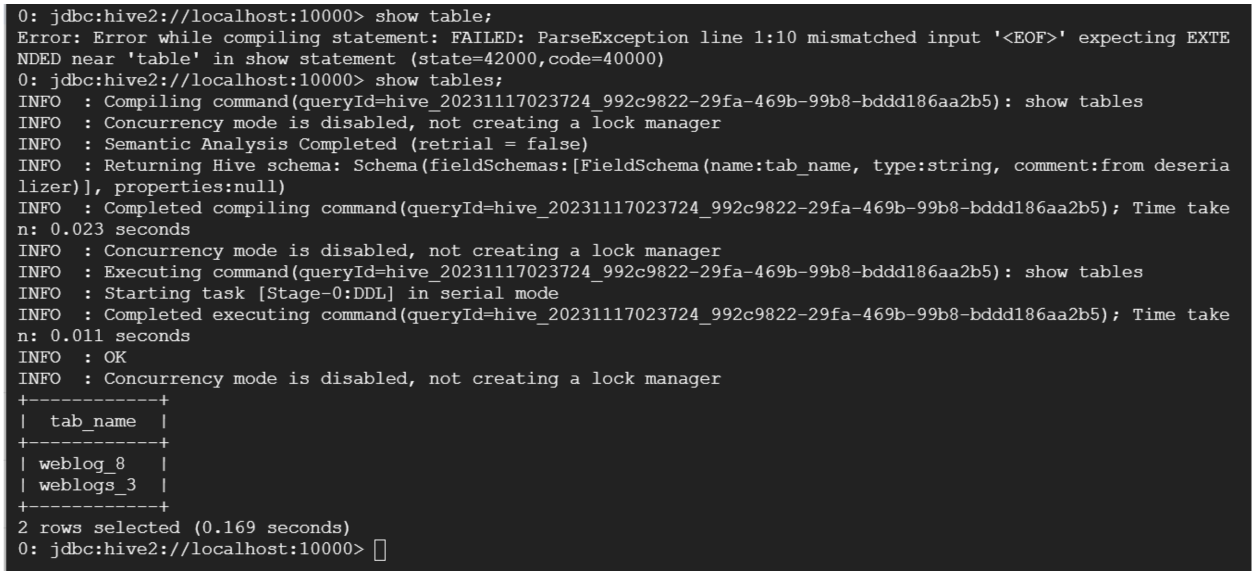
Step 3. We created a table in the spark using the “show tables” commands.



```
23/11/16 23:27:03 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1700170504670_0005
spark-sql> show tables;
default weblog_8        false
default weblogs_3       false
Time taken: 3.698 seconds, Fetched 2 row(s)
spark-sql> 
```

Step 4. Let’s run Hive Spark, we can make a comparison between Hive and Spark.

- Open a new SSH through the cluster SSH.
- Start running Hiving using the following commands. (beeline -u jdbc:hive2://localhost:10000)
- We can show tables in Having using the following commands: (show tables;)
- Below the screen the table has two see spark (weblogs_3 and weblog_8)
- Then, both Hive and Spark work with the Metastore.
- Hare screenshot

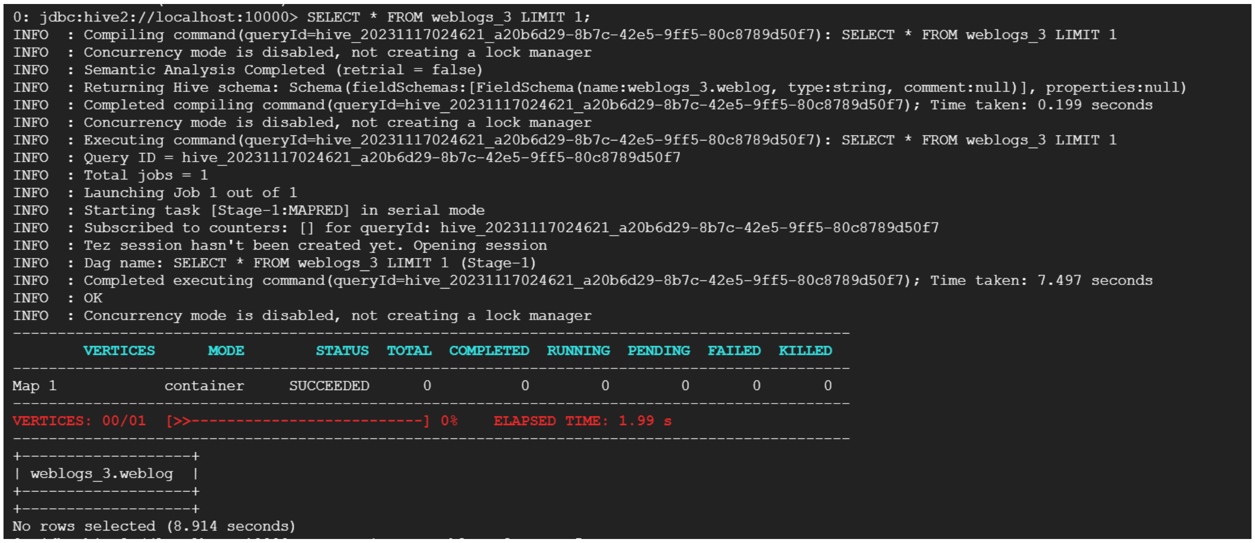


```
0: jdbc:hive2://localhost:10000> show table;
Error: Error while compiling statement: FAILED: ParseException line 1:10 mismatched input '<EOF>' expecting EXTE
NDED near 'table' in show statement (state=42000,code=40000)
0: jdbc:hive2://localhost:10000> show tables;
INFO  : Compiling command(queryId=hive_20231117023724_992c9822-29fa-469b-99b8-bddd186aa2b5): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deseria
lizer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231117023724_992c9822-29fa-469b-99b8-bddd186aa2b5); Time take
n: 0.023 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231117023724_992c9822-29fa-469b-99b8-bddd186aa2b5): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231117023724_992c9822-29fa-469b-99b8-bddd186aa2b5); Time take
n: 0.011 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+
|  tab_name  |
+------------+
| weblog_8   |
| weblogs_3  |
+------------+
2 rows selected (0.169 seconds)
0: jdbc:hive2://localhost:10000>
```

