

Group A

LAB 2b

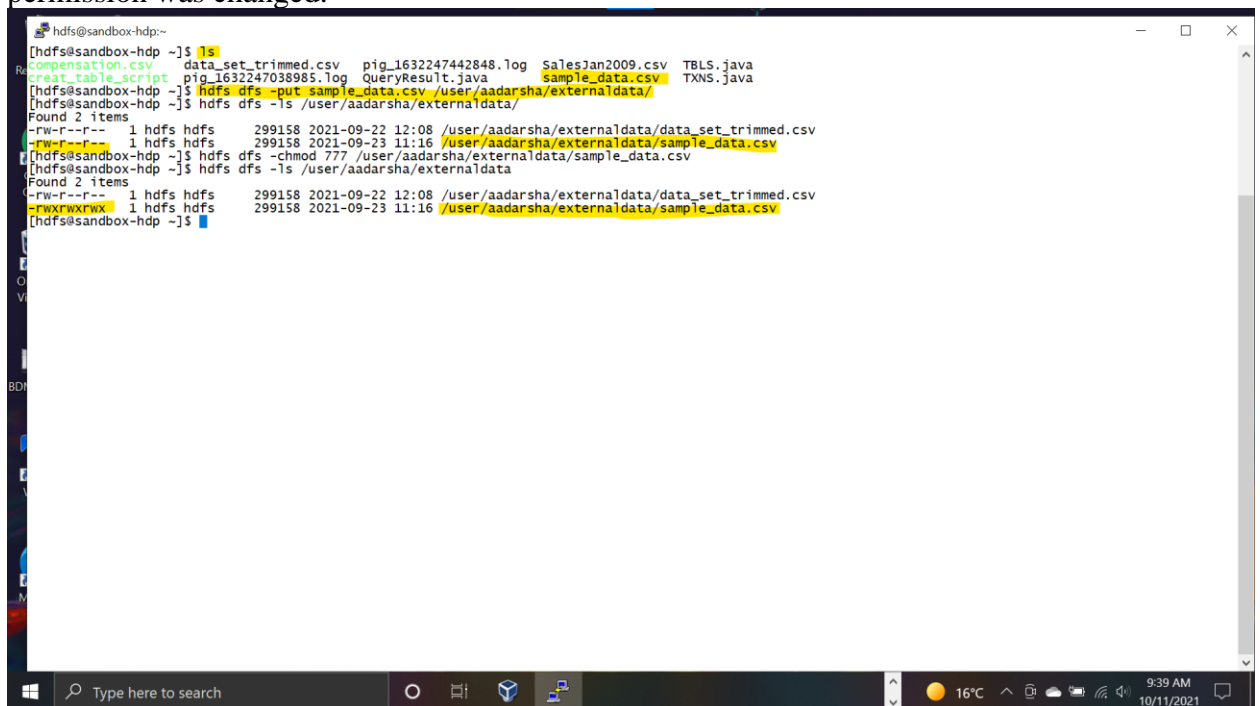
Hive Part1

Participants:

1. Aadarsha chapagain
2. Jyoti shukla
3. Rishi Phaneendra Varma
4. Priti Bhale
5. Sreya Treasa Johny
6. Piyush Bhatia

Objective: *The objective is to find the number of male staff who has been serving more than 15 years and are still active.*

1. The public data set was uploaded with winscp and copied to hdfs file system and permission was changed.



```
hdfs@sandbox-hdp:~$ ls
compensation.csv  data_set_trimmed.csv  pig_1632247442848.log  SalesJan2009.csv  TBLS.java
creat_table_script  pig_1632247038985.log  QueryResult.java      sample_data.csv   TXNS.java
hdfs@sandbox-hdp:~$ hdfs dfs -put sample_data.csv /user/aadarsha/externaldata/
hdfs@sandbox-hdp:~$ hdfs dfs -ls /user/aadarsha/externaldata/
Found 2 items
-rw-r--r-- 1 hdfs hdfs 299158 2021-09-22 12:08 /user/aadarsha/externaldata/data_set_trimmed.csv
-rw-r--r-- 1 hdfs hdfs 299158 2021-09-23 11:16 /user/aadarsha/externaldata/sample_data.csv
hdfs@sandbox-hdp:~$ hdfs dfs -chmod 777 /user/aadarsha/externaldata/sample_data.csv
hdfs@sandbox-hdp:~$ hdfs dfs -ls /user/aadarsha/externaldata
Found 2 items
-rw-r--r-- 1 hdfs hdfs 299158 2021-09-22 12:08 /user/aadarsha/externaldata/data_set_trimmed.csv
-rwxrwxrwx 1 hdfs hdfs 299158 2021-09-23 11:16 /user/aadarsha/externaldata/sample_data.csv
hdfs@sandbox-hdp:~$
```

