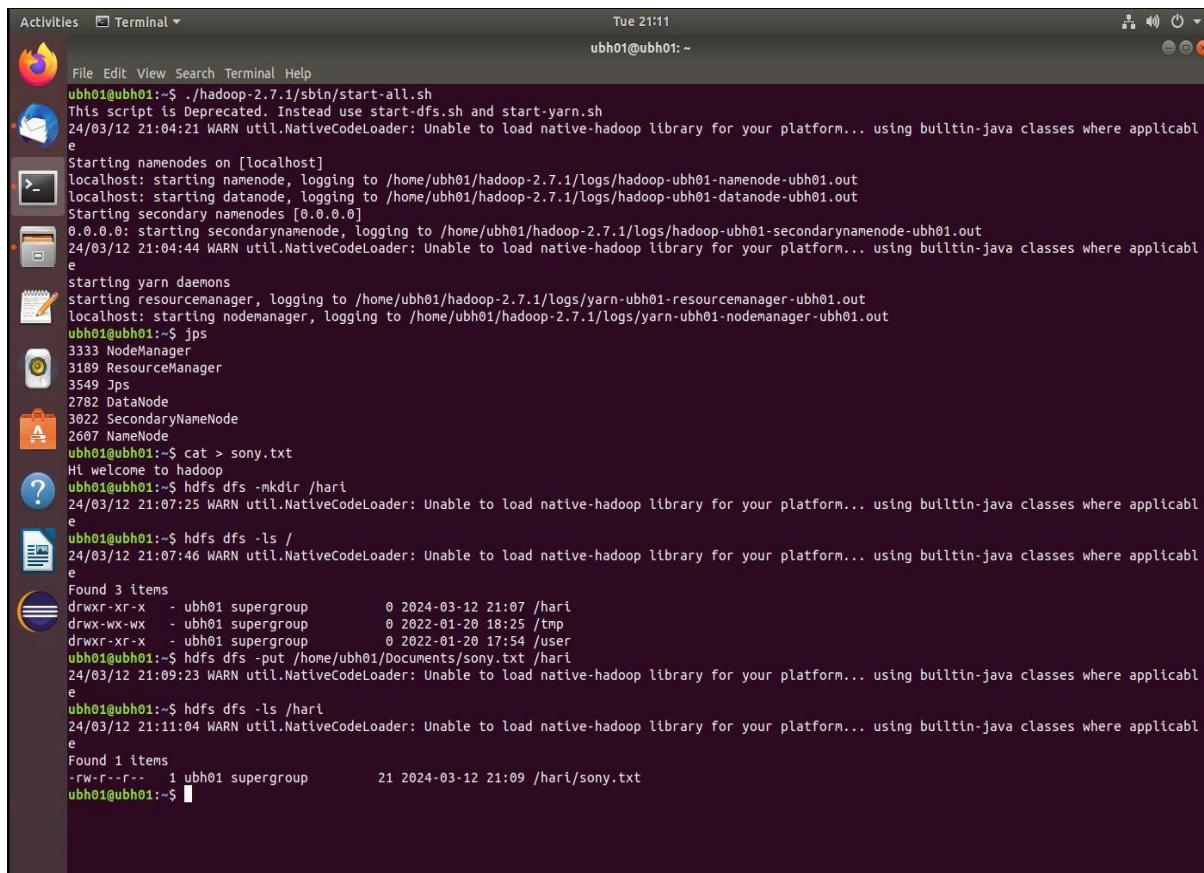


BIG DATA

CSDAIA24GP003
Chakka Sudheer Kumar
Emp-id:2320418

Hadoop

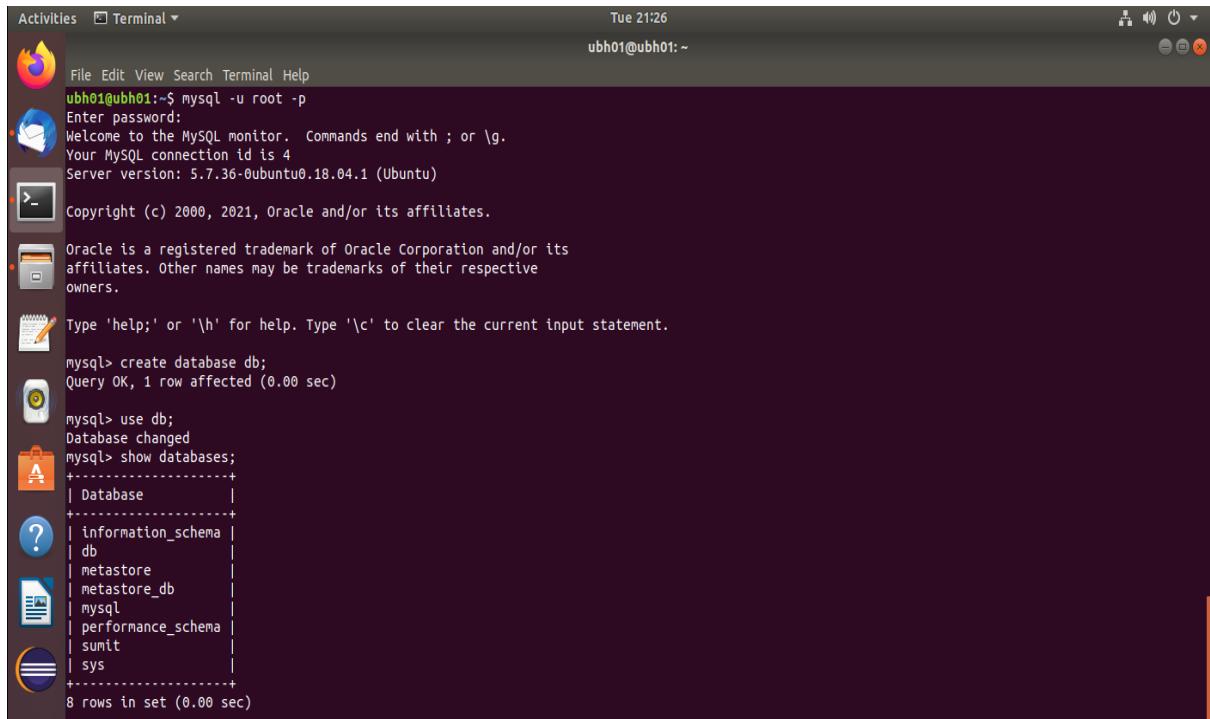
- Hadoop is an open-source framework used for storing and processing big data.
- It provides a distributed file system called Hadoop Distributed File System (HDFS).
 1. To start the Hadoop we use **--./hadoop-2.7.1/sbin/start-all.sh**
 2. To check whether Hadoop daemons are running or not --- **jps**
 3. Create a file in local -- **cat > filename.txt**
 4. Create a directory in local --- **mkdir name**
 5. Just check whether file is copied or not using --- **ls**