A) We ran the following query in the have below.

SELECT * FROM weblogs_3 LIMIT 1;

- Hare screenshot

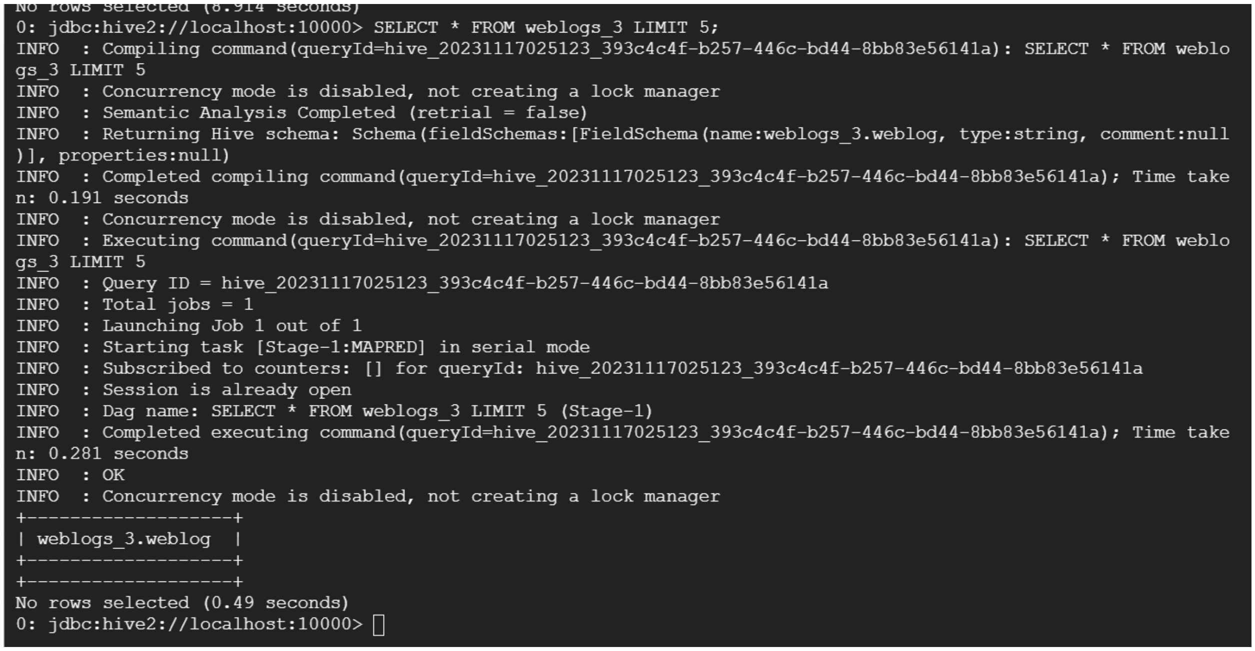


```
0: jdbc:hive2://localhost:10000> SELECT * FROM weblogs_3 LIMIT 1;
INFO  : Compiling command(queryId=hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7): SELECT * FROM weblogs_3 LIMIT 1
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:weblogs_3.weblog, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7); Time taken: 0.199 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7): SELECT * FROM weblogs_3 LIMIT 1
INFO  : Query ID = hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT * FROM weblogs_3 LIMIT 1 (Stage-1)
INFO  : Completed executing command(queryId=hive_20231117024621_a20b6d29-8b7c-42e5-9ff5-80c8789d50f7); Time taken: 7.497 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/01  [>>--------------------------] 0%    ELAPSED TIME: 1.99 s
----------------------------------------------------------------------------------------------
+-------------------+
| weblogs_3.weblog  |
+-------------------+
+-------------------+
No rows selected (8.914 seconds)
```

B) We can also run the following query in have below.