2. Managed table without partition

```
hdfs@sandbox-hdp:~$
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=hive;serviceDiscoveryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2
21/09/23 11:30:12 [main]: INFO jdbc.HiveConnection: Connected to sandbox-hdp.hortonworks.com:10000
Connected to: Apache Hive (version 3.1.0.3.0.1.0-187)
Driver: Hive JDBC (version 3.1.0.3.0.1.0-187)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.0.3.0.1.0-187 by Apache Hive
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Create table staff_man(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
STATUS string,
BUSINESS_UNIT string,
gender_full string)
row format delimited fields terminated by ' ' stored as textfile tblproperties("skip.header.line.count"=1);
INFO : Compiling command(queryId=hive_20210923113015_66ef67a4-116d-44ca-b24a-542562758af9): Create table staff_man(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
STATUS string,
BUSINESS_UNIT string,
gender_full string)
row format delimited fields terminated by ' ' stored as textfile tblproperties("skip.header.line.count"=1)
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20210923113015_66ef67a4-116d-44ca-b24a-542562758af9); Time taken: 0.103 seconds
INFO : Executing command(queryId=hive_20210923113015_66ef67a4-116d-44ca-b24a-542562758af9): Create table staff_man(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
STATUS string,
BUSINESS_UNIT string,
gender_full string)
row format delimited fields terminated by ' ' stored as textfile tblproperties("skip.header.line.count"=1)
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20210923113015_66ef67a4-116d-44ca-b24a-542562758af9); Time taken: 0.214 seconds
INFO : OK
No rows affected (0.53 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

Analysis:

To achieve the objective the following query was run.

select count() from staff_man where gender_full='Male' and length_of_service > 15 and status='ACTIVE';*

```
hdfs@sandbox-hdp:~$
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select count(*) from staff_man where gender_full='Male' and length_of_service > 15 and status='ACTIVE';
INFO : Compiling command(queryId=hive_20210923120645_3667f56f-165f-43ff-84a9-ea3c8be1314e): select count(*) from staff_man where gender_full='Male' and length_of_service > 15 and status='ACTIVE'
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchema(name: c0, type: bigint, comment: null), properties: null)
INFO : Completed compiling command(queryId=hive_20210923120645_3667f56f-165f-43ff-84a9-ea3c8be1314e); Time taken: 0.217 seconds
INFO : Executing command(queryId=hive_20210923120645_3667f56f-165f-43ff-84a9-ea3c8be1314e): select count(*) from staff_man where gender_full='Male' and length_of_service > 15 and status='ACTIVE'
INFO : Query ID = hive_20210923120645_3667f56f-165f-43ff-84a9-ea3c8be1314e
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_20210923120645_3667f56f-165f-43ff-84a9-ea3c8be1314e
INFO : Session is already open
INFO : Dag name: select count(*) from staff...status='ACTIVE' (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_163227054967_0039)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1          1          0          0          0          0
Reducer 2 ..... container  SUCCEEDED    1          1          0          0          0          0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 14.37s
-----
INFO : Status: DAG finished successfully in 14.37 seconds
INFO : Query Execution Summary
INFO : OPERATION      DURATION
INFO : -----
INFO : Compile Query      0.22s
INFO : Prepare Plan      0.09s
INFO : Get Query Coordinator (AM) 0.00s
INFO : Submit Plan      0.07s
INFO : Start DAG        0.63s
INFO : Run DAG          14.37s
INFO : -----
INFO : Task Execution Summary
INFO : VERTICES      DURATION(ms)  CPU_TIME(ms)  GC_TIME(ms)  INPUT_RECORDS  OUTPUT_RECORDS
INFO : -----
INFO : Map 1          9942.00      5,910         548           4,964           1
INFO : Reducer 2      1872.00      1,250         0              1              0
INFO : -----
INFO : org.apache.tez.common.counters.DAGCounter:
```


4. External table without partition

Keyword external is used to create a external table;