The screenshot shows a terminal window in an Ubuntu desktop environment. The terminal output is as follows:

```
Tue 21:11
ubh01@ubh01: ~
File Edit View Search Terminal Help
ubh01@ubh01:~$ ./hadoop-2.7.1/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
24/03/12 21:04:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-namenode-ubh01.out
localhost: starting datanode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-datanode-ubh01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-secondarynamenode-ubh01.out
24/03/12 21:04:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-resourcemanager-ubh01.out
localhost: starting nodemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-nodemanager-ubh01.out
ubh01@ubh01:~$ jps
3333 NodeManager
3189 ResourceManager
3549 Jps
2782 DataNode
3022 SecondaryNameNode
2607 NameNode
ubh01@ubh01:~$ cat > sony.txt
Hi welcome to hadoop
ubh01@ubh01:~$ hdfs dfs -mkdir /hari
24/03/12 21:07:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubh01@ubh01:~$ hdfs dfs -ls /
24/03/12 21:07:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x  - ubh01 supergroup      0 2024-03-12 21:07 /hari
drwxrwxr-x  - ubh01 supergroup      0 2022-01-20 18:25 /tmp
drwxr-xr-x  - ubh01 supergroup      0 2022-01-20 17:54 /user
ubh01@ubh01:~$ hdfs dfs -put /home/ubh01/Documents/sony.txt /hari
24/03/12 21:09:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubh01@ubh01:~$ hdfs dfs -ls /hari
24/03/12 21:11:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--  1 ubh01 supergroup      21 2024-03-12 21:09 /hari/sony.txt
ubh01@ubh01:~$
```

*** Start MYSQL and create a database with named db



```
ubh01@ubh01:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 4
Server version: 5.7.36-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

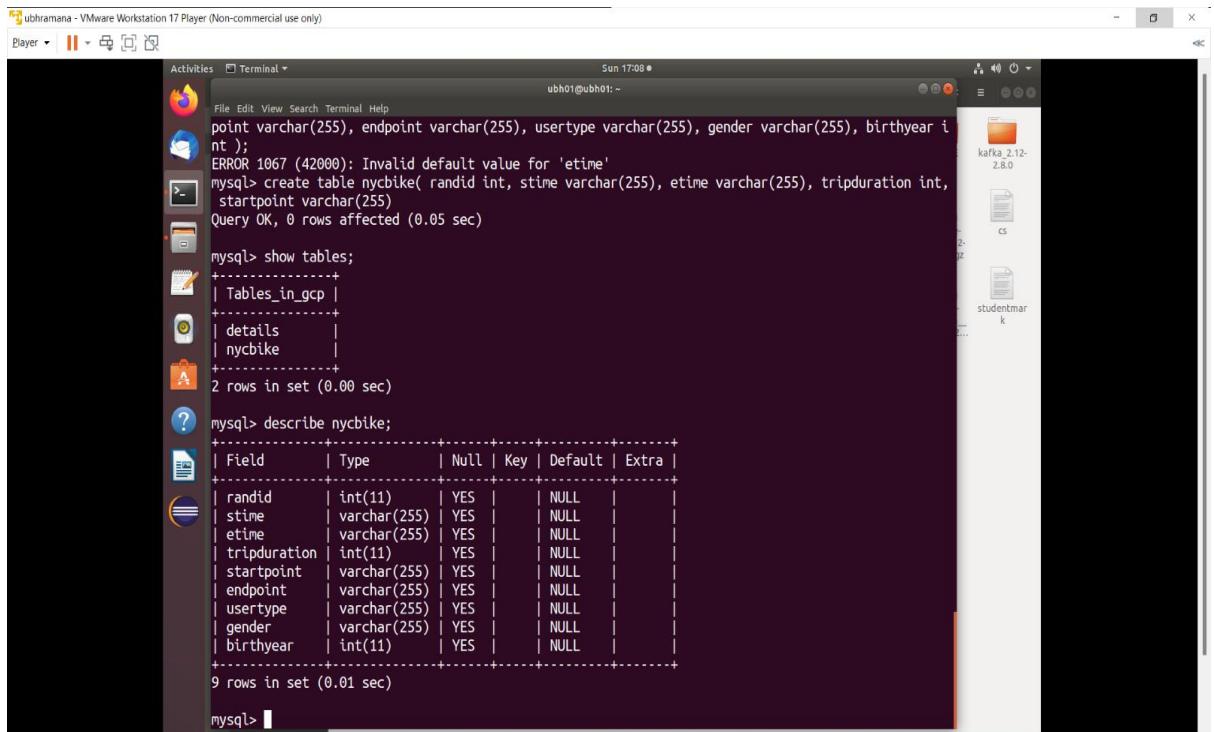
mysql> create database db;
Query OK, 1 row affected (0.00 sec)

mysql> use db;
Database changed
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| db |
| metastore |
| metastore_db |
| mysql |
| performance_schema |
| sumit |
| sys |
+-----+
8 rows in set (0.00 sec)
```