SELECT * FROM weblogs_3 LIMIT 5;

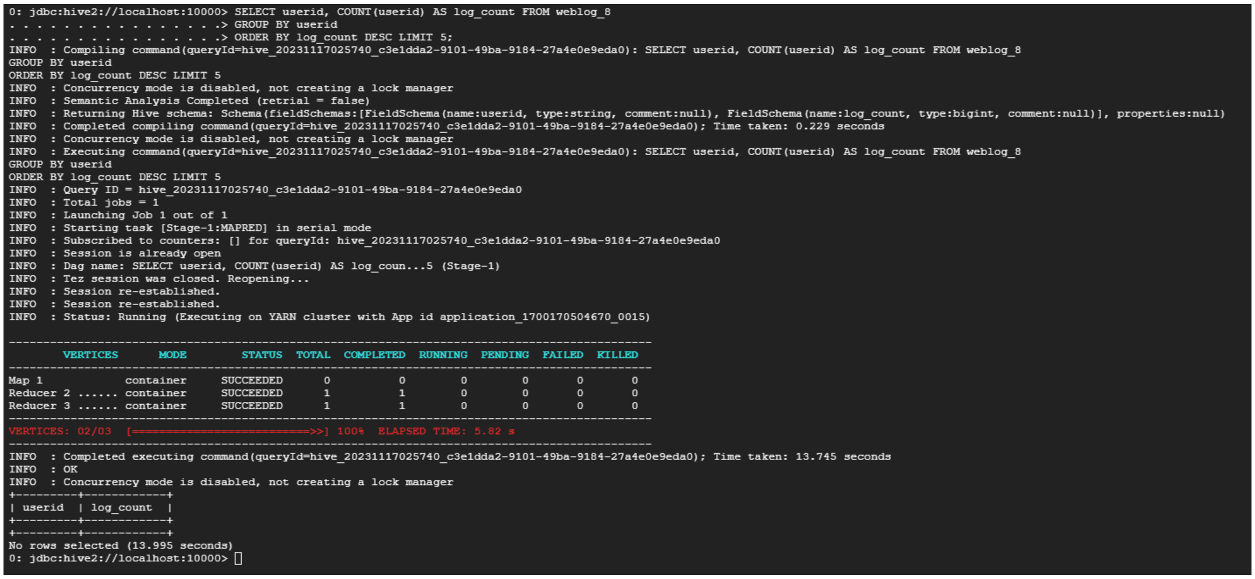
- Hare screenshot



```
No rows selected (8.914 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM weblogs_3 LIMIT 5;
INFO  : Compiling command(queryId=hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a): SELECT * FROM weblo
gs_3 LIMIT 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:weblogs_3.weblog, type:string, comment:null
)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a); Time take
n: 0.191 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a): SELECT * FROM weblo
gs_3 LIMIT 5
INFO  : Query ID = hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a
INFO  : Session is already open
INFO  : Dag name: SELECT * FROM weblogs_3 LIMIT 5 (Stage-1)
INFO  : Completed executing command(queryId=hive_20231117025123_393c4c4f-b257-446c-bd44-8bb83e56141a); Time take
n: 0.281 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------+
| weblogs_3.weblog  |
+--------------------+
+--------------------+
No rows selected (0.49 seconds)
0: jdbc:hive2://localhost:10000> 
```

C) We can use complex queries and now nun it will show.

- SELECT userid, COUNT(userid) AS log_count FROM weblog_8 GROUP BY userid ORDER BY log_count DESC LIMIT 5;
- Hare screenshot.