```
hdfs@sandbox-hdp:~$
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Create external table staff_ext(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
gender_full string,
STATUS string,
BUSINESS_UNIT string)
row format delimited fields terminated by ',' stored as textfile tblproperties("skip.header.line.count"="1")
INFO : Compiling command(queryId=hive_20210923125159_b70ac629-10c3-4514-957d-29e58d596c96): Create external table staff_ext(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
gender_full string,
STATUS string,
BUSINESS_UNIT string)
row format delimited fields terminated by ',' stored as textfile tblproperties("skip.header.line.count"="1")
INFO : Semantic Analysis Completed (Retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20210923125159_b70ac629-10c3-4514-957d-29e58d596c96): Time taken: 0.036 seconds
INFO : Executing command(queryId=hive_20210923125159_b70ac629-10c3-4514-957d-29e58d596c96): Create external table staff_ext(age int,
length_of_service int,
city_name string,
department_name string,
job_title string,
gender_full string,
STATUS string,
BUSINESS_UNIT string)
row format delimited fields terminated by ',' stored as textfile tblproperties("skip.header.line.count"="1")
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20210923125159_b70ac629-10c3-4514-957d-29e58d596c96): Time taken: 0.086 seconds
INFO : OK
No rows affected (0.208 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> load data inpath '/user/aadarsha/externaldata/sample_data.csv' into table staff_ext;
```

Analysis

select count() from staff_ext where gender_full='Male' and length_of_service >15 and status='ACTIVE';*

```
hdfs@sandbox-hdp:~$
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select count(*) from staff_ext where gender_full='Male' and length_of_service >15 and status='ACTIVE';
INFO : Compiling command(queryId=hive_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44): select count(*) from staff_ext where gender_full='Male' and le
length_of_service >15 and status='ACTIVE'
INFO : Semantic Analysis Completed (Retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[FieldSchema(name:c0, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44): Time taken: 0.179 seconds
INFO : Executing command(queryId=hive_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44): select count(*) from staff_ext where gender_full='Male' and le
length_of_service >15 and status='ACTIVE'
INFO : Query ID = hive_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44
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_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44
INFO : Session is already open
INFO : Dag name: select count(*) from staff_ext w...='ACTIVE' (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_1632227054967_0043)

-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FAILED      KILLED
-----
Map 1 ..... container SUCCEEDED      1          1          0          0          0          0
Reducer 2 ..... container SUCCEEDED      1          1          0          0          0          0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 4.13 s
-----
INFO : Status: DAG finished successfully in 4.09 seconds
INFO :
INFO : Query Execution Summary
INFO : -----
INFO : OPERATION      DURATION
INFO : -----
INFO : Compile Query      0.18s
INFO : Prepare Plan      0.07s
INFO : Get Query Coordinator (AM) 0.00s
INFO : Submit Plan      3.24s
INFO : Start DAG      1.03s
INFO : Run DAG      4.09s
INFO : -----
INFO : Task Execution Summary
INFO : -----
INFO : VERTICES      DURATION(ms)      CPU_TIME(ms)      GC_TIME(ms)      INPUT_RECORDS      OUTPUT_RECORDS
INFO : -----
INFO : Map 1          2063.00          2,470          102          4,964          1
INFO : Reducer 2      441.00          540          24          1          0
INFO : -----

INFO : INPUT_DIRECTORIES_Map_1: 1
INFO : INPUT_FILES_Map_1: 1
INFO : RAW_INPUT_SPLITS_Map_1: 1
INFO : Completed executing command(queryId=hive_20210923125756_f0dc5706-39e7-4a8d-a630-33f760981c44): Time taken: 8.447 seconds
INFO : OK
INFO : _c0
1894
1 row selected (8.713 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

Here it is seen that for dynamic partition insert command is used instead of load command and the field by which partition is done is mentioned at last.