- Created table with table name student

Syntax:

```
Create table student(name varchar(20),roll int primary key, age int);
```



```
ubh01@ubh01:~$ mysql -u root -p
point varchar(255), endpoint varchar(255), usertype varchar(255), gender varchar(255), birthyear int );
ERROR 1067 (42000): Invalid default value for 'etime'
mysql> create table nycbike( randid int, stime varchar(255), etime varchar(255), tripduration int,
startpoint varchar(255)
Query OK, 0 rows affected (0.05 sec)

mysql> show tables;
+-----+
| Tables_in_gcp |
+-----+
| details |
| nycbike |
+-----+
2 rows in set (0.00 sec)

mysql> describe nycbike;
+-----+-----+-----+-----+-----+-----+
| Field | Type | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| randid | int(11) | YES | | NULL | |
| stime | varchar(255) | YES | | NULL | |
| etime | varchar(255) | YES | | NULL | |
| tripduration | int(11) | YES | | NULL | |
| startpoint | varchar(255) | YES | | NULL | |
| endpoint | varchar(255) | YES | | NULL | |
| usertype | varchar(255) | YES | | NULL | |
| gender | varchar(255) | YES | | NULL | |
| birthyear | int(11) | YES | | NULL | |
+-----+-----+-----+-----+-----+-----+
9 rows in set (0.01 sec)
```

- Inserted the values manually or load values from csv file as shown below:

```
mysql> load data local infile '/home/ubh01/new_york_city2.csv' into table nycbike fields terminated by ',' enclosed by '\"' lines terminated by '\n' ignore 1 rows (randid,stime,etime,tripduration,startpoint,endpoint,usertype,gender,birthyear);
Query OK, 837 rows affected, 98 warnings (0.28 sec)
Records: 837 Deleted: 0 Skipped: 0 Warnings: 98
mysql>
```

SQOOP:

It is a command line tool that helps to import and export data from database.

- Sqoop import:**

Sqoop import is used to import the data from RDBMS to HDFS.

***To start the Hadoop we use **./hadoop-2.7.1/sbin/start-all.sh**

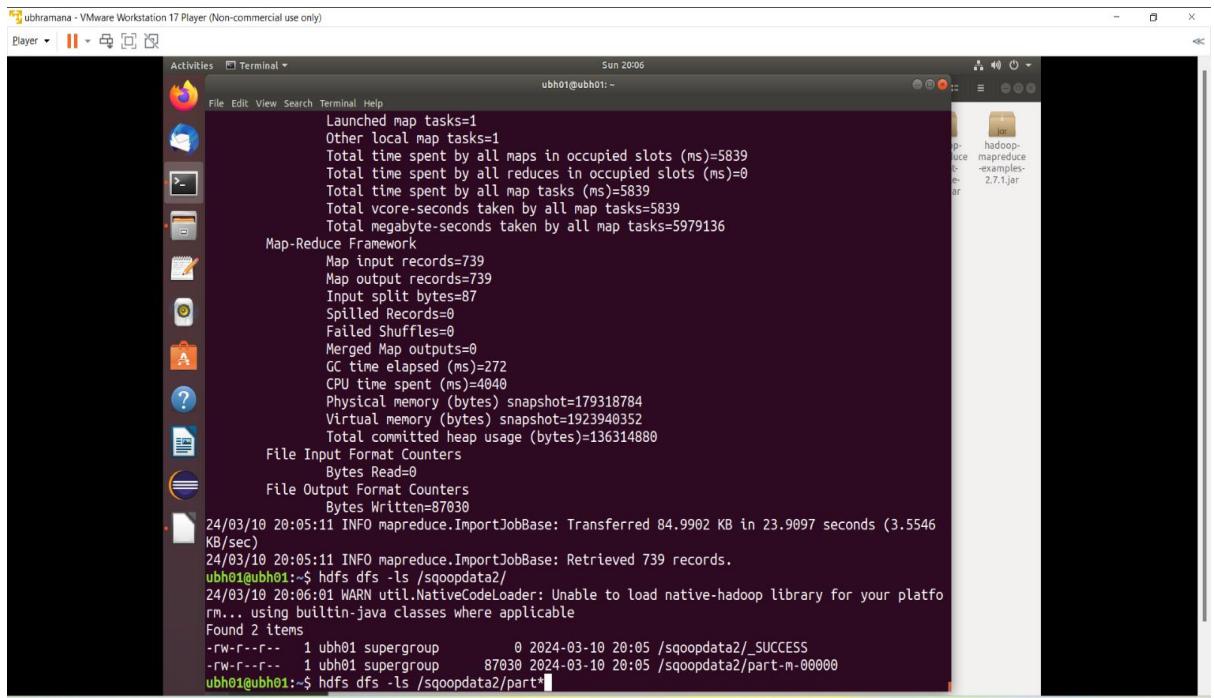
- Using the below command the table nycbike from database gcp is imported to hdfs through sqoop import

Syntax:

Sqoop import –connectjdbc:mysql://localhost:3306/gcp –username root -p –query –table student –target-dir /sqoopdata

```
ubh01@ubh01:~$ ./hadoop-2.7.1/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
24/03/12 21:32:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 2607. Stop it first.
localhost: datanode running as process 2782. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3022. Stop it first.
24/03/12 21:32:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
resourcemanager running as process 3189. Stop it first.
localhost: nodemanager running as process 3333. Stop it first.
ubh01@ubh01:~$ sqoop import --connect jdbc:mysql://localhost:3306/db --username root -P --table student --target-dir /sqoopdata
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/.hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/.zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
24/03/12 21:34:56 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
24/03/12 21:35:04 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/12 21:35:04 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
24/03/12 21:35:05 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
24/03/12 21:35:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
24/03/12 21:35:06 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/48773eeaf4f6d38a14bd327e9308eeaf/student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/12 21:35:09 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ubh01/compile/48773eeaf4f6d38a14bd327e9308eeaf/student.jar
24/03/12 21:35:09 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/03/12 21:35:09 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/03/12 21:35:09 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/03/12 21:35:09 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/03/12 21:35:09 INFO mapreduce.ImportJobBase: Beginning import of student
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.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.slf4j.impl.Log4jLoggerFactory]
24/03/12 21:35:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

- Listing the output files of sqoopdata file which includes _SUCCESS and PART FILES using ls command.

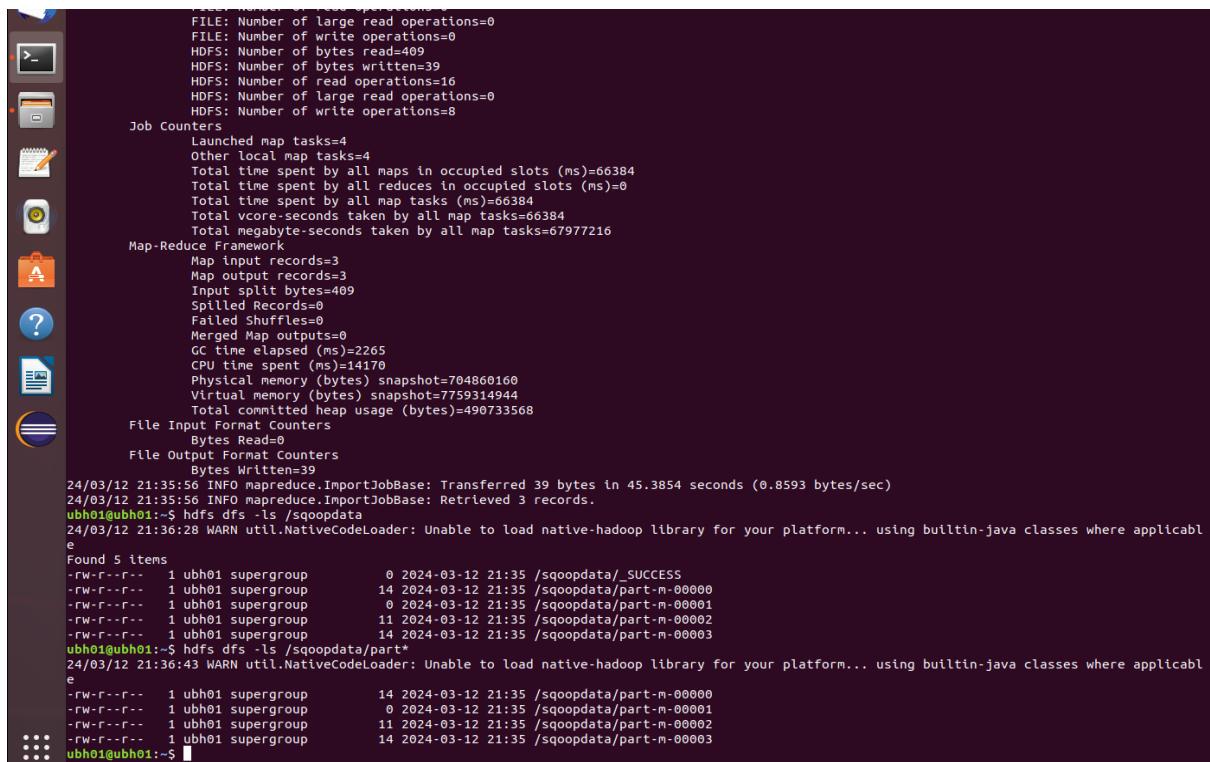


```

ubhramana - VMware Workstation 17 Player (Non-commercial use only)
Activities Terminal Sun 20:06
ubh01@ubh01:~$ 
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=5839
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=5839
Total vcore-seconds taken by all map tasks=5839
Total megabyte-seconds taken by all map tasks=5979136
Map-Reduce Framework
    Map input records=739
    Map output records=739
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=272
    CPU time spent (ms)=4040
    Physical memory (bytes) snapshot=179318784
    Virtual memory (bytes) snapshot=1923940352
    Total committed heap usage (bytes)=136314880
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=87030
24/03/10 20:05:11 INFO mapreduce.ImportJobBase: Transferred 84.9902 KB in 23.9097 seconds (3.5546 KB/sec)
24/03/10 20:05:11 INFO mapreduce.ImportJobBase: Retrieved 739 records.
ubh01@ubh01:~$ hdfs dfs -ls /sqoopdata2/
24/03/10 20:06:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 ubh01 supergroup      0 2024-03-10 20:05 /sqoopdata2/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup  87030 2024-03-10 20:05 /sqoopdata2/part-m-00000
ubh01@ubh01:~$ hdfs dfs -ls /sqoopdata2/part*

```

- Display the content inside PART FILE using cat command.



```

FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=409
HDFS: Number of bytes written=39
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
    Launched map tasks=4
    Other local map tasks=4
    Total time spent by all maps in occupied slots (ms)=66384
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=66384
    Total vcore-seconds taken by all map tasks=66384
    Total megabyte-seconds taken by all map tasks=67977216
Map-Reduce Framework
    Map input records=3
    Map output records=3
    Input split bytes=409
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=2265
    CPU time spent (ms)=14170
    Physical memory (bytes) snapshot=704860160
    Virtual memory (bytes) snapshot=7759314944
    Total committed heap usage (bytes)=490733568
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=39
24/03/12 21:35:56 INFO mapreduce.ImportJobBase: Transferred 39 bytes in 45.3854 seconds (0.8593 bytes/sec)
24/03/12 21:35:56 INFO mapreduce.ImportJobBase: Retrieved 3 records.
ubh01@ubh01:~$ hdfs dfs -ls /sqoopdata
24/03/12 21:36:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 5 items
-rw-r--r-- 1 ubh01 supergroup      0 2024-03-12 21:35 /sqoopdata/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup  14 2024-03-12 21:35 /sqoopdata/part-m-00000
-rw-r--r-- 1 ubh01 supergroup      0 2024-03-12 21:35 /sqoopdata/part-m-00001
-rw-r--r-- 1 ubh01 supergroup   11 2024-03-12 21:35 /sqoopdata/part-m-00002
-rw-r--r-- 1 ubh01 supergroup  14 2024-03-12 21:35 /sqoopdata/part-m-00003
ubh01@ubh01:~$ hdfs dfs -ls /sqoopdata/part*
24/03/12 21:36:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r-- 1 ubh01 supergroup      14 2024-03-12 21:35 /sqoopdata/part-m-00000
-rw-r--r-- 1 ubh01 supergroup      0 2024-03-12 21:35 /sqoopdata/part-m-00001
-rw-r--r-- 1 ubh01 supergroup   11 2024-03-12 21:35 /sqoopdata/part-m-00002
-rw-r--r-- 1 ubh01 supergroup  14 2024-03-12 21:35 /sqoopdata/part-m-00003
ubh01@ubh01:~$ 

```

- Sqoop import command with **where** clause.

Syntax:

```
Sqoop import --connect jdbc:mysql://localhost:3306/db --username root -p --query "select * from student where 'roll'=4" --target-dir /sqoopdata2
```

```
ubh01@ubh01:~$ sqoop import --connect jdbc:mysql://localhost:3306/db --username root -p --table student --where "roll=4" --target-dir /sqoopdata2
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set SHCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
24/03/12 22:06:18 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
24/03/12 22:06:26 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/12 22:06:26 INFO tool.CodeGenTool: Beginning code generation
Loading class com.mysql.jdbc.Driver. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
24/03/12 22:06:27 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
24/03/12 22:06:28 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
24/03/12 22:06:28 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/5b3faddaf773391734c853895500ee60/student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/12 22:06:33 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ubh01/compile/5b3faddaf773391734c853895500ee60/student.jar
24/03/12 22:06:33 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/03/12 22:06:33 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/03/12 22:06:33 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/03/12 22:06:33 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/03/12 22:06:33 INFO mapreduce.ImportJobBase: Beginning import of student
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.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.slf4j.impl.Log4jLoggerFactory]
24/03/12 22:06:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/03/12 22:06:34 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/03/12 22:06:35 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/03/12 22:06:35 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/12 22:06:41 INFO db.DBInputFormat: Using read committed transaction isolation
24/03/12 22:06:41 INFO db.DataDrivenDBInputFormat: BoundingAlgebraQuery: SELECT MIN(`roll`), MAX(`roll`) FROM `student` WHERE ( `roll` = 4 )
24/03/12 22:06:41 INFO db.IntegerSplitter: Split size: 0; Num splits: 4 from: 4 to: 4
24/03/12 22:06:41 INFO mapreduce.JobSubmitter: number of splits:1
24/03/12 22:06:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710261014202_0001
24/03/12 22:06:42 INFO impl.YarnClientImpl: Submitted application application_1710261014202_0001
24/03/12 22:06:42 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1710261014202_0001/
24/03/12 22:06:42 INFO mapreduce.Job: Running job: job_1710261014202_0001
24/03/12 22:06:58 INFO mapreduce.Job: Job job_1710261014202_0001 running in uber mode : false
24/03/12 22:06:58 INFO mapreduce.Job: map 0% reduce 0%
24/03/12 22:07:08 INFO mapreduce.Job: map 100% reduce 0%
24/03/12 22:07:09 INFO mapreduce.Job: Job job_1710261014202_0001 completed successfully
```

- If we want to verify whether values that satisfies where condition imported to hdfs or not, run the below hdfs command **--hdfs dfs -cat /sqoopdata2/part***

```
24/03/12 22:06:58 INFO mapreduce.Job: map 0% reduce 0%
24/03/12 22:07:08 INFO mapreduce.Job: map 100% reduce 0%
24/03/12 22:07:09 INFO mapreduce.Job: Job job_1710261014202_0001 completed successfully
24/03/12 22:07:09 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=134667
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=103
    HDFS: Number of bytes written=11
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=6579
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=6579
    Total vcore-seconds taken by all map tasks=6579
    Total megabyte-seconds taken by all map tasks=6736896
  Map-Reduce Framework
    Map input records=1
    Map output records=1
    Input split bytes=103
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=3500
    CPU time spent (ms)=3500
    Physical memory (bytes) snapshot=179347456
    Virtual memory (bytes) snapshot=1939496960
    Total committed heap usage (bytes)=115867648
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=11
24/03/12 22:07:09 INFO mapreduce.ImportJobBase: Transferred 11 bytes in 34.2373 seconds (0.3213 bytes/sec)
24/03/12 22:07:09 INFO mapreduce.ImportJobBase: Retrieved 1 records.
ubh01@ubh01:~$ hdfs dfs -cat /sqoopdata2/part*
24/03/12 22:08:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cat: '/sqoopdata2/part*': Is a directory
ubh01@ubh01:~$ hdfs dfs -cat /sqoopdata2/part*
24/03/12 22:08:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
privat 4,11
ubh01@ubh01:~$
```

- Sqoop import with –query command where we will specify mysql query and \\$CONDITIONS that is a placeholder in query which divide the data with more where clauses internally and –split-by is used to split the data by particular column that is primary key.

```
ubh01@ubh01:~$ sqoop import --connect jdbc:mysql://localhost:3306/db --username root -P --query "select * from student where age=31 and \$CONDITIONS" --split-by roll --target-dir /sqoopdata3
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set SHCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../accumulo does not exist! Accumulo imports will fail.
Please set SACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../zookeeper does not exist! Accumulo imports will fail.
Please set ZOOKEEPER_HOME to the root of your Zookeeper installation.
24/03/12 22:13:55 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
24/03/12 22:14:03 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/12 22:14:03 INFO tool.CodeGenTool: Beginning code generation
Loading class com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
24/03/12 22:14:04 INFO manager.SqlManager: Executing SQL statement: select * from student where age=31 and (1 = 0)
24/03/12 22:14:04 INFO manager.SqlManager: Executing SQL statement: select * from student where age=31 and (1 = 0)
24/03/12 22:14:04 INFO manager.SqlManager: Executing SQL statement: select * from student where age=31 and (1 = 0)
24/03/12 22:14:04 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/3f725e71c53b379ff1f208666e112f990/QueryResult.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/12 22:14:07 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ubh01/compile/3f725e71c53b379ff1f208666e112f990/QueryResult.jar
24/03/12 22:14:07 INFO mapreduce.ImportJobBase: Beginning query import.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j.impl.StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.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.slf4j.impl.Log4jLoggerFactory]
24/03/12 22:14:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/03/12 22:14:07 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/03/12 22:14:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/03/12 22:14:08 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/12 22:14:12 INFO db.DBInputFormat: Using read committed transaction isolation
24/03/12 22:14:12 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(roll), MAX(roll) FROM (select * from student where age=31 and (1 = 1)) AS t1
24/03/12 22:14:12 INFO db.IntegersSplitter: Split size: 0; Num splits: 4 from: 6 to: 6
24/03/12 22:14:12 INFO mapreduce.JobSubmitter: number of splits:1
24/03/12 22:14:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710261014202_0002
24/03/12 22:14:13 INFO impl.YarnClientImpl: Submitted application application_1710261014202_0002
24/03/12 22:14:13 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1710261014202_0002/
24/03/12 22:14:13 INFO mapreduce.Job: Running job: job_1710261014202_0002
```

- If we want to verify whether values that satisfies QUERY imported to hdfs or not, run the below hdfs command **--hdfs dfs -cat /sqoopdata3/part***

```
24/03/12 22:14:13 INFO impl.YarnClientImpl: Submitted application application_1710261014202_0002
24/03/12 22:14:13 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1710261014202_0002/
24/03/12 22:14:13 INFO mapreduce.Job: Running job: job_1710261014202_0002
24/03/12 22:14:22 INFO mapreduce.Job: Job job_1710261014202_0002 running in uber mode : false
24/03/12 22:14:22 INFO mapreduce.Job: map 0% reduce 0%
24/03/12 22:14:32 INFO mapreduce.Job: Job job_1710261014202_0002 completed successfully
24/03/12 22:14:32 INFO mapreduce.Job: Counters: 30
File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=134601
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=99
    HDFS: Number of bytes written=14
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=6387
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=6387
    Total vcore-seconds taken by all map tasks=6387
    Total megabyte-seconds taken by all map tasks=6540288
Map-Reduce Framework
    Map input records=1
    Map output records=1
    Input split bytes=99
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=280
    CPU time spent (ms)=3050
    Physical memory (bytes) snapshot=176640000
    Virtual memory (bytes) snapshot=1938382848
    Total committed heap usage (bytes)=116916224
File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=14
24/03/12 22:14:32 INFO mapreduce.ImportJobBase: Transferred 14 bytes in 23.9802 seconds (0.5838 bytes/sec)
24/03/12 22:14:32 INFO mapreduce.ImportJobBase: Retrieved 1 records.
ubh01@ubh01:~$ hdfs dfs -cat /sqoopdata3/part*
24/03/12 22:15:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
priyaswi,6,31
ubh01@ubh01:~$
```

- **Sqoop export:**

- Sqoop export is used to import the data from HDFS to RDBMS.
- Here we have to mention export keyword and –export-dir and before this we have to create a empty schema or table in mysql.

```
ubh01@ubh01:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 11
Server version: 5.7.36-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from students1;
ERROR 1146 (42002): Table 'db.students1' doesn't exist
mysql> create table students1(name varchar(20),roll int primary key,age int);
Query OK, 0 rows affected (0.02 sec)

mysql> exit;
Bye
ubh01@ubh01:~$ sqoop export --connect jdbc:mysql://localhost:3306/db --username root -P --table students1 --export-dir /sqoopdata
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin_hadoop-2.6.0/../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
24/03/12 22:29:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
24/03/12 22:29:17 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/12 22:29:17 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
24/03/12 22:29:18 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `students1` AS t LIMIT 1
24/03/12 22:29:18 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `students1` AS t LIMIT 1
24/03/12 22:29:18 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/04e7b59d329fc2d205f6441c9534c6877/students1.java uses or overrides a deprecated API.
24/03/12 22:29:21 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ubh01/compile/04e7b59d329fc2d205f6441c9534c6877/students1.jar
24/03/12 22:29:21 INFO mapreduce.ExportJobBase: Beginning export of students1
```

- If we want to verify whether values are exported from hdfs to mysql or not by running the below command in mysql—select * from students1;

```
Total vcore-seconds taken by all map tasks=50398
Total megabyte-seconds taken by all map tasks=51607552
Map-Reduce Framework
  Map Input records=3
  Map output records=3
  Input split bytes=683
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2270
  CPU time spent (ms)=11000
  Physical memory (bytes) snapshot=701575168
  Virtual memory (bytes) snapshot=7746822144
  Total committed heap usage (bytes)=475529216
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=0
24/03/12 22:29:51 INFO mapreduce.ExportJobBase: Transferred 760 bytes in 28.9485 seconds (26.2535 bytes/sec)
24/03/12 22:29:51 INFO mapreduce.ExportJobBase: Exported 3 records.
ubh01@ubh01:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 17
Server version: 5.7.36-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

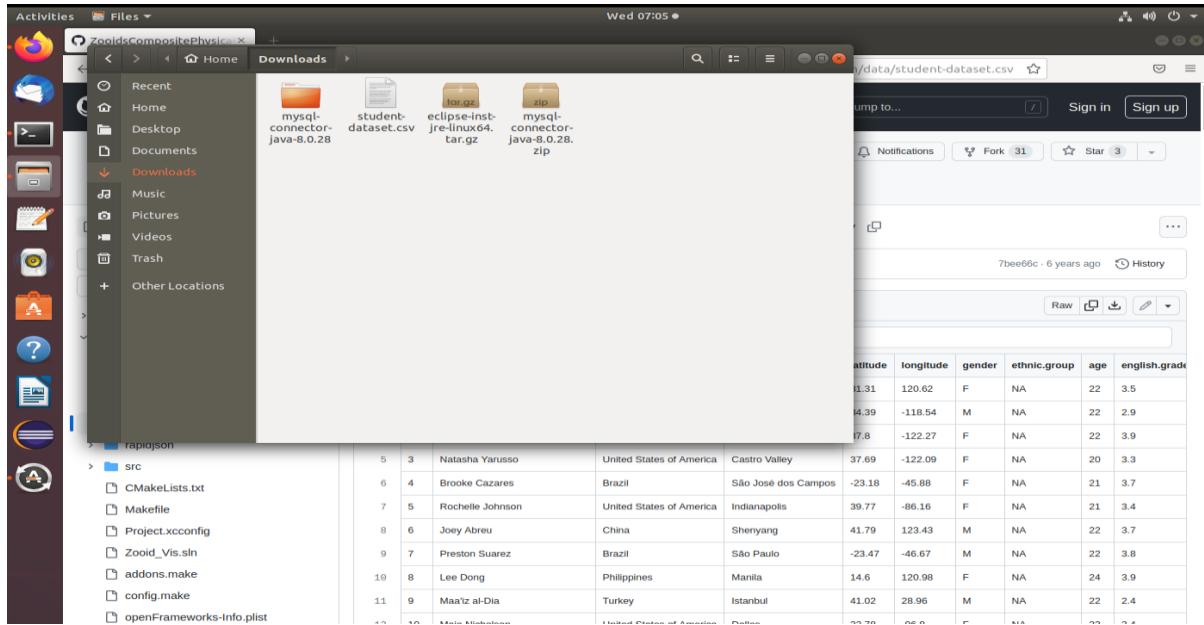
mysql> use db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from students1;
+-----+-----+-----+
| name | roll | age |
+-----+-----+-----+
| priyanka | 1 | 12 |
| priya | 4 | 11 |
| priyaswi | 6 | 31 |
+-----+-----+-----+
3 rows in set (0.00 sec)
```

HIVE

Hive is a data warehouse system which is used for querying and analysing large datasets stored in HDFS.

In hive here we are downloading the dataset(student-dataset.csv) from github.



- To enter into the hive, use **hive** command.
- Create a database and use database.

```
ubh01@ubh01:~$ ./hadoop-2.7.1/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
24/03/13 07:07:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-namenode-ubh01.out
localhost: starting datanode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-datanode-ubh01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-secondarynamenode-ubh01.out
24/03/13 07:07:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-resourcemanager-ubh01.out
localhost: starting nodemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-nodemanager-ubh01.out
ubh01@ubh01:~$ jps
3328 ResourceManager
2949 DataNode
2773 NameNode
3911 Jps
3162 SecondaryNameNode
3469 NodeManager
ubh01@ubh01:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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]

Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> use gcp;
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
FAILED: SemanticException [Error 10072]: Database does not exist: gcp
hive> create database gcp;
OK
Time taken: 0.514 seconds
hive> use gcp;
OK
Time taken: 0.036 seconds
hive> show databases;
OK
default
gcp
sumitdb
```

- Here we are creating a hive table and inserted values into table using load command.
- To verify whether the data is loaded are not, use command--**select * from student3;**

```

hive> create table student3(id int,name varchar(20),nationality varchar(20),city varchar(20),gender varchar(20),age int)row format delimited field
s terminated by ',' 'tblproperties('skip.header.line.count'=1');
OK
Time taken: 0.989 seconds
hive> load data local inpath '/home/ubh01/Downloads/student-dataset.csv' into table student3;
FAILED: ParseException line 1:10 extraneous input 'loacal' expecting INPATH near '<EOF>'
hive> load data local inpath '/home/ubh01/Downloads/student-dataset.csv' into table student3;
Loading data to table gcp.student3
OK
Time taken: 2.014 seconds
hive> select * from student3;
OK
0      Kiana Lor      China  Suzhou   F      22
1      Joshua Lonaker United States of Ame Santa Clarita   M      22
2      Dakota Blanco  United States of Ame Oakland   F      22
3      Natasha Yarusso United States of Ame Castro Valley   F      20
4      Brooke Cazares Brazil  São José dos Campos   F      21
5      Rochelle Johnson United States of Ame Indianapolis   F      21
6      Joey Abreu     China  Shenyang   M      22
7      Preston Suarez Brazil  São Paulo   M      22
8      Lee Dong       Philippines Manila   F      24
9      Maa'iz al-Dia   Turkey  İstanbul   M      22
10     Maja Nicholson United States of Ame Dallas   F      23
11     Sasha Jansen   United States of Ame Chicago   F      21
12     Alexander Sherman United States of Ame Omaha   M      20
13     Edgar Sanchez  Mexico  Tijuana   M      23
14     Kolbt Strunk   United States of Ame Mission Viejo   M      21
15     Brittany Sath  Japan   Tokyo   F      21
16     Megan Smith    United States of Ame Los Angeles   F      21
17     Ericka Arreola Mexico  Mexico other   23
18     David Pulc     Canada  Toronto   M      24
19     Kyle Luckey    United States of Ame Moreno Valley   M      23
20     Rojesh Her     Japan   Tokyo   M      22
21     David Weber   China  Peking   M      20
22     Rachel Jambor United States of Ame Chicago   F      22
23     Mus'ab al-Moustafa Pakistan Rawalpindi   M      23
24     Sila Nguyen    China  Hebi   M      23

```

Hive Partitioning:

Create a partition table and follow either static or dynamic partitioning to insert values from main table following conditions.

- Static partitioning:
 - It is setup when the table is created and remains static, meaning the partitions don't change unless explicitly altered by the user.
 - Here we are inserting values of a main table into partition table using static partitioning.

```

hive> use gcp;
OK
Time taken: 0.028 seconds
hive> create table student3(id int,name varchar(20),nationality varchar(20),city varchar(20),age int)partitioned by(gender varchar(10));
FAILED: ParseException line 1:95 missing EOF at 'partitioned' near ')'
hive> create table student3(id int,name varchar(20),nationality varchar(20),city varchar(20),age int)partitioned by(gender varchar(10));
FAILED: ParseException line 1:95 missing EOF at 'partitioned' near ')'
hive> use gcp;
OK
Time taken: 0.028 seconds
hive> create table student3(id int,name varchar(20),nationality varchar(20),city varchar(20),age int)partitioned by(gender varchar(10));
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table student3 already exists)
hive> use gcp;
OK
Time taken: 0.03 seconds
hive> create table sp3(id int,name varchar(20),nationality varchar(20),city varchar(20),age int)partitioned by(gender varchar(10));
OK
Time taken: 0.111 seconds
hive> inser into table sp3 partition(gender='M')select name,nationality,city,gender from student3 where gender='M';
NoVisibleArgumentException[24@[]]
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1300)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.Runjar.run(Runjar.java:221)
        at org.apache.hadoop.util.Runjar.main(Runjar.java:136)
FAILED: ParseException line 1:8 cannot recognize input near 'inser' 'into' 'table'
hive> inser into table sp3 partition(gender='M')select name,nationality,city,gender from student3 where gender='M';
FAILED: SemanticException [Error 10044]: Line 1:18 Cannot insert into target table because column number/types are different ''M'': Table insclaus
e-0 has 5 columns, but query has 4 columns.
hive> inser into table sp3 partition(gender='M')select name,nationality,city,gender from student3 where gender='M';

```

- Here is the partition table for partition gender = 'M' shown below

```
2024-03-13 07:35:27,621 Stage-1 Map = 100%, reduce = 0%, Cumulative CPU 5.94 sec
MapReduce Total cumulative CPU time: 5 seconds 940 msec
Ended Job = job_1710293849024_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/gcp.db/sp3/gender=M/.hive-staging_hive_2024-03-13_07-34-56_408_1044200944667402
076-1/-ext-10000
Loading data to table gcp.sp3 partition (gender=M)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 5.94 sec HDFS Read: 21322 HDFS Write: 5216 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 940 msec
OK
Time taken: 33.616 seconds
hive> select * from sp3;
OK
NULL United States of Ame Santa Clarita M 22 M
NULL China Shenyang M 22 M
NULL Brazil São Paulo M 22 M
NULL Turkey İstanbul M 22 M
NULL United States of Ame Omaha M 20 M
NULL Mexico Tijuana M 23 M
NULL United States of Ame Mission Viejo M 21 M
NULL Canada Toronto M 24 M
NULL United States of Ame Moreno Valley M 23 M
NULL Japan Tokyo M 22 M
NULL China Peking M 20 M
NULL Pakistan Rawalpindi M 23 M
NULL China Hebi M 23 M
NULL United States of Ame South Hill M 21 M
NULL United States of Ame Phoenix M 21 M
NULL United States of Ame Los Angeles M 21 M
NULL United States of Ame Detroit M 23 M
NULL Peru Lima M 22 M
NULL United States of Ame New York M 24 M
NULL Japan Tokyo M 22 M
NULL Japan Tokyo M 20 M
NULL United States of Ame Los Angeles M 23 M
NULL Mexico Tehuacán M 23 M
NULL Spain Madrid M 21 M
NULL United States of Ame Charlotte M 23 M
NULL United States of Ame Los Angeles M 20 M
NULL Pakistan Khalabat M 22 M
NULL China Shanghai M 21 M
NULL United States of Ame San Jose M 23 M
NULL Poland Warsaw M 21 M
NULL United States of Ame Stockton M 22 M
NULL United States of Ame Sacramento M 21 M
NULL United States of Ame Los Angeles M 21 M
NULL China Wuhan M 21 M
NULL Tunisia Manzil Tamim M 23 M
```

- Dynamic partitioning:**

- It involves automatically creating partitions for data based on the values of specified columns during the insertion process.
- Here we are inserting values of a main table into partition table using dynamic partitioning.

```
ubh01@ubh01:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j.impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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]

Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> use gcp;
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
OK
Time taken: 4.799 seconds
hive> create table dp(id int,name varchar(10),nationality varchar(20),city varchar(10),age int)partitioned by(gender varchar(10));
OK
Time taken: 0.645 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> insert into table dp partition(gender)select * from student3 where gender='M';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ubh01_20240313193005_e6aae7b8-5046-439a-8332-25648258ab61
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1710337172011_0001, Tracking URL = http://ubh01:8088/proxy/application_1710337172011_0001/
KILL Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job -kill job_1710337172011_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-03-13 19:30:18,909 Stage-1 map = 0%, reduce = 0%
2024-03-13 19:30:26,594 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.33 sec
MapReduce Total cumulative CPU time: 5 seconds 330 msec
Ended Job = job_1710337172011_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/gcp.db/dp/.hive-staging_hive_2024-03-13_19-30-05_070_4190010835581164805-1/-ext-10000
Loading data to table gcp.dp partition (gender=null)

Loaded : 8/8 partitions.
Time taken to load dynamic partitions: 1.569 seconds
Time taken for adding to write entity : 0.003 seconds
MapReduce Jobs Launched:
```

- To verify whether the data is loaded are not, use command--**select * from dp;**

```

hive> select * from dp;
OK
+-----+-----+-----+-----+-----+-----+-----+
| id   | name | nationality | city | gender | age |    |
+-----+-----+-----+-----+-----+-----+-----+
| 173  | Sean Bruso | United Kingdom | London | NULL | 