Group E
Data Technology Solutions

Anirudh Siddula
Davinderjit Singh
Kamaldeep Kaur Dhami
Kirandeep Kaur
Bibek Shah Shankar

Lab 2b:

Objectives:

- Showcase the performance boost of partitioning on tables
- Create Internal and External tables with and without partitioning
- Loading the table using dynamic partition in hive

Dataset Objective:

- Find out the make and model of most affordable motorcycle in India which has at least 40 bhp of power and is not older than 2019

Dataset Source : Kaggle(https://www.kaggle.com/ropali/used-bike-price-in-india)

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:

creating directory 2b for assignment

Source: Kaggle
https://www.kaggle.com/ropali/used-bike-price-in-india

Ln 18, Col 37 150% Windows (CRLF) UTF-8

root@sandbox-hdp:~

```
[root@sandbox-hdp ~]# hdfs dfs -ls /lab/
Found 1 items
drwxrwxrwx   - hdfs hdfs          0 2021-10-06 08:35 /lab/sqoop
[root@sandbox-hdp ~]# hdfs dfs -mkdir /lab/2b
[root@sandbox-hdp ~]# hdfs dfs -ls /lab/
Found 2 items
drwxr-xr-x   - root hdfs          0 2021-10-06 09:46 /lab/2b
drwxrwxrwx   - hdfs hdfs          0 2021-10-06 08:35 /lab/sqoop
[root@sandbox-hdp ~]#
```

10:24 AM
2021-10-11

Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:

Copying bikes.csv from Sandbox's local to HDFS's "/lab/2b" directory

Ln 19, Col 10 150% Windows (CRLF) UTF-8

root@sandbox-hdp:~

```
[root@sandbox-hdp ~]# hdfs dfs -ls /lab/
Found 1 items
drwxrwxrwx   - hdfs hdfs          0 2021-10-06 08:35 /lab/sqoop
[root@sandbox-hdp ~]# hdfs dfs -mkdir /lab/2b
[root@sandbox-hdp ~]# hdfs dfs -ls /lab/
Found 2 items
drwxr-xr-x   - root hdfs          0 2021-10-06 09:46 /lab/2b
drwxrwxrwx   - hdfs hdfs          0 2021-10-06 08:35 /lab/sqoop
[root@sandbox-hdp ~]# hdfs dfs -copyFromLocal /bikes.csv /lab/2b/
[root@sandbox-hdp ~]# hdfs dfs -ls /lab/2b/
Found 1 items
-rw-r--r--   1 root hdfs     727407 2021-10-06 09:48 /lab/2b/bikes.csv
[root@sandbox-hdp ~]#
```

10:28 AM
2021-10-11

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Creating a Database "DTS" and using the same database for the assignment

Ln 17, Col 45 150% Windows (CRLF) UTF-8

root@sandbox-hdp:~

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show databases;
INFO  : Compiling command(queryId=hive_20211006103141_93bcb67c-c079-46a2-b078-39954acf072d): show databases
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)],
 properties:null)
INFO  : Completed compiling command(queryId=hive_20211006103141_93bcb67c-c079-46a2-b078-39954acf072d); Time taken: 0.029 seco
nds
INFO  : Executing command(queryId=hive_20211006103141_93bcb67c-c079-46a2-b078-39954acf072d): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211006103141_93bcb67c-c079-46a2-b078-39954acf072d); Time taken: 0.008 seco
nds
INFO  : OK
+----------------------+
|    database_name     |
+----------------------+
| default              |
| foodmart             |
| hive                 |
| information_schema   |
| sys                  |
+----------------------+
5 rows selected (0.081 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> create database dts;
INFO  : Compiling command(queryId=hive_20211006103147_01b50be8-b347-4aa7-a6dd-98b5fabdb131): create database dts
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211006103147_01b50be8-b347-4aa7-a6dd-98b5fabdb131); Time taken: 0.038 seco
nds
INFO  : Executing command(queryId=hive_20211006103147_01b50be8-b347-4aa7-a6dd-98b5fabdb131): create database dts
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211006103147_01b50be8-b347-4aa7-a6dd-98b5fabdb131); Time taken: 0.048 seco
nds
INFO  : OK
No rows affected (0.116 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> use database dts;
Error: Error while compiling statement: FAILED: ParseException line 1:4 cannot recognize input near 'database' 'dts' '<EOF>'
in switch database statement (state=42000,code=40000)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> use dts;
INFO  : Compiling command(queryId=hive_20211006103204_e78a29c4-2bad-47a1-9f19-96356913eddb): use dts
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211006103204_e78a29c4-2bad-47a1-9f19-96356913eddb); Time taken: 0.041 seco
nds
INFO  : Executing command(queryId=hive_20211006103204_e78a29c4-2bad-47a1-9f19-96356913eddb): use dts
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211006103204_e78a29c4-2bad-47a1-9f19-96356913eddb); Time taken: 0.016 seco
nds
INFO  : OK
No rows affected (0.076 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

11:10 AM
2021-10-11

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Copying bikes.csv from HDFS to a path where hive has read permissions to access.

Ln 18, Col 37 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
[root@sandbox-hdp ~]# su hdfs
[hdfs@sandbox-hdp root]$ clear
[hdfs@sandbox-hdp root]$ hdfs dfs -chmod 777 /lab/2b/bikes.csv
[hdfs@sandbox-hdp root]$ hdfs dfs -ls /user/hive
Found 5 items
drwxr-xr-x   - hive hdfs          0 2018-11-29 19:04 /user/hive/{hive_metastore_warehouse_dir}
drwx------   - hive hdfs          0 2021-10-06 07:55 /user/hive/.Trash
drwxr-xr-x   - hive hdfs          0 2018-11-29 17:56 /user/hive/.hiveJars
drwxr-xr-x   - hive hdfs          0 2021-10-06 11:02 /user/hive/repl
drwxr-xr-x   - hdfs hdfs          0 2021-10-06 08:17 /user/hive/warehouse
[hdfs@sandbox-hdp root]$ hdfs dfs -cp /lab/2b/bikes.csv /user/hive/
[hdfs@sandbox-hdp root]$ hdfs dfs -chown hive:hdfs/user/hive/bikes.csv
-chown: Not enough arguments: expected 2 but got 1
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] [-e] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-v] [-x] <path> ...]
        [-expunge]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-head <file>]
        [-help [cmd ...]]
        [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touch [-a] [-m] [-t TIMESTAMP ] [-c] <path> ...]
        [-touchz <path> ...]
        [-truncate [-w] <length> <path> ...]
        [-usage [cmd ...]]
```

11:46 AM
2021-10-11

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Changing the bikes.csv owner to hive using chown command.

Ln 18, Col 15 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] [-e] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-v] [-x] <path> ...]
        [-expunge]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-head <file>]
        [-help [cmd ...]]
        [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touch [-a] [-m] [-t TIMESTAMP ] [-c] <path> ...]
        [-touchz <path> ...]
        [-truncate [-w] <length> <path> ...]
        [-usage [cmd ...]]

Generic options supported are:
-conf <configuration file>        specify an application configuration file
-D <property=value>               define a value for a given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurati
ons.
-jt <local|resourcemanager:port>  specify a ResourceManager
-files <file1,...>                specify a comma-separated list of files to be copied to the map reduce cluster
-libjars <jar1,...>               specify a comma-separated list of jar files to be included in the classpath
-archives <archive1,...>          specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]

Usage: hadoop fs [generic options] -chown [-R] [OWNER][:[GROUP]] PATH...
[hdfs@sandbox-hdp root]$ hdfs dfs -chown hive:hdfs /user/hive/bikes.csv
[hdfs@sandbox-hdp root]$
```

11:47 AM
2021-10-11

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
creating hive managed internal table with no partitions and loading bikes.csv file into the table bikes_int

Ln 21, Col 1 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> create table if not exists bikes_int(model_name string, model_year smallint, kms_driven string, owner string
, location string, milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\054' STORED AS TEXTFILE tblproperties("skip.header.l
ine.count"="1");
INFO  : Compiling command(queryId=hive_20211007004501_e4105935-f3c2-45c3-88a9-ce59c3003302): create table if not exists bikes_int(model_name string, model_
year smallint, kms_driven string, owner string, location string, milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\054' S
TORED AS TEXTFILE tblproperties("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007004501_e4105935-f3c2-45c3-88a9-ce59c3003302); Time taken: 0.037 seconds
INFO  : Executing command(queryId=hive_20211007004501_e4105935-f3c2-45c3-88a9-ce59c3003302): create table if not exists bikes_int(model_name string, model_
year smallint, kms_driven string, owner string, location string, milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\054' S
TORED AS TEXTFILE tblproperties("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007004501_e4105935-f3c2-45c3-88a9-ce59c3003302); Time taken: 0.102 seconds
INFO  : OK
No rows affected (0.185 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> LOAD DATA INPATH '/user/hive/bikes.csv' OVERWRITE INTO table bikes_int;
INFO  : Compiling command(queryId=hive_20211007004507_ae3a07b1-b313-4090-8215-734a3efbalae): LOAD DATA INPATH '/user/hive/bikes.csv' OVERWRITE INTO table b
ikes_int
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007004507_ae3a07b1-b313-4090-8215-734a3efbalae); Time taken: 0.072 seconds
INFO  : Executing command(queryId=hive_20211007004507_ae3a07b1-b313-4090-8215-734a3efbalae): LOAD DATA INPATH '/user/hive/bikes.csv' OVERWRITE INTO table b
ikes_int
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table dts.bikes_int from hdfs://sandbox-hdp.hortonworks.com:8020/user/hive/bikes.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211007004507_ae3a07b1-b313-4090-8215-734a3efbalae); Time taken: 0.508 seconds
INFO  : OK
No rows affected (0.641 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_int limit 3;
INFO  : Compiling command(queryId=hive_20211007004548_f4ebdb0d-c305-44ab-9a27-e19ac913e933): select * from bikes_int limit 3
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_int.model_name, type:string, comment:null), FieldSchema(name:bikes_int.model_yea
r, type:smallint, comment:null), FieldSchema(name:bikes_int.kms_driven, type:string, comment:null), FieldSchema(name:bikes_int.owner, type:string, comment:
null), FieldSchema(name:bikes_int.location, type:string, comment:null), FieldSchema(name:bikes_int.milage, type:string, comment:null), FieldSchema(name:bik
es_int.power, type:string, comment:null), FieldSchema(name:bikes_int.price, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007004548_f4ebdb0d-c305-44ab-9a27-e19ac913e933); Time taken: 0.36 seconds
INFO  : Executing command(queryId=hive_20211007004548_f4ebdb0d-c305-44ab-9a27-e19ac913e933): select * from bikes_int limit 3
INFO  : Completed executing command(queryId=hive_20211007004548_f4ebdb0d-c305-44ab-9a27-e19ac913e933); Time taken: 0.008 seconds
INFO  : OK
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-------------
------+--------------------+
|      bikes_int.model_name     | bikes_int.model_year  | bikes_int.kms_driven  | bikes_int.owner   | bikes_int.location    | bikes_int.milage      | bikes_int.p
ower  | bikes_int.price  |
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-------------
------+--------------------+
| Bajaj Avenger Cruise 220 2017  | 2017                  | 17000 Km              | first owner       | hyderabad             | "                     | NULL
      | NULL             |
|                               | NULL                  | NULL                  | NULL              | NULL                  | NULL                  | NULL
      | NULL             |
|  35 kmpl"                     | NULL                  | 63500                 | NULL              | NULL                  | NULL                  | NULL
      | NULL             |
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-------------
------+--------------------+
3 rows selected (0.471 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