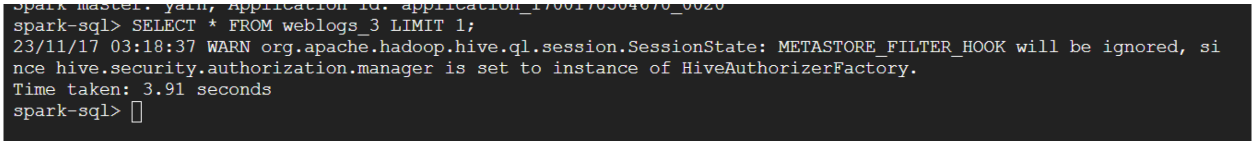


```
0: jdbc:hive2://localhost:10000> SELECT userid, COUNT(userid) AS log_count FROM weblog_8
. . . . . . . . . . . . . . . .> GROUP BY userid
. . . . . . . . . . . . . . . .> ORDER BY log_count DESC LIMIT 5;
INFO  : Compiling command(queryId=hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0): SELECT userid, COUNT(userid) AS log_count FROM weblog_8
GROUP BY userid
ORDER BY log_count DESC LIMIT 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:userid, type:string, comment:null), FieldSchema(name:log_count, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0); Time taken: 0.229 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0): SELECT userid, COUNT(userid) AS log_count FROM weblog_8
GROUP BY userid
ORDER BY log_count DESC LIMIT 5
INFO  : Query ID = hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0
INFO  : Session is already open
INFO  : Dag name: SELECT userid, COUNT(userid) AS log_coun...5 (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1700170504670_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/03  [==========================>>] 100%  ELAPSED TIME: 5.82 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20231117025740_c3e1dda2-9101-49ba-9184-27a4e0e9eda0); Time taken: 13.745 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------+------------+
| userid  | log_count  |
+---------+------------+
+---------+------------+
No rows selected (13.995 seconds)
0: jdbc:hive2://localhost:10000> 
```

Step 5. We can go back to the terminal, which is running Spark API. We will return the queries.

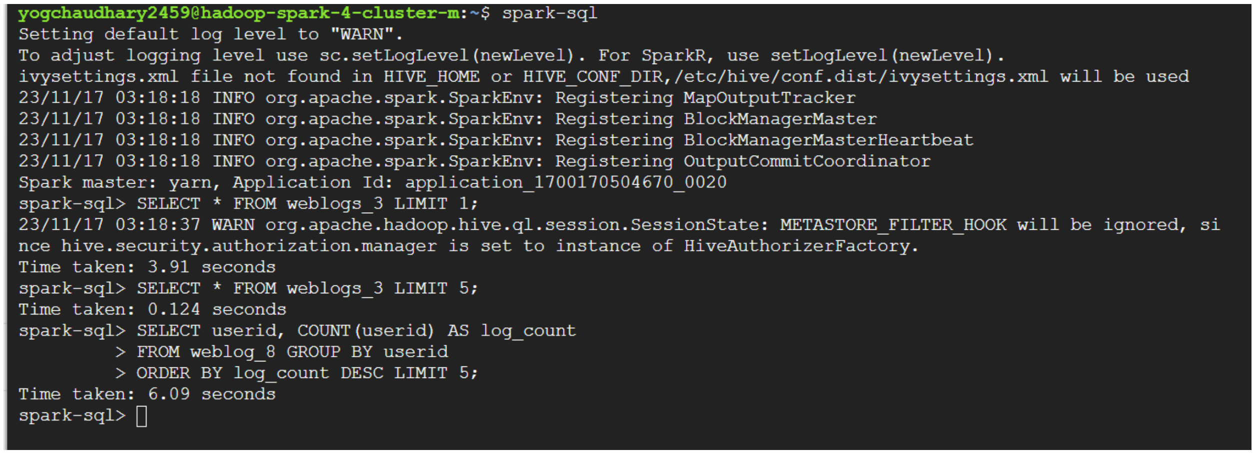
- **SELECT * FROM weblogs_3 LIMIT 1;**



```
Spark master: yarn, Application Id: application_1700170504670_0020
spark-sql> SELECT * FROM weblogs_3 LIMIT 1;
23/11/17 03:18:37 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, si
nce hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
Time taken: 3.91 seconds
spark-sql> 
```

Now, we can run both queries in the below.

- **SELECT * FROM weblogs_3 LIMIT 5;**
- **SELECT userid, COUNT(userid) AS log_count FROM weblog_8 GROUP BY userid ORDER BY log_count DESC LIMIT 5;**



```
yogchaudhary2459@hadoop-spark-4-cluster-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
23/11/17 03:18:18 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/11/17 03:18:18 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/11/17 03:18:18 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/11/17 03:18:18 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1700170504670_0020
spark-sql> SELECT * FROM weblogs_3 LIMIT 1;
23/11/17 03:18:37 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, si
nce hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
Time taken: 3.91 seconds
spark-sql> SELECT * FROM weblogs_3 LIMIT 5;
Time taken: 0.124 seconds
spark-sql> SELECT userid, COUNT(userid) AS log_count
         > FROM weblog_8 GROUP BY userid
         > ORDER BY log_count DESC LIMIT 5;
Time taken: 6.09 seconds
spark-sql> 
```

Finally, we have compared the time in Hive and in Spark. We can see running the same command lines in both Spark and Hive environments depends on the Metastore components of the Hive to understand the data structure that can perform the queries on the data.