select count() from staff_dynamic_part where gender_full='Male' and status='ACTIVE' and length_of_service>15;*

```

root@sandbox-hdp:~#
INFO : Completed compiling command(queryId=hive_20210923171743_cb436b58-c126-4adf-8c1c-905cea774519); Time taken: 0.264 seconds
INFO : Executing command(queryId=hive_20210923171743_cb436b58-c126-4adf-8c1c-905cea774519); show tables
INFO : Starting task [Stage-0:DOL] in serial mode
INFO : Completed executing command(queryId=hive_20210923171743_cb436b58-c126-4adf-8c1c-905cea774519); Time taken: 0.009 seconds
INFO : OK
+-----+
| tab_name |
+-----+
| employee_data |
| salesjan2009_clustered |
| salesjan2009_ext |
| salesjan2009_parquet1 |
| salesjan2009_skewed |
| staff_dynamic_part |
| staff_ext |
| staff_ext_part |
| staff_man |
| staff_man_part |
+-----+
10 rows selected (0.302 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select count(*) from staff_dynamic_part where gender_full='Male' and status='ACTIVE' and length_of_service>15;
INFO : Compiling command(queryId=hive_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb); select count(*) from staff_dynamic_part where gender_full='Male' and status='ACTIVE' and length_of_service>15
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[FieldsSchema(name:c0, type:bigint, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb); Time taken: 0.456 seconds
INFO : Executing command(queryId=hive_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb); select count(*) from staff_dynamic_part where gender_full='Male' and status='ACTIVE' and length_of_service>15
INFO : Query ID = hive_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb
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_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb
INFO : Session is already open
INFO : Dag name: select count(*) from ...length_of_service>15 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1632227054967_0052)
+-----+
| VERTICES | MODE | STATUS | TOTAL | COMPLETED | RUNNING | PENDING | FAILED | KILLED |
+-----+
| Map 1 | container | SUCCEEDED | 1 | 1 | 0 | 0 | 0 | 0 |
| Reducer 2 | container | SUCCEEDED | 1 | 1 | 0 | 0 | 0 | 0 |
+-----+
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 7.10 s
INFO : Status: DAG finished successfully in 7.05 seconds
INFO :

```

And as per our objective:

```

root@sandbox-hdp:~#
INFO : RECORDS_OUT_OPERATOR_FS_14: 1
INFO : RECORDS_OUT_OPERATOR_GBY_11: 1
INFO : RECORDS_OUT_OPERATOR_GBY_13: 1
INFO : RECORDS_OUT_OPERATOR_MAP_0: 0
INFO : RECORDS_OUT_OPERATOR_RS_12: 1
INFO : RECORDS_OUT_OPERATOR_SEL_10: 1893
INFO : RECORDS_OUT_OPERATOR_TS_0: 2079
INFO : TaskCounter_Map_1_INPUT_staff_dynamic_part:
INFO : INPUT_RECORDS_PROCESSED: 2079
INFO : INPUT_SPLIT_LENGTH_BYTES: 109562
INFO : TaskCounter_Map_1_OUTPUT_Reducer_2:
INFO : ADDITIONAL_SPILLS_BYTES_READ: 0
INFO : ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO : ADDITIONAL_SPILL_COUNT: 0
INFO : OUTPUT_BYTES: 5
INFO : OUTPUT_BYTES_PHYSICAL: 51
INFO : OUTPUT_BYTES_WITH_OVERHEAD: 13
INFO : OUTPUT_LARGE_RECORDS: 0
INFO : OUTPUT_RECORDS: 1
INFO : SPILLED_RECORDS: 0
INFO : TaskCounter_Reducer_2_INPUT_Map_1:
INFO : FIRST_EVENT_RECEIVED: 38
INFO : INPUT_RECORDS_PROCESSED: 1
INFO : LAST_EVENT_RECEIVED: 38
INFO : NUM_FAILED_SHUFFLE_INPUTS: 0
INFO : NUM_SHUFFLED_INPUTS: 1
INFO : SHUFFLE_BYTES: 27
INFO : SHUFFLE_BYTES_DECOMPRESSED: 13
INFO : SHUFFLE_BYTES_DISK_DIRECT: 27
INFO : SHUFFLE_BYTES_TO_DISK: 0
INFO : SHUFFLE_BYTES_TO_MEM: 0
INFO : SHUFFLE_PHASE_TIME: 57
INFO : TaskCounter_Reducer_2_OUTPUT_out_Reducer_2:
INFO : OUTPUT_RECORDS: 0
INFO : org.apache.hadoop.hive.q1.exec.tez.HiveInputCounters:
INFO : GROUPED_INPUT_SPLITS_Map_1: 1
INFO : INPUT_DIRECTORIES_Map_1: 1
INFO : INPUT_FILES_Map_1: 1
INFO : RAW_INPUT_SPLITS_Map_1: 1
INFO : Completed executing command(queryId=hive_20210923171919_770030f3-6d17-4408-aba4-2a4071e84feb); Time taken: 8.317 seconds
INFO : OK
+-----+
| _c0 |
+-----+
| 1893 |
+-----+
1 row selected (8.949 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>

```