8:24 PM
2021-10-12

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Since hive does not support CTAS with partitioning (as shown in example), We have created new table with partitioning to insert the data from the non partitioned table with partitions into the new internal partitioned table

Ln 21, Col 11 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> create table temp1 partitioned by (name string) as select "Anirudh" as name
;
Error: Error while compiling statement: FAILED: SemanticException [Error 10068]: CREATE-TABLE-AS-SELECT does not support p
artitioning in the target table (state=42000,code=10068)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> CREATE TABLE IF NOT EXISTS BIKES_INT_P(model_name STRING, kms_driven STRING
, owner STRING, location STRING, milage STRING, power STRING, price STRING) PARTITIONED BY (model_year SMALLINT) ROW FORMA
T DELIMITED FIELDS TERMINATED BY ',' STORED as TEXTFILE;
INFO  : Compiling command(queryId=hive_20211007045053_c6942a63-bb94-4e77-bafc-019b6eab3fa4): CREATE TABLE IF NOT EXISTS BI
KES_INT_P(model_name STRING, kms_driven STRING, owner STRING, location STRING, milage STRING, power STRING, price STRING)
PARTITIONED BY (model_year SMALLINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED as TEXTFILE
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007045053_c6942a63-bb94-4e77-bafc-019b6eab3fa4); Time taken: 0.056 s
econds
INFO  : Executing command(queryId=hive_20211007045053_c6942a63-bb94-4e77-bafc-019b6eab3fa4): CREATE TABLE IF NOT EXISTS BI
KES_INT_P(model_name STRING, kms_driven STRING, owner STRING, location STRING, milage STRING, power STRING, price STRING)
PARTITIONED BY (model_year SMALLINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED as TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007045053_c6942a63-bb94-4e77-bafc-019b6eab3fa4); Time taken: 0.085 s
econds
INFO  : OK
No rows affected (0.199 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

2:21 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Showing the partitioning columns in partitioned table using describe command. There are no partition information made available since there is no data in the table.

Ln 20, Col 25 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe bikes_int_p;
INFO  : Compiling command(queryId=hive_20211007045717_76f6c558-f649-44c5-8d6c-fae9656af808): describe bikes_int_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), Fi
eldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, comment:from des
erializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007045717_76f6c558-f649-44c5-8d6c-fae9656af808); Time taken: 0.044 s
econds
INFO  : Executing command(queryId=hive_20211007045717_76f6c558-f649-44c5-8d6c-fae9656af808): describe bikes_int_p
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007045717_76f6c558-f649-44c5-8d6c-fae9656af808); Time taken: 0.038 s
econds
INFO  : OK
+--------------------------+------------+----------+
|         col_name         | data_type  | comment  |
+--------------------------+------------+----------+
| model_name               | string     |          |
| kms_driven               | string     |          |
| owner                    | string     |          |
| location                 | string     |          |
| milage                   | string     |          |
| power                    | string     |          |
| price                    | string     |          |
| model_year               | smallint   |          |
|                          | NULL       | NULL     |
| # Partition Information  | NULL       | NULL     |
| # col_name               | data_type  | comment  |
| model_year               | smallint   |          |
+--------------------------+------------+----------+
12 rows selected (0.127 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show partitions bikes_int_p;
INFO  : Compiling command(queryId=hive_20211007045732_ea450f86-9340-4ec6-977f-013d792cbdff): show partitions bikes_int_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:string, comment:from deserializer)],
properties:null)
INFO  : Completed compiling command(queryId=hive_20211007045732_ea450f86-9340-4ec6-977f-013d792cbdff); Time taken: 0.062 s
econds
INFO  : Executing command(queryId=hive_20211007045732_ea450f86-9340-4ec6-977f-013d792cbdff): show partitions bikes_int_p
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007045732_ea450f86-9340-4ec6-977f-013d792cbdff); Time taken: 0.03 se
conds
INFO  : OK
+------------+
| partition  |
+------------+
+------------+
No rows selected (0.108 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

2:30 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Inserting the data from non partitioned table into partitioned table with partitioning.

Ln 18, Col 42 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> insert into bikes_int_p select * from (select model_name, kms_driven, owner
, location, milage, power, price, model_year from bikes int) as tmp;
INFO  : Compiling command(queryId=hive_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b): insert into bikes_int_p selec
t * from (select model_name, kms_driven, owner, location, milage, power, price, model_year from bikes_int) as tmp
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_col0, type:string, comment:null), FieldSchema(name:_
col1, type:string, comment:null), FieldSchema(name:_col2, type:string, comment:null), FieldSchema(name:_col3, type:string,
 comment:null), FieldSchema(name:_col4, type:string, comment:null), FieldSchema(name:_col5, type:string, comment:null), Fi
eldSchema(name:_col6, type:string, comment:null), FieldSchema(name:_col7, type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b); Time taken: 0.602 s
econds
INFO  : Executing command(queryId=hive_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b): insert into bikes_int_p selec
t * from (select model_name, kms_driven, owner, location, milage, power, price, model_year from bikes_int) as tmp
INFO  : Query ID = hive_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b
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_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: insert into bikes_int_p select * from ...tmp (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1633487809125_0026)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 59.54 s
----------------------------------------------------------------------------------------------
INFO  : Status: DAG finished successfully in 47.20 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                          0.60s
INFO  : Prepare Plan                           7.90s
INFO  : Get Query Coordinator (AM)             0.00s
INFO  : Submit Plan                            0.56s
INFO  : Start DAG                              1.58s
INFO  : Run DAG                               47.20s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)    CPU_TIME(ms)    GC_TIME(ms)    INPUT_RECORDS    OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :      Map 1          21830.00          16,270            250           23,549                38
INFO  :  Reducer 2          22204.00          13,350            802               38                 0
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
```

2:36 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Inserting the data from non partitioned table into partitioned table with partitioning.

Ln 18, Col 42 | 150% | Windows (CRLF) | UTF-8

hdfs@sandbox-hdp:/root

```
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)    CPU_TIME(ms)    GC_TIME(ms)    INPUT_RECORDS    OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :      Map 1          21830.00        16,270            250          23,549                38
INFO  :  Reducer 2          22204.00        13,350            802              38                 0
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
INFO  :    NUM_SUCCEEDED_TASKS: 3
INFO  :    TOTAL_LAUNCHED_TASKS: 3
INFO  :    AM_CPU_MILLISECONDS: 8850
INFO  :    AM_GC_TIME_MILLIS: 29
INFO  : File System Counters:
INFO  :    FILE_BYTES_READ: 24255
INFO  :    FILE_BYTES_WRITTEN: 17003
INFO  :    HDFS_BYTES_READ: 727406
INFO  :    HDFS_BYTES_WRITTEN: 1003136
INFO  :    HDFS_READ_OPS: 160
INFO  :    HDFS_WRITE_OPS: 120
INFO  :    HDFS_OP_CREATE: 42
INFO  :    HDFS_OP_GET_FILE_STATUS: 159
INFO  :    HDFS_OP_MKDIRS: 76
INFO  :    HDFS_OP_OPEN: 1
INFO  :    HDFS_OP_RENAME: 2
INFO  : org.apache.tez.common.counters.TaskCounter:
INFO  :    REDUCE_INPUT_GROUPS: 38
INFO  :    REDUCE_INPUT_RECORDS: 38
INFO  :    COMBINE_INPUT_RECORDS: 0
INFO  :    SPILLED_RECORDS: 76
INFO  :    NUM_SHUFFLED_INPUTS: 2
INFO  :    NUM_SKIPPED_INPUTS: 0
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    MERGED_MAP_OUTPUTS: 2
INFO  :    GC_TIME_MILLIS: 1052
INFO  :    TASK_DURATION_MILLIS: 38840
INFO  :    CPU_MILLISECONDS: 29620
INFO  :    PHYSICAL_MEMORY_BYTES: 1901068288
INFO  :    VIRTUAL_MEMORY_BYTES: 8049266688
INFO  :    COMMITTED_HEAP_BYTES: 1901068288
INFO  :    INPUT_RECORDS_PROCESSED: 23549
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 727406
INFO  :    OUTPUT_RECORDS: 38
INFO  :    OUTPUT_LARGE_RECORDS: 0
INFO  :    OUTPUT_BYTES: 34945
INFO  :    OUTPUT_BYTES_WITH_OVERHEAD: 35094
INFO  :    OUTPUT_BYTES_PHYSICAL: 16947
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 16947
INFO  :    ADDITIONAL_SPILL_COUNT: 0
INFO  :    SHUFFLE_CHUNK_COUNT: 1
```

2:40 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Inserting the data from non partitioned table into partitioned table with partitioning.

Ln 18, Col 42 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    ADDITIONAL_SPILL_COUNT: 0
INFO  :    OUTPUT_BYTES: 34945
INFO  :    OUTPUT_BYTES_PHYSICAL: 16947
INFO  :    OUTPUT_BYTES_WITH_OVERHEAD: 35094
INFO  :    OUTPUT_LARGE_RECORDS: 0
INFO  :    OUTPUT_RECORDS: 38
INFO  :    SHUFFLE_CHUNK_COUNT: 1
INFO  :    SPILLED_RECORDS: 38
INFO  : TaskCounter_Reducer_2_INPUT_Map_1:
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 16947
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    COMBINE_INPUT_RECORDS: 0
INFO  :    FIRST_EVENT_RECEIVED: 1012
INFO  :    LAST_EVENT_RECEIVED: 1012
INFO  :    MERGED_MAP_OUTPUTS: 2
INFO  :    MERGE_PHASE_TIME: 2096
INFO  :    NUM_DISK_TO_DISK_MERGES: 0
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    NUM_MEM_TO_DISK_MERGES: 0
INFO  :    NUM_SHUFFLED_INPUTS: 2
INFO  :    NUM_SKIPPED_INPUTS: 0
INFO  :    REDUCE_INPUT_GROUPS: 38
INFO  :    REDUCE_INPUT_RECORDS: 38
INFO  :    SHUFFLE_BYTES: 16947
INFO  :    SHUFFLE_BYTES_DECOMPRESSED: 35094
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 16947
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_PHASE_TIME: 1454
INFO  :    SPILLED_RECORDS: 38
INFO  : TaskCounter_Reducer_2_OUTPUT_out_Reducer_2:
INFO  :    OUTPUT_RECORDS: 0
INFO  : org.apache.hadoop.hive.ql.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  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table dts.bikes_int_p partition (model_year=null) from hdfs://sandbox-hdp.hortonworks.com:8020/war
ehouse/tablespace/managed/hive/dts.db/bikes_int_p
INFO  :

INFO  :          Time taken to load dynamic partitions: 6.654 seconds
INFO  :          Time taken for adding to write entity : 0.004 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211007050654_2181dd1a-5a43-4942-9779-d3a6c99b1e0b); Time taken: 69.514
seconds
INFO  : OK
No rows affected (70.207 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

2:40 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Partitioning on years is considered good as there will be limited unique years and hence creating partitions on model_year is a good choice.

The screenshot shows the paritition information on the table.

Ln 22, Col 14 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show partitions bikes_int_p;
INFO  : Compiling command(queryId=hive_20211007051349_c4f26786-bfcc-45a4-8730-98e048c5bfd7): show partitions bikes_int_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:string, comment:from deserializer)],
properties:null)
INFO  : Completed compiling command(queryId=hive_20211007051349_c4f26786-bfcc-45a4-8730-98e048c5bfd7); Time taken: 0.183 s
econds
INFO  : Executing command(queryId=hive_20211007051349_c4f26786-bfcc-45a4-8730-98e048c5bfd7): show partitions bikes_int_p
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007051349_c4f26786-bfcc-45a4-8730-98e048c5bfd7); Time taken: 0.046 s
econds
INFO  : OK
+-----------------------------------------+
|               partition                 |
+-----------------------------------------+
| model_year=1950                         |
| model_year=1970                         |
| model_year=1978                         |
| model_year=1982                         |
| model_year=1985                         |
| model_year=1986                         |
| model_year=1990                         |
| model_year=1991                         |
| model_year=1993                         |
| model_year=1994                         |
| model_year=1996                         |
| model_year=1997                         |
| model_year=1998                         |
| model_year=1999                         |
| model_year=2000                         |
| model_year=2001                         |
| model_year=2002                         |
| model_year=2003                         |
| model_year=2004                         |
| model_year=2005                         |
| model_year=2006                         |
| model_year=2007                         |
| model_year=2008                         |
| model_year=2009                         |
| model_year=2010                         |
| model_year=2011                         |
| model_year=2012                         |
| model_year=2013                         |
| model_year=2014                         |
| model_year=2015                         |
| model_year=2016                         |
| model_year=2017                         |
| model_year=2018                         |
| model_year=2019                         |
| model_year=2020                         |
| model_year=2021                         |
| model_year=7                            |
| model_year=__HIVE_DEFAULT_PARTITION__   |
+-----------------------------------------+
38 rows selected (0.259 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

2:45 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
looking at data fetch time for non partitioned table.

Ln 19, Col 1 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_int where model_year = 2018;
```

3:29 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
looking at data fetch time for non partitioned table when running the command mentioned in "()".

Ln 19, Col 1 150% Windows (CRLF) UTF-8

hdfs@sandbox-hdp:/root

```
| UM Renegade Commando 2017                          | 2017                 | 3000 Km              | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Bajaj V15 150cc 2017                               | 2017                 | 2525 Km              | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Bajaj Pulsar 180cc 2017                            | 2017                 | 15000 Km             | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Yamaha YZF-R15 2.0 150cc 2017                      | 2017                 | 2500 Km              | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Yamaha YZF-R15 150cc 2017                          | 2017                 | 9000 Km              | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| TVS Apache RTR 160cc 2017                          | 2017                 | Mileage 60 Kmpl      | first owner        |
patna                | "                    | NULL                 | NULL                 |
| Yamaha FZs 150cc 2017                              | 2017                 | 400 Km               | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2017                   | 2017                 | 15000 Km             | second owner       |
delhi                | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2017                   | 2017                 | 7193 Km              | first owner        |
faridabad            | "                    | NULL                 | NULL                 |
| Bajaj Avenger Street 150 2017                      | 2017                 | 7000 Km              | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Honda CB Shine 125cc 2017                          | 2017                 | Mileage 65 Kmpl      | first owner        |
ahmedabad            | "                    | NULL                 | NULL                 |
| Bajaj Pulsar 150cc 2017                            | 2017                 | Mileage 65 Kmpl      | first owner        |
ahmedabad            | "                    | NULL                 | NULL                 |
| Bajaj Pulsar RS200 2017                            | 2017                 | Mileage 35 Kmpl      | first owner        |
siliguri             | "                    | NULL                 | NULL                 |
| Bajaj CT 100 100cc 2017                            | 2017                 | 31968 Km             | first owner        |
gurgaon              | "                    | NULL                 | NULL                 |
| Bajaj Pulsar 220cc 2017                            | 2017                 | 16000 Km             | first owner        |
bangalore            | "                    | NULL                 | NULL                 |
| Bajaj Pulsar RS200 2017                            | 2017                 | Mileage 35 Kmpl      | first owner        |
jabalpur             | "                    | NULL                 | NULL                 |
| Hero Passion Pro 100cc 2017                        | 2017                 | 3000 Km              | first owner        |
tiruvallur           | "                    | NULL                 | NULL                 |
| Hero Passion Pro 100cc 2017                        | 2017                 | 22000 Km             | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Bajaj V12 125cc 2017                               | 2017                 | 15621 Km             | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Bajaj Dominar 400 2017                             | 2017                 | Mileage 28 Kms       | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Hero CD Deluxe 100cc 2017                          | 2017                 | 28000 Km             | first owner        |
delhi                | "                    | NULL                 | NULL                 |
| Bajaj Dominar 400 2017                             | 2017                 | Mileage 28 Kms       | first owner        |
rohtak               | "                    | NULL                 | NULL                 |
| Bajaj Avenger Cruise 220 2017                      | 2017                 | 5800 Km              | first owner        |
jalandhar            | "                    | NULL                 | NULL                 |
| Hero Splendor iSmart 110cc 2017                    | 2017                 | 3000 Km              | first owner        |
mumbai               | "                    | NULL                 | NULL                 |
| Honda CB Shine 125cc 2017                          | 2017                 | Mileage 65 Kmpl      | first owner        |
mumbai               | "                    | NULL                 | NULL                 |
| TVS Apache RTR 180cc 2017                          | 2017                 | Mileage 45 Kmpl      | first owner        |
gurgaon              | "                    | NULL                 | NULL                 |
+----------------------------------------------------+----------------------+----------------------+--------------------
---+----------------------+------------------+------------------+------------------+
1,160 rows selected (0.624 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes int where model year = 2017;
```

3:32 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
looking at data fetch time for partitioned table when running the command mentioned in "()".

We observe that records are captured few milliseconds faster than it did when querying the non partitioned table.
The time difference will only increase if the dataset size keeps on increasing.

If the data is accessed frequently using a certain column we should consider partitioning as it a huge performace and resouce booster.

Ln 24, Col 43 | 150% | Windows (CRLF) | UTF-8

hdfs@sandbox-hdp:/root

```
| UM Renegade Commando 2017                 | 3000 Km           | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Bajaj V15 150cc 2017                      | 2525 Km           | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Bajaj Pulsar 180cc 2017                   | 15000 Km          | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Yamaha YZF-R15 2.0 150cc 2017             | 2500 Km           | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Yamaha YZF-R15 150cc 2017                 | 9000 Km           | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| TVS Apache RTR 160cc 2017                 | Mileage 60 Kmpl   | first owner    | patna
  | "              | NULL          | NULL          | 2017           |
| Yamaha FZs 150cc 2017                     | 400 Km            | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Royal Enfield Classic 350cc 2017          | 15000 Km          | second owner   | delhi
  | "              | NULL          | NULL          | 2017           |
| Royal Enfield Classic 350cc 2017          | 7193 Km           | first owner    | faridabad
  | "              | NULL          | NULL          | 2017           |
| Bajaj Avenger Street 150 2017             | 7000 Km           | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Honda CB Shine 125cc 2017                 | Mileage 65 Kmpl   | first owner    | ahmedabad
  | "              | NULL          | NULL          | 2017           |
| Bajaj Pulsar 150cc 2017                   | Mileage 65 Kmpl   | first owner    | ahmedabad
  | "              | NULL          | NULL          | 2017           |
| Bajaj Pulsar RS200 2017                   | Mileage 35 Kmpl   | first owner    | siliguri
  | "              | NULL          | NULL          | 2017           |
| Bajaj CT 100 100cc 2017                   | 31968 Km          | first owner    | gurgaon
  | "              | NULL          | NULL          | 2017           |
| Bajaj Pulsar 220cc 2017                   | 16000 Km          | first owner    | bangalore
  | "              | NULL          | NULL          | 2017           |
| Bajaj Pulsar RS200 2017                   | Mileage 35 Kmpl   | first owner    | jabalpur
  | "              | NULL          | NULL          | 2017           |
| Hero Passion Pro 100cc 2017               | 3000 Km           | first owner    | tiruvallur
  | "              | NULL          | NULL          | 2017           |
| Hero Passion Pro 100cc 2017               | 22000 Km          | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Bajaj V12 125cc 2017                      | 15621 Km          | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Bajaj Dominar 400 2017                    | Mileage 28 Kms    | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Hero CD Deluxe 100cc 2017                 | 28000 Km          | first owner    | delhi
  | "              | NULL          | NULL          | 2017           |
| Bajaj Dominar 400 2017                    | Mileage 28 Kms    | first owner    | rohtak
  | "              | NULL          | NULL          | 2017           |
| Bajaj Avenger Cruise 220 2017             | 5800 Km           | first owner    | jalandhar
  | "              | NULL          | NULL          | 2017           |
| Hero Splendor iSmart 110cc 2017           | 3000 Km           | first owner    | mumbai
  | "              | NULL          | NULL          | 2017           |
| Honda CB Shine 125cc 2017                 | Mileage 65 Kmpl   | first owner    | mumbai
  | "              | NULL          | NULL          | 2017           |
| TVS Apache RTR 180cc 2017                 | Mileage 45 Kmpl   | first owner    | gurgaon
  | "              | NULL          | NULL          | 2017           |
+-------------------------------------------+-------------------+----------------+------------------
-----+-------------------+-------------------+-------------------+-----------------------+
1,160 rows selected (0.543 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_int_p where model_year = 2017;
```

3:37 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Creating a directory external in /user/hive and giving Full access with chmod 777 to the directory copying the csv files to the external directory and also giving it 777 access along with changing the owner to hive.

Ln 21, Col 19 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
[hive@sandbox-hdp root]$ hdfs dfs -mkdir /user/hive/external
[hive@sandbox-hdp root]$ hdfs dfs -chmod 777 /user/hive/external
[hive@sandbox-hdp root]$ hdfs dfs -copyFromLocal /bikes.csv /user/hive/external/
[hive@sandbox-hdp root]$ hdfs dfs -chmod 777 /user/hive/external/bikes.csv
hd[hive@sandbox-hdp root]$ hdfs dfs -chown hive:hdfs /user/hive/external/bikes.csv
[hive@sandbox-hdp root]$ hive
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
21/10/07 06:31:10 [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>
Closing: 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=hive;serviceDiscoveryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2
[hive@sandbox-hdp root]$ hive
```

4:00 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
1. Creating an External table with no partitions in hive from csv placed in "/user/hive/external/" directory and not including the csv header using TBLPROPERTIES.

2. Doing a select * on the external table to confirm the data coming up.

Ln 23, Col 1 | 150% | Windows (CRLF) | UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> CREATE EXTERNAL TABLE IF NOT EXISTS BIKES_EXT(model_name string,model_year string, kms_driven string, owner string, location string,milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TBLPROPERTIES("skip.header.line.count"="1");
INFO  : Compiling command(queryId=hive_20211007064431_4279e575-feec-4e79-92c3-387ba57a1335): CREATE EXTERNAL TABLE IF NOT EXISTS BIKES_EXT(model_name string,model_year string, kms_driven string, owner string, location string,milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007064431_4279e575-feec-4e79-92c3-387ba57a1335); Time taken: 0.043 seconds
INFO  : Executing command(queryId=hive_20211007064431_4279e575-feec-4e79-92c3-387ba57a1335): CREATE EXTERNAL TABLE IF NOT EXISTS BIKES_EXT(model_name string,model_year string, kms_driven string, owner string, location string,milage string, power string, price int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007064431_4279e575-feec-4e79-92c3-387ba57a1335); Time taken: 0.114 seconds
INFO  : OK
No rows affected (0.198 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext limit 2;
INFO  : Compiling command(queryId=hive_20211007064444_209e73eb-dea0-4925-a274-48cff45e53a0): select * from bikes_ext limit 2
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext.model_name, type:string, comment:null), FieldSchema(name:bikes_ext.model_year, type:string, comment:null), FieldSchema(name:bikes_ext.kms_driven, type:string, comment:null), FieldSchema(name:bikes_ext.owner, type:string, comment:null), FieldSchema(name:bikes_ext.location, type:string, comment:null), FieldSchema(name:bikes_ext.milage, type:string, comment:null), FieldSchema(name:bikes_ext.power, type:string, comment:null), FieldSchema(name:bikes_ext.price, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007064444_209e73eb-dea0-4925-a274-48cff45e53a0); Time taken: 0.188 seconds
INFO  : Executing command(queryId=hive_20211007064444_209e73eb-dea0-4925-a274-48cff45e53a0): select * from bikes_ext limit 2
INFO  : Completed executing command(queryId=hive_20211007064444_209e73eb-dea0-4925-a274-48cff45e53a0); Time taken: 0.006 seconds
INFO  : OK
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-----------------------+
|      bikes_ext.model_name     | bikes_ext.model_year  | bikes_ext.kms_driven  | bikes_ext.owner   | bikes_ext.location    | bikes_ext.milage      | bikes_ext.power       | bikes_ext.price  |
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-----------------------+------------------+
| Bajaj Avenger Cruise 220 2017  | 2017                  | 17000 Km              | first owner       | hyderabad             | "                     | NULL                  | NULL             |
| NULL                          | NULL                  | NULL                  | NULL              | NULL                  | NULL                  | NULL                  | NULL             |
+-------------------------------+-----------------------+-----------------------+-------------------+-----------------------+-----------------------+-----------------------+------------------+
2 rows selected (0.239 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

4:13 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
There are multiple ways of creating a partitioned external table in hive. One of which is shown in the screenshots.

We have created the External table with partition but the table does not have any data yet and we have to insert data into the external table

Ln 24, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> CREATE EXTERNAL TABLE IF NOT EXISTS BIKES_EXT_P(model_name string, kms_driv
en string, owner string, location string,milage string, power string, price int) PARTITIONED BY (model_year smallint) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TBLPROPERTIES("skip.header.line.count"="1");
INFO  : Compiling command(queryId=hive_20211007070429_6d9b7c3d-065d-41f7-b95b-e91528f7da1e): CREATE EXTERNAL TABLE IF NOT
EXISTS BIKES_EXT_P(model_name string, kms_driven string, owner string, location string,milage string, power string, price
int) PARTITIONED BY (model_year smallint) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TB
LPROPERTIES("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007070429_6d9b7c3d-065d-41f7-b95b-e91528f7da1e); Time taken: 0.055 s
econds
INFO  : Executing command(queryId=hive_20211007070429_6d9b7c3d-065d-41f7-b95b-e91528f7da1e): CREATE EXTERNAL TABLE IF NOT
EXISTS BIKES_EXT_P(model_name string, kms_driven string, owner string, location string,milage string, power string, price
int) PARTITIONED BY (model_year smallint) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TB
LPROPERTIES("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007070429_6d9b7c3d-065d-41f7-b95b-e91528f7da1e); Time taken: 0.089 s
econds
INFO  : OK
No rows affected (0.197 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext_p limit 2;
INFO  : Compiling command(queryId=hive_20211007070838_44eb9896-8bda-464d-9345-0b057ea444b2): select * from bikes_ext_p lim
it 2
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext_p.model_name, type:string, comment:null), F
ieldSchema(name:bikes_ext_p.kms_driven, type:string, comment:null), FieldSchema(name:bikes_ext_p.owner, type:string, comme
nt:null), FieldSchema(name:bikes_ext_p.location, type:string, comment:null), FieldSchema(name:bikes_ext_p.milage, type:str
ing, comment:null), FieldSchema(name:bikes_ext_p.power, type:string, comment:null), FieldSchema(name:bikes_ext_p.price, ty
pe:int, comment:null), FieldSchema(name:bikes_ext_p.model_year, type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007070838_44eb9896-8bda-464d-9345-0b057ea444b2); Time taken: 0.193 s
econds
INFO  : Executing command(queryId=hive_20211007070838_44eb9896-8bda-464d-9345-0b057ea444b2): select * from bikes_ext_p lim
it 2
INFO  : Completed executing command(queryId=hive_20211007070838_44eb9896-8bda-464d-9345-0b057ea444b2); Time taken: 0.005 s
econds
INFO  : OK
+-------------------------+-------------------------+--------------------+-----------------------+-----------------------+--
-------------------+--------------------+-------------------------+
| bikes_ext_p.model_name  | bikes_ext_p.kms_driven  | bikes_ext_p.owner  | bikes_ext_p.location  | bikes_ext_p.milage  | b
ikes_ext_p.power  | bikes_ext_p.price  | bikes_ext_p.model_year  |
+-------------------------+-------------------------+--------------------+-----------------------+-----------------------+--
-------------------+--------------------+-------------------------+
+-------------------------+-------------------------+--------------------+-----------------------+-----------------------+--
-------------------+--------------------+-------------------------+
No rows selected (0.224 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

4:37 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Manually adding partitions to the external table.

Ln 17, Col 50 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2017) LOCATION '/user/hive/external';
INFO  : Compiling command(queryId=hive_20211007071628_805bb582-9fdf-4bab-ba34-29bd7df2473e): ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2017) LOCATION '/user/hive/external'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071628_805bb582-9fdf-4bab-ba34-29bd7df2473e); Time taken: 0.079 seconds
INFO  : Executing command(queryId=hive_20211007071628_805bb582-9fdf-4bab-ba34-29bd7df2473e): ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2017) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071628_805bb582-9fdf-4bab-ba34-29bd7df2473e); Time taken: 0.238 seconds
INFO  : OK
No rows affected (0.361 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext_p limit 2;
INFO  : Compiling command(queryId=hive_20211007071640_e7b6e168-d027-4d6f-a8a8-9dee573258c7): select * from bikes_ext_p limit 2
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext_p.model_name, type:string, comment:null), FieldSchema(name:bikes_ext_p.kms_driven, type:string, comment:null), FieldSchema(name:bikes_ext_p.owner, type:string, comment:null), FieldSchema(name:bikes_ext_p.location, type:string, comment:null), FieldSchema(name:bikes_ext_p.milage, type:string, comment:null), FieldSchema(name:bikes_ext_p.power, type:string, comment:null), FieldSchema(name:bikes_ext_p.price, type:int, comment:null), FieldSchema(name:bikes_ext_p.model_year, type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071640_e7b6e168-d027-4d6f-a8a8-9dee573258c7); Time taken: 0.211 seconds
INFO  : Executing command(queryId=hive_20211007071640_e7b6e168-d027-4d6f-a8a8-9dee573258c7): select * from bikes_ext_p limit 2
INFO  : Completed executing command(queryId=hive_20211007071640_e7b6e168-d027-4d6f-a8a8-9dee573258c7); Time taken: 0.006 seconds
INFO  : OK
+-------------------------------+-------------------------+--------------------+-------------------------+-------------------+---------------------+--+
|     bikes_ext_p.model_name    | bikes_ext_p.kms_driven  | bikes_ext_p.owner  | bikes_ext_p.location    | bikes_ext_p.milage | bikes_ext_p.power  | bikes_ext_p.price  | bikes_ext_p.model_year  |
+-------------------------------+-------------------------+--------------------+-------------------------+-------------------+---------------------+--+
| Bajaj Avenger Cruise 220 2017 | 2017                    | 17000 Km           | first owner             | hyderabad         | "                   | NULL               | 2017                    |
| NULL                          | NULL                    | NULL               | NULL                    | NULL              | NULL                | NULL               | 2017                    |
+-------------------------------+-------------------------+--------------------+-------------------------+-------------------+---------------------+--+
2 rows selected (0.257 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2018) LOCATION '/user/hive/external';
INFO  : Compiling command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15): ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2018) LOCATION '/user/hive/external'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15); Time taken: 0.06 seconds
INFO  : Executing command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15): ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2018) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15); Time taken: 0.092 s
```

4:48 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Manually adding partitions to the external table.

Ln 17, Col 50 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15); Time taken: 0.06 se
conds
INFO  : Executing command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2018) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071829_b5858904-1392-40aa-baed-91e24155ef15); Time taken: 0.092 s
econds
INFO  : OK
No rows affected (0.175 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2019) LOCATION '/user/hiv
e/external';
INFO  : Compiling command(queryId=hive_20211007071836_a1e5707d-b927-4d2e-883f-f712f852ecc7): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2019) LOCATION '/user/hive/external'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071836_a1e5707d-b927-4d2e-883f-f712f852ecc7); Time taken: 0.084 s
econds
INFO  : Executing command(queryId=hive_20211007071836_a1e5707d-b927-4d2e-883f-f712f852ecc7): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2019) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071836_a1e5707d-b927-4d2e-883f-f712f852ecc7); Time taken: 0.079 s
econds
INFO  : OK
No rows affected (0.189 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2020) LOCATION '/user/hiv
e/external';
INFO  : Compiling command(queryId=hive_20211007071844_42bcbeb4-eb84-487b-8a49-33d6afc937f7): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2020) LOCATION '/user/hive/external'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071844_42bcbeb4-eb84-487b-8a49-33d6afc937f7); Time taken: 0.056 s
econds
INFO  : Executing command(queryId=hive_20211007071844_42bcbeb4-eb84-487b-8a49-33d6afc937f7): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2020) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071844_42bcbeb4-eb84-487b-8a49-33d6afc937f7); Time taken: 0.078 s
econds
INFO  : OK
No rows affected (0.16 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ALTER TABLE bikes_ext_p ADD PARTITION (model_year=2016) LOCATION '/user/hiv
e/external';
INFO  : Compiling command(queryId=hive_20211007071903_aa52046c-e017-4ac1-9b8f-a78ad580b1fc): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2016) LOCATION '/user/hive/external'
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007071903_aa52046c-e017-4ac1-9b8f-a78ad580b1fc); Time taken: 0.075 s
econds
INFO  : Executing command(queryId=hive_20211007071903_aa52046c-e017-4ac1-9b8f-a78ad580b1fc): ALTER TABLE bikes_ext_p ADD P
ARTITION (model_year=2016) LOCATION '/user/hive/external'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007071903_aa52046c-e017-4ac1-9b8f-a78ad580b1fc); Time taken: 0.068 s
econds
INFO  : OK
No rows affected (0.169 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

4:49 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Table structure of two external tables created.

Ln 17, Col 48 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe bikes_ext;
INFO  : Compiling command(queryId=hive_20211007074351_90a608e8-abf3-49b4-bbc3-23d51d815bda): describe bikes_ext
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), Fi
eldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, comment:from des
erializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007074351_90a608e8-abf3-49b4-bbc3-23d51d815bda); Time taken: 0.046 s
econds
INFO  : Executing command(queryId=hive_20211007074351_90a608e8-abf3-49b4-bbc3-23d51d815bda): describe bikes_ext
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007074351_90a608e8-abf3-49b4-bbc3-23d51d815bda); Time taken: 0.015 s
econds
INFO  : OK
+-------------+------------+----------+
|  col_name   | data_type  | comment  |
+-------------+------------+----------+
| model_name  | string     |          |
| model_year  | string     |          |
| kms_driven  | string     |          |
| owner       | string     |          |
| location    | string     |          |
| milage      | string     |          |
| power       | string     |          |
| price       | int        |          |
+-------------+------------+----------+
8 rows selected (0.1 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe bikes_ext_p;
INFO  : Compiling command(queryId=hive_20211007074401_f38dc08d-bb49-489e-a545-aadac5054904): describe bikes_ext_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), Fi
eldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, comment:from des
erializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007074401_f38dc08d-bb49-489e-a545-aadac5054904); Time taken: 0.054 s
econds
INFO  : Executing command(queryId=hive_20211007074401_f38dc08d-bb49-489e-a545-aadac5054904): describe bikes_ext_p
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007074401_f38dc08d-bb49-489e-a545-aadac5054904); Time taken: 0.094 s
econds
INFO  : OK
+--------------------------+------------+----------+
|         col_name         | data_type  | comment  |
+--------------------------+------------+----------+
| model_name               | string     |          |
| kms_driven               | string     |          |
| owner                    | string     |          |
| location                 | string     |          |
| milage                   | string     |          |
| power                    | string     |          |
| price                    | int        |          |
| model_year               | smallint   |          |
|                          | NULL       | NULL     |
| # Partition Information  | NULL       | NULL     |
| # col_name               | data_type  | comment  |
| model_year               | smallint   |          |
+--------------------------+------------+----------+
12 rows selected (0.175 seconds)
```

5:12 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Time taken to query non partitioned external table shown in "()"
- 1,101 rows selected (0.615 seconds)

Ln 19, Col 29 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
| Royal Enfield Classic 350cc 2018                    | 2018                 | 21000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 24000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| TVS Apache RTR 200 4V Carburetor 2018               | 2018                 | Mileage 40 Kmpl        | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Hero Passion Xpro Disc 2018                         | 2018                 | 5500 Km                | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Bajaj CT 100 100cc 2018                             | 2018                 | 5661 Km                | first owner       |
bangalore           | "                    | NULL                 | NULL                 |
| Suzuki Gixxer SF 150cc ABS 2018                     | 2018                 | 9000 Km                | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Hero Passion Pro i3S Alloy 100cc 2018               | 2018                 | 8000 Km                | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Hero Super Splendor 125cc 2018                      | 2018                 | 8049 Km                | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 12552 Km               | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| TVS Apache RTR 200 4V FI 2018                       | 2018                 | Mileage 40 Kmpl        | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Yamaha YZF-R15 V3 150cc 2018                        | 2018                 | 6900 Km                | first owner       |
chennai             | "                    | NULL                 | NULL                 |
| Bajaj Pulsar 220cc 2018                             | 2018                 | 22000 Km               | first owner       |
bangalore           | "                    | NULL                 | NULL                 |
| Bajaj Platina  Alloy ES-100cc 2018                  | 2018                 | Mileage 104 Kmpl       | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| Hero HF Deluxe i3s 100cc 2018                       | 2018                 | 13000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Bajaj Dominar 400 ABS 2018                          | 2018                 | Mileage 28 Kms         | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| UM Renegade Commando Classic 2018                   | 2018                 | 2911 Km                | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| TVS Star City Plus 110cc 2018                       | 2018                 | 1363 Km                | first owner       |
chitradurga         | "                    | NULL                 | NULL                 |
| TVS Sport 100cc 2018                                | 2018                 | 8048 Km                | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| KTM Duke 390cc 2018                                 | 2018                 | 4500 Km                | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| KTM Duke 390cc 2018                                 | 2018                 | 12000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Thunderbird 350cc 2018                | 2018                 | 20000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 14354 Km               | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
+-----------------------------------------------------+----------------------+------------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
|                bikes_ext.model_name                 | bikes_ext.model_year  | bikes_ext.kms_driven  |    bikes_ext.owner
   |  bikes_ext.location  | bikes_ext.milage  | bikes_ext.power  | bikes_ext.price  |
+-----------------------------------------------------+----------------------+------------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
| TVS Apache RTR 160 4V Carburetor With Rear Disc 2018 | 2018                 | 16510 Km               | first owner
| ahmedabad           | "                    | NULL                 | NULL                 |
+-----------------------------------------------------+----------------------+------------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
1,101 rows selected (0.615 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext where model_year = 2018;
```

5:20 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Another way to partition table is enable dynamic partition flags using the following commands

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

Ln 18, Col 15 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> set hive.exec.dynamic.partition=true
. . . . . .> ;
No rows affected (0.014 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> set hive.exec.dynamic.partition.mode=nonstrict;
No rows affected (0.008 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Closing: 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=
hive;serviceDiscoveryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2
[hive@sandbox-hdp root]$ hive
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/StaticLoggerBi
nder.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/StaticLoggerBin
der.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
21/10/07 07:55:38 [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>
```

5:24 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Another way to partition table is enable dynamic partition flags using the following commands

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

Ln 18, Col 15 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> set hive.exec.dynamic.partition=true
. . . . . .> ;
No rows affected (0.014 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> set hive.exec.dynamic.partition.mode=nonstrict;
No rows affected (0.008 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Closing: 0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=
hive;serviceDiscoveryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2
[hive@sandbox-hdp root]$ hive
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/StaticLoggerBi
nder.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/StaticLoggerBin
der.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
21/10/07 07:55:38 [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> show partitions bikes_ext_p;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Table not found bikes_ext_p (state=42S02,
code=10001)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show tables;
INFO  : Compiling command(queryId=hive_20211007080200_a3fc8420-97b5-4b9f-91f8-0c89247c0a3b): show tables
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], p
roperties:null)
INFO  : Completed compiling command(queryId=hive_20211007080200_a3fc8420-97b5-4b9f-91f8-0c89247c0a3b); Time taken: 0.023 s
econds
INFO  : Executing command(queryId=hive_20211007080200_a3fc8420-97b5-4b9f-91f8-0c89247c0a3b): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007080200_a3fc8420-97b5-4b9f-91f8-0c89247c0a3b); Time taken: 0.028 s
econds
INFO  : OK
+-------------+
|  tab_name   |
+-------------+
| bikes_ext   |
+-------------+
1 row selected (0.147 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> use dts;
INFO  : Compiling command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8): use dts
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8); Time taken: 0.021 s
econds
INFO  : Executing command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8): use dts
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8); Time taken: 0.016 s
econds
INFO  : OK
```

5:41 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Manually added static partitions in external table.

Ln 18, Col 7 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
INFO  : OK
+------------+
|  tab_name  |
+------------+
| bikes_ext  |
+------------+
1 row selected (0.147 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> use dts;
INFO  : Compiling command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8): use dts
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8); Time taken: 0.021 s
econds
INFO  : Executing command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8): use dts
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007080209_2d6bae4b-0738-41f8-9a6e-34cd90fb8ba8); Time taken: 0.016 s
econds
INFO  : OK
No rows affected (0.048 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show partitions bikes_ext_p;
INFO  : Compiling command(queryId=hive_20211007080231_a621ab0a-3ca1-4267-ae72-df3db2d2311f): show partitions bikes_ext_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:string, comment:from deserializer)],
properties:null)
INFO  : Completed compiling command(queryId=hive_20211007080231_a621ab0a-3ca1-4267-ae72-df3db2d2311f); Time taken: 0.104 s
econds
INFO  : Executing command(queryId=hive_20211007080231_a621ab0a-3ca1-4267-ae72-df3db2d2311f): show partitions bikes_ext_p
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007080231_a621ab0a-3ca1-4267-ae72-df3db2d2311f); Time taken: 0.06 se
conds
INFO  : OK
+--------------------+
|     partition      |
+--------------------+
| model_year=2016    |
| model_year=2017    |
| model_year=2018    |
| model_year=2019    |
| model_year=2020    |
+--------------------+
5 rows selected (0.188 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> msck repair table bikes_ext_p;
INFO  : Compiling command(queryId=hive_20211007080439_ff16bb89-3682-4d32-846f-d61b3eddd6e6): msck repair table bikes_ext_p
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007080439_ff16bb89-3682-4d32-846f-d61b3eddd6e6); Time taken: 0.063 s
econds
INFO  : Executing command(queryId=hive_20211007080439_ff16bb89-3682-4d32-846f-d61b3eddd6e6): msck repair table bikes_ext_p
INFO  : Starting task [Stage-0:DDL] in serial mode
ERROR : FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask
INFO  : Completed executing command(queryId=hive_20211007080439_ff16bb89-3682-4d32-846f-d61b3eddd6e6); Time taken: 0.087 s
econds
Error: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTas
k (state=08S01,code=1)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Msck repair table bikes_ext_p;
INFO  : Compiling command(queryId=hive_20211007080611_eaf1d2f1-3ecb-40c2-bffa-1f105c3f5b19): Msck repair table bikes_ext_p
```

5:43 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Creating dynamically partitioned external table

Ln 18, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007080611_eaf1d2f1-3ecb-40c2-bffa-1f105c3f5b19); Time taken: 0.062 s
econds
INFO  : Executing command(queryId=hive_20211007080611_eaf1d2f1-3ecb-40c2-bffa-1f105c3f5b19): Msck repair table bikes_ext_p
INFO  : Starting task [Stage-0:DDL] in serial mode
ERROR : FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask
INFO  : Completed executing command(queryId=hive_20211007080611_eaf1d2f1-3ecb-40c2-bffa-1f105c3f5b19); Time taken: 0.054 s
econds
Error: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTas
k (state=08S01,code=1)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> set hive.msck.path.validation=ignore;
Error: Error while processing statement: Cannot modify hive.msck.path.validation at runtime. It is not in list of params t
hat are allowed to be modified at runtime (state=42000,code=1)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> CREATE EXTERNAL TABLE IF NOT EXISTS BIKES_EXT_DP(model_name string, kms_dri
ven string, owner string, location string,milage string, power string, price int) PARTITIONED BY (model_year smallint) ROW
 FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' TBLPROPERTIES("skip.header.line.count"="1");
INFO  : Compiling command(queryId=hive_20211007081017_17ccd257-af45-46ee-b92b-c522c86d44e5): CREATE EXTERNAL TABLE IF NOT
EXISTS BIKES_EXT_DP(model_name string, kms_driven string, owner string, location string,milage string, power string, price
 int) PARTITIONED BY (model_year smallint) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' T
BLPROPERTIES("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211007081017_17ccd257-af45-46ee-b92b-c522c86d44e5); Time taken: 0.088 s
econds
INFO  : Executing command(queryId=hive_20211007081017_17ccd257-af45-46ee-b92b-c522c86d44e5): CREATE EXTERNAL TABLE IF NOT
EXISTS BIKES_EXT_DP(model_name string, kms_driven string, owner string, location string,milage string, power string, price
 int) PARTITIONED BY (model_year smallint) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/hive/external/' T
BLPROPERTIES("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007081017_17ccd257-af45-46ee-b92b-c522c86d44e5); Time taken: 0.145 s
econds
INFO  : OK
No rows affected (0.316 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext_dp limit 2;
INFO  : Compiling command(queryId=hive_20211007081030_507bcbf6-ed99-468f-bc16-fd52f14468c4): select * from bikes_ext_dp li
mit 2
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext_dp.model_name, type:string, comment:null),
FieldSchema(name:bikes_ext_dp.kms_driven, type:string, comment:null), FieldSchema(name:bikes_ext_dp.owner, type:string, co
mment:null), FieldSchema(name:bikes_ext_dp.location, type:string, comment:null), FieldSchema(name:bikes_ext_dp.milage, typ
e:string, comment:null), FieldSchema(name:bikes_ext_dp.power, type:string, comment:null), FieldSchema(name:bikes_ext_dp.pr
ice, type:int, comment:null), FieldSchema(name:bikes_ext_dp.model_year, type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007081030_507bcbf6-ed99-468f-bc16-fd52f14468c4); Time taken: 0.314 s
econds
INFO  : Executing command(queryId=hive_20211007081030_507bcbf6-ed99-468f-bc16-fd52f14468c4): select * from bikes_ext_dp li
mit 2
INFO  : Completed executing command(queryId=hive_20211007081030_507bcbf6-ed99-468f-bc16-fd52f14468c4); Time taken: 0.008 s
econds
INFO  : OK
+--------------------------+--------------------------+---------------------+--------------------------+---------------------
--+---------------------+---------------------+-------------------------+
| bikes_ext_dp.model_name  | bikes_ext_dp.kms_driven  | bikes_ext_dp.owner  | bikes_ext_dp.location  | bikes_ext_dp.milage
  | bikes_ext_dp.power  | bikes_ext_dp.price  | bikes_ext_dp.model_year  |
+--------------------------+--------------------------+---------------------+--------------------------+---------------------
--+---------------------+---------------------+-------------------------+
```

5:45 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
showing partitions created for the new dynamically partitioned table and inserting the data into external table from another table.

Ln 21, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe bikes_ext_dp;
INFO  : Compiling command(queryId=hive_20211007081119_bdbaccda-0418-4105-ac16-61bc60936e7b): describe bikes_ext_dp
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), Fi
eldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, comment:from des
erializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007081119_bdbaccda-0418-4105-ac16-61bc60936e7b); Time taken: 0.064 s
econds
INFO  : Executing command(queryId=hive_20211007081119_bdbaccda-0418-4105-ac16-61bc60936e7b): describe bikes_ext_dp
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007081119_bdbaccda-0418-4105-ac16-61bc60936e7b); Time taken: 0.046 s
econds
INFO  : OK
+--------------------------+------------+----------+
|         col_name         | data_type  | comment  |
+--------------------------+------------+----------+
| model_name               | string     |          |
| kms_driven               | string     |          |
| owner                    | string     |          |
| location                 | string     |          |
| milage                   | string     |          |
| power                    | string     |          |
| price                    | int        |          |
| model_year               | smallint   |          |
|                          | NULL       | NULL     |
| # Partition Information  | NULL       | NULL     |
| # col_name               | data_type  | comment  |
| model_year               | smallint   |          |
+--------------------------+------------+----------+
12 rows selected (0.147 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> insert into bikes_ext_dp select * from (select model_name, kms_driven,owner
,location,milage,power,price,model_year from bikes_int) as tmp;
INFO  : Compiling command(queryId=hive_20211007081234_812eaeeb-8887-4a5a-9ee5-2035d32481fe): insert into bikes_ext_dp sele
ct * from (select model_name, kms_driven,owner,location,milage,power,price,model_year from bikes_int) as tmp
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tmp.model_name, type:string, comment:null), FieldSche
ma(name:tmp.kms_driven, type:string, comment:null), FieldSchema(name:tmp.owner, type:string, comment:null), FieldSchema(na
me:tmp.location, type:string, comment:null), FieldSchema(name:tmp.milage, type:string, comment:null), FieldSchema(name:tmp
.power, type:string, comment:null), FieldSchema(name:tmp.price, type:int, comment:null), FieldSchema(name:tmp.model_year,
type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007081234_812eaeeb-8887-4a5a-9ee5-2035d32481fe); Time taken: 0.377 s
econds
INFO  : Executing command(queryId=hive_20211007081234_812eaeeb-8887-4a5a-9ee5-2035d32481fe): insert into bikes_ext_dp sele
ct * from (select model_name, kms_driven,owner,location,milage,power,price,model_year from bikes_int) as tmp
INFO  : Query ID = hive_20211007081234_812eaeeb-8887-4a5a-9ee5-2035d32481fe
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_20211007081234_812eaeeb-8887-4a5a-9ee5-2035d32481fe
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: insert into bikes_ext_dp select * from...tmp (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1633487809125_0028)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
```

5:50 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Partitions in dynamic partitioned external table after inserting data into the table.

Ln 18, Col 37 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show partitions bikes_ext_dp;
INFO  : Compiling command(queryId=hive_20211007082351_64a60286-dc8e-48e1-9b85-1c8285b53475): show partitions bikes_ext_dp
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:string, comment:from deserializer)],
properties:null)
INFO  : Completed compiling command(queryId=hive_20211007082351_64a60286-dc8e-48e1-9b85-1c8285b53475); Time taken: 0.229 s
econds
INFO  : Executing command(queryId=hive_20211007082351_64a60286-dc8e-48e1-9b85-1c8285b53475): show partitions bikes_ext_dp
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007082351_64a60286-dc8e-48e1-9b85-1c8285b53475); Time taken: 0.056 s
econds
INFO  : OK
+-----------------------------------------+
|               partition                 |
+-----------------------------------------+
| model_year=1950                         |
| model_year=1970                         |
| model_year=1978                         |
| model_year=1982                         |
| model_year=1985                         |
| model_year=1986                         |
| model_year=1990                         |
| model_year=1991                         |
| model_year=1993                         |
| model_year=1994                         |
| model_year=1996                         |
| model_year=1997                         |
| model_year=1998                         |
| model_year=1999                         |
| model_year=2000                         |
| model_year=2001                         |
| model_year=2002                         |
| model_year=2003                         |
| model_year=2004                         |
| model_year=2005                         |
| model_year=2006                         |
| model_year=2007                         |
| model_year=2008                         |
| model_year=2009                         |
| model_year=2010                         |
| model_year=2011                         |
| model_year=2012                         |
| model_year=2013                         |
| model_year=2014                         |
| model_year=2015                         |
| model_year=2016                         |
| model_year=2017                         |
| model_year=2018                         |
| model_year=2019                         |
| model_year=2020                         |
| model_year=2021                         |
| model_year=7                            |
| model_year=__HIVE_DEFAULT_PARTITION__   |
+-----------------------------------------+
38 rows selected (0.32 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

5:55 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Time taken for querying external table with no partitions shown in "()".

-1,101 rows selected (1.273 seconds)

Ln 20, Col 37 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
| Royal Enfield Classic 350cc 2018                    | 2018                 | 21000 Km              | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 24000 Km              | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| TVS Apache RTR 200 4V Carburetor 2018               | 2018                 | Mileage 40 Kmpl       | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Hero Passion Xpro Disc 2018                         | 2018                 | 5500 Km               | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Bajaj CT 100 100cc 2018                             | 2018                 | 5661 Km               | first owner       |
bangalore           | "                    | NULL                 | NULL                 |
| Suzuki Gixxer SF 150cc ABS 2018                     | 2018                 | 9000 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Hero Passion Pro i3S Alloy 100cc 2018               | 2018                 | 8000 Km               | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Hero Super Splendor 125cc 2018                      | 2018                 | 8049 Km               | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 12552 Km              | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| TVS Apache RTR 200 4V FI 2018                       | 2018                 | Mileage 40 Kmpl       | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| Yamaha YZF-R15 V3 150cc 2018                        | 2018                 | 6900 Km               | first owner       |
chennai             | "                    | NULL                 | NULL                 |
| Bajaj Pulsar 220cc 2018                             | 2018                 | 22000 Km              | first owner       |
bangalore           | "                    | NULL                 | NULL                 |
| Bajaj Platina  Alloy ES-100cc 2018                  | 2018                 | Mileage 104 Kmpl      | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
| Hero HF Deluxe i3s 100cc 2018                       | 2018                 | 13000 Km              | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Bajaj Dominar 400 ABS 2018                          | 2018                 | Mileage 28 Kms        | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| UM Renegade Commando Classic 2018                   | 2018                 | 2911 Km               | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| TVS Star City Plus 110cc 2018                       | 2018                 | 1363 Km               | first owner       |
chitradurga         | "                    | NULL                 | NULL                 |
| TVS Sport 100cc 2018                                | 2018                 | 8048 Km               | first owner       |
delhi               | "                    | NULL                 | NULL                 |
| KTM Duke 390cc 2018                                 | 2018                 | 4500 Km               | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| KTM Duke 390cc 2018                                 | 2018                 | 12000 Km              | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Thunderbird 350cc 2018                | 2018                 | 20000 Km              | first owner       |
mumbai              | "                    | NULL                 | NULL                 |
| Royal Enfield Classic 350cc 2018                    | 2018                 | 14354 Km              | first owner       |
faridabad           | "                    | NULL                 | NULL                 |
+-----------------------------------------------------+----------------------+-----------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
|                bikes_ext.model_name                 | bikes_ext.model_year  | bikes_ext.kms_driven  |    bikes_ext.owner
   |  bikes_ext.location  | bikes_ext.milage  | bikes_ext.power  | bikes_ext.price  |
+-----------------------------------------------------+----------------------+-----------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
| TVS Apache RTR 160 4V Carburetor With Rear Disc 2018 | 2018                 | 16510 Km              | first owner
| ahmedabad           | "                    | NULL                 | NULL                 |
+-----------------------------------------------------+----------------------+-----------------------+-------------------
---+----------------------+-------------------+------------------+------------------+
1,101 rows selected (1.273 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext where model_year=2018
```

6:12 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Time taken for querying external table with dynamic partitions shown in "()".

-1,100 rows selected (0.447 seconds)

Which clearly concludes that partitioning plays as a major performance metric for external tables.

Ln 23, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

TVS Apache RTR 160 4V Carburetor 2018	11258 Km	first owner	faridabad
"	NULL	NULL	2018	
Suzuki Intruder 150cc 2018	1700 Km	first owner	faridabad
"	NULL	NULL	2018	
Hero Splendor Plus 100cc 2018	18000 Km	first owner	delhi
"	NULL	NULL	2018	
Royal Enfield Classic 350cc 2018	21000 Km	first owner	mumbai
"	NULL	NULL	2018	
Royal Enfield Classic 350cc 2018	24000 Km	first owner	mumbai
"	NULL	NULL	2018	
TVS Apache RTR 200 4V Carburetor 2018	Mileage 40 Kmpl	first owner	delhi
"	NULL	NULL	2018	
Hero Passion Xpro Disc 2018	5500 Km	first owner	delhi
"	NULL	NULL	2018	
Bajaj CT 100 100cc 2018	5661 Km	first owner	bangalore
"	NULL	NULL	2018	
Suzuki Gixxer SF 150cc ABS 2018	9000 Km	first owner	mumbai
"	NULL	NULL	2018	
Hero Passion Pro i3S Alloy 100cc 2018	8000 Km	first owner	delhi
"	NULL	NULL	2018	
Hero Super Splendor 125cc 2018	8049 Km	first owner	faridabad
"	NULL	NULL	2018	
Royal Enfield Classic 350cc 2018	12552 Km	first owner	faridabad
"	NULL	NULL	2018	
TVS Apache RTR 200 4V FI 2018	Mileage 40 Kmpl	first owner	delhi
"	NULL	NULL	2018	
Yamaha YZF-R15 V3 150cc 2018	6900 Km	first owner	chennai
"	NULL	NULL	2018	
Bajaj Pulsar 220cc 2018	22000 Km	first owner	bangalore
"	NULL	NULL	2018	
Bajaj Platina Alloy ES-100cc 2018	Mileage 104 Kmpl	first owner	faridabad
"	NULL	NULL	2018	
Hero HF Deluxe i3s 100cc 2018	13000 Km	first owner	mumbai
"	NULL	NULL	2018	
Bajaj Dominar 400 ABS 2018	Mileage 28 Kms	first owner	mumbai
"	NULL	NULL	2018	
UM Renegade Commando Classic 2018	2911 Km	first owner	delhi
"	NULL	NULL	2018	
TVS Star City Plus 110cc 2018	1363 Km	first owner	chitradurga
"	NULL	NULL	2018	
TVS Sport 100cc 2018	8048 Km	first owner	delhi
"	NULL	NULL	2018	
KTM Duke 390cc 2018	4500 Km	first owner	mumbai
"	NULL	NULL	2018	
KTM Duke 390cc 2018	12000 Km	first owner	mumbai
"	NULL	NULL	2018	
Royal Enfield Thunderbird 350cc 2018	20000 Km	first owner	mumbai
"	NULL	NULL	2018	
Royal Enfield Classic 350cc 2018	14354 Km	first owner	faridabad
"	NULL	NULL	2018	
TVS Apache RTR 160 4V Carburetor With Rear Disc 2018	16510 Km	first owner	ahmedabad
"	NULL	NULL	2018	

1,100 rows selected (0.447 seconds)

0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext_dp where model_year=2018;

6:16 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
We have cleaned the Dataset having empty lines in between and reloaded in all tables.

Ln 21, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
| model_year=1985  |
| model_year=1986  |
| model_year=1990  |
| model_year=1991  |
| model_year=1993  |
| model_year=1994  |
| model_year=1996  |
| model_year=1997  |
| model_year=1998  |
| model_year=1999  |
| model_year=2000  |
| model_year=2001  |
| model_year=2002  |
| model_year=2003  |
| model_year=2004  |
| model_year=2005  |
| model_year=2006  |
| model_year=2007  |
| model_year=2008  |
| model_year=2009  |
| model_year=2010  |
| model_year=2011  |
| model_year=2012  |
| model_year=2013  |
| model_year=2014  |
| model_year=2015  |
| model_year=2016  |
| model_year=2017  |
| model_year=2018  |
| model_year=2019  |
| model_year=2020  |
| model_year=2021  |
+------------------+
36 rows selected (0.269 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show tables;
INFO  : Compiling command(queryId=hive_20211007100334_d7f3131f-5aac-47ce-bcfe-4ad11424035b): show tables
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], p
roperties:null)
INFO  : Completed compiling command(queryId=hive_20211007100334_d7f3131f-5aac-47ce-bcfe-4ad11424035b); Time taken: 0.027 s
econds
INFO  : Executing command(queryId=hive_20211007100334_d7f3131f-5aac-47ce-bcfe-4ad11424035b): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211007100334_d7f3131f-5aac-47ce-bcfe-4ad11424035b); Time taken: 0.018 s
econds
INFO  : OK
+---------------+
|   tab_name    |
+---------------+
| bikes_ext     |
| bikes_ext_dp  |
| bikes_int     |
| bikes_int_p   |
+---------------+
4 rows selected (0.065 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

7:32 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Since Our Objectives is to findout the most affordable bikes in India which is newer than 2018 and has atleast 40 bhp of power (filtered using Regular Expressions) ,less than 200,000 on price and ordered on price.

We have got the results using the query:

```
select * from bikes_ext where model_year > 2018
and cast(price as int)<200000 and
cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3})
 bhp',1) as int) > 40;
```

Ln 20, Col 22 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext where model_year > 2018 and cast(price as int)<200000
 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) > 40 ORDER BY cast(price as int);
NFO  : Compiling command(queryId=hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173): select * from bikes_ext where
odel_year > 2018 and cast(price as int)<200000 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) > 40
ORDER BY cast(price as int)
NFO  : Semantic Analysis Completed (retrial = false)
NFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext.model_name, type:string, comment:null), Fiel
Schema(name:bikes_ext.model_year, type:smallint, comment:null), FieldSchema(name:bikes_ext.kms_driven, type:string, commen
:null), FieldSchema(name:bikes_ext.owner, type:string, comment:null), FieldSchema(name:bikes_ext.location, type:string, co
ment:null), FieldSchema(name:bikes_ext.milage, type:string, comment:null), FieldSchema(name:bikes_ext.power, type:string,
omment:null), FieldSchema(name:bikes_ext.price, type:int, comment:null)], properties:null)
NFO  : Completed compiling command(queryId=hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173); Time taken: 0.348 se
onds
NFO  : Executing command(queryId=hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173): select * from bikes_ext where
odel_year > 2018 and cast(price as int)<200000 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) > 40
ORDER BY cast(price as int)
NFO  : Query ID = hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173
NFO  : Total jobs = 1
NFO  : Launching Job 1 out of 1
NFO  : Starting task [Stage-1:MAPRED] in serial mode
NFO  : Subscribed to counters: [] for queryId: hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173
NFO  : Session is already open
NFO  : Dag name: select * from bikes_ext where model_y...int) (Stage-1)
NFO  : Tez session was closed. Reopening...
NFO  : Session re-established.
NFO  : Session re-established.
NFO  : Status: Running (Executing on YARN cluster with App id application_1633487809125_0033)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
ap 1 ..... container     SUCCEEDED      1          1        0        0       0       0
educer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
ERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.86 s
----------------------------------------------------------------------------------------------
NFO  : Status: DAG finished successfully in 6.78 seconds
NFO  :
NFO  : Query Execution Summary
NFO  : ----------------------------------------------------------------------------------------------
NFO  : OPERATION                            DURATION
NFO  : ----------------------------------------------------------------------------------------------
NFO  : Compile Query                          0.35s
NFO  : Prepare Plan                           0.10s
NFO  : Get Query Coordinator (AM)             0.00s
NFO  : Submit Plan                            5.87s
NFO  : Start DAG                              1.55s
NFO  : Run DAG                                6.78s
INFO : ----------------------------------------------------------------------------------------------
INFO :
INFO : Task Execution Summary
INFO : ----------------------------------------------------------------------------------------------
INFO :   VERTICES      DURATION(ms)    CPU_TIME(ms)    GC_TIME(ms)    INPUT_RECORDS    OUTPUT_RECORDS
INFO : ----------------------------------------------------------------------------------------------
INFO :      Map 1           4122.00          7,580            144          10,241                5
INFO :  Reducer 2            523.00            840             20               5                0
```

7:55 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
Since Our Objectives is to findout the most affordable bikes in India which is newer than 2018 and has atleast 40 bhp of power (filtered using Regular Expressions) ,less than 200,000 on price and ordered on price.

We have got the results using the query:

```
select * from bikes_ext where model_year > 2018
and cast(price as int)<200000 and
cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3})
 bhp',1) as int) > 40;
```

Time Taken : 5 rows selected (14.784 seconds)

Ln 29, Col 46 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
INFO  :    OUTPUT_LARGE_RECORDS: 0
INFO  :    OUTPUT_RECORDS: 5
INFO  :    SHUFFLE_CHUNK_COUNT: 1
INFO  :    SPILLED_RECORDS: 5
INFO  : TaskCounter_Reducer_2_INPUT_Map_1:
INFO  :    ADDITIONAL_SPILLS_BYTES_READ: 234
INFO  :    ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :    COMBINE_INPUT_RECORDS: 0
INFO  :    FIRST_EVENT_RECEIVED: 92
INFO  :    LAST_EVENT_RECEIVED: 92
INFO  :    MERGED_MAP_OUTPUTS: 1
INFO  :    MERGE_PHASE_TIME: 138
INFO  :    NUM_DISK_TO_DISK_MERGES: 0
INFO  :    NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :    NUM_MEM_TO_DISK_MERGES: 0
INFO  :    NUM_SHUFFLED_INPUTS: 1
INFO  :    NUM_SKIPPED_INPUTS: 0
INFO  :    REDUCE_INPUT_GROUPS: 5
INFO  :    REDUCE_INPUT_RECORDS: 5
INFO  :    SHUFFLE_BYTES: 234
INFO  :    SHUFFLE_BYTES_DECOMPRESSED: 383
INFO  :    SHUFFLE_BYTES_DISK_DIRECT: 234
INFO  :    SHUFFLE_BYTES_TO_DISK: 0
INFO  :    SHUFFLE_BYTES_TO_MEM: 0
INFO  :    SHUFFLE_PHASE_TIME: 122
INFO  :    SPILLED_RECORDS: 5
INFO  : TaskCounter_Reducer_2_OUTPUT_out_Reducer_2:
INFO  :    OUTPUT_RECORDS: 0
INFO  : org.apache.hadoop.hive.ql.exec.tez.HiveInputCounters:
INFO  :    GROUPED_INPUT_SPLITS_Map_1: 1
INFO  :    INPUT_DIRECTORIES_Map_1: 1
INFO  :    INPUT_FILES_Map_1: 37
INFO  :    RAW_INPUT_SPLITS_Map_1: 37
INFO  : Completed executing command(queryId=hive_20211007102548_45a36447-5893-4a98-8a38-3ae4f147a173); Time taken: 14.346 seconds
INFO  : OK
```

bikes_ext.model_name	bikes_ext.model_year	bikes_ext.kms_driven	bikes_ext.owner	bikes_ext.location	bikes_ext.milage	bikes_ext.power	bikes_ext.price
KTM Duke 390cc 2020	2020	18000 Km	first owner	mumbai	25 kmpl	42.90 bhp	150000
KTM RC 390cc 2019	2019	14000 Km	first owner	gurgaon	26kmpl	42.30 bhp	173175
KTM Duke 390cc 2019	2019	167 Km	first owner	nashik	25 kmpl	42.90 bhp	180000
KTM RC 390cc 2019	2019	1440 Km	first owner	delhi	26kmpl	42.30 bhp	190000
KTM RC 390cc 2019	2019	2500 Km	first owner	mumbai	26kmpl	42.30 bhp	197125

```
5 rows selected (14.784 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext where model_year > 2018 and cast(price as int)<20000
```

7:59 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
On running the same query on partitioned external table showing us significant time save

Time Taken :

Ln 19, Col 14 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from bikes_ext_dp where model_year > 2018 and cast(price as int)<20
0000 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) > 40 ORDER BY cast(price as int);
INFO  : Compiling command(queryId=hive_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631): select * from bikes_ext_dp whe
re model_year > 2018 and cast(price as int)<200000 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) >
 40 ORDER BY cast(price as int)
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:bikes_ext_dp.model_name, type:string, comment:null), F
ieldSchema(name:bikes_ext_dp.kms_driven, type:string, comment:null), FieldSchema(name:bikes_ext_dp.owner, type:string, comm
ent:null), FieldSchema(name:bikes_ext_dp.location, type:string, comment:null), FieldSchema(name:bikes_ext_dp.milage, type:s
tring, comment:null), FieldSchema(name:bikes_ext_dp.power, type:string, comment:null), FieldSchema(name:bikes_ext_dp.price,
 type:int, comment:null), FieldSchema(name:bikes_ext_dp.model_year, type:smallint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631); Time taken: 0.412 se
conds
INFO  : Executing command(queryId=hive_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631): select * from bikes_ext_dp whe
re model_year > 2018 and cast(price as int)<200000 and cast(REGEXP_EXTRACT(power, '(\\d{1,2}\\.?\\d{1,3}) bhp',1) as int) >
 40 ORDER BY cast(price as int)
INFO  : Query ID = hive_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631
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_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631
INFO  : Session is already open
INFO  : Dag name: select * from bikes_ext_dp where mode...int) (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1633487809125_0033)

----------------------------------------------------------------------------------------------
        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: 7.88 s
----------------------------------------------------------------------------------------------
INFO  : Status: DAG finished successfully in 7.82 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                          0.41s
INFO  : Prepare Plan                           0.12s
INFO  : Get Query Coordinator (AM)             0.00s
INFO  : Submit Plan                            0.07s
INFO  : Start DAG                              0.56s
INFO  : Run DAG                                7.82s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)    CPU_TIME(ms)    GC_TIME(ms)    INPUT_RECORDS    OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :      Map 1           4744.00          7,950            255              90                5
INFO  :  Reducer 2            463.00            930              0               5                0
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
```

8:02 PM
2021-10-13

*Group_E - Notepad

File Edit Format View Help

Group E
Data Technology Solutions

Anirudh Siddula
Kirandeep Kaur
Kamaldeep Kaur Dhami
Davinderjit Singh
Bibek Shah Shankar

Description:
On running the same query on partitioned external table showing us significant time save

Time Taken : 5 rows selected (9.068 seconds)

Conclusion: The most affordables bikes in India which matches or parameters is most likely from KTM manufacturer and the two model being KTM RC 390cc and KTM Duke 390cc

Ln 25, Col 1 150% Windows (CRLF) UTF-8

hive@sandbox-hdp:/root

```
INFO  :     OUTPUT_LARGE_RECORDS: 0
INFO  :     OUTPUT_RECORDS: 5
INFO  :     SHUFFLE_CHUNK_COUNT: 1
INFO  :     SPILLED_RECORDS: 5
INFO  : TaskCounter_Reducer_2_INPUT_Map_1:
INFO  :     ADDITIONAL_SPILLS_BYTES_READ: 232
INFO  :     ADDITIONAL_SPILLS_BYTES_WRITTEN: 0
INFO  :     COMBINE_INPUT_RECORDS: 0
INFO  :     FIRST_EVENT_RECEIVED: 99
INFO  :     LAST_EVENT_RECEIVED: 99
INFO  :     MERGED_MAP_OUTPUTS: 1
INFO  :     MERGE_PHASE_TIME: 158
INFO  :     NUM_DISK_TO_DISK_MERGES: 0
INFO  :     NUM_FAILED_SHUFFLE_INPUTS: 0
INFO  :     NUM_MEM_TO_DISK_MERGES: 0
INFO  :     NUM_SHUFFLED_INPUTS: 1
INFO  :     NUM_SKIPPED_INPUTS: 0
INFO  :     REDUCE_INPUT_GROUPS: 5
INFO  :     REDUCE_INPUT_RECORDS: 5
INFO  :     SHUFFLE_BYTES: 232
INFO  :     SHUFFLE_BYTES_DECOMPRESSED: 383
INFO  :     SHUFFLE_BYTES_DISK_DIRECT: 232
INFO  :     SHUFFLE_BYTES_TO_DISK: 0
INFO  :     SHUFFLE_BYTES_TO_MEM: 0
INFO  :     SHUFFLE_PHASE_TIME: 140
INFO  :     SPILLED_RECORDS: 5
INFO  : TaskCounter_Reducer_2_OUTPUT_out_Reducer_2:
INFO  :     OUTPUT_RECORDS: 0
INFO  : org.apache.hadoop.hive.ql.exec.tez.HiveInputCounters:
INFO  :     GROUPED_INPUT_SPLITS_Map_1: 1
INFO  :     INPUT_DIRECTORIES_Map_1: 3
INFO  :     INPUT_FILES_Map_1: 3
INFO  :     RAW_INPUT_SPLITS_Map_1: 3
INFO  : Completed executing command(queryId=hive_20211007102930_737dc141-aaee-4cb6-99e2-6e4ab57dc631); Time taken: 8.58 seconds
INFO  : OK
```

bikes_ext_dp.model_name	bikes_ext_dp.kms_driven	bikes_ext_dp.owner	bikes_ext_dp.location	bikes_ext_dp.milage	bikes_ext_dp.power	bikes_ext_dp.price	bikes_ext_dp.model_year
KTM Duke 390cc 2020	18000 Km	first owner	mumbai	25 kmpl	42.90 bhp	150000	2020
KTM RC 390cc 2019	14000 Km	first owner	gurgaon	26kmpl	42.30 bhp	173175	2019
KTM Duke 390cc 2019	167 Km	first owner	nashik	25 kmpl	42.90 bhp	180000	2019
KTM RC 390cc 2019	1440 Km	first owner	delhi	26kmpl	42.30 bhp	190000	2019
KTM RC 390cc 2019	2500 Km	first owner	mumbai	26kmpl	42.30 bhp	197125	2019

```
5 rows selected (9.068 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

8:06 PM
2021-10-13

Conclusion:

Hence we can conclude that partitioning has improved the performance of queries significantly as seen in the screenshots attached in slides.

Dataset Conclusion:

The most affordable bikes in India which matches or parameters is most likely from KTM manufacturer and the two model being **KTM RC 390c** and **KTM Duke 390cc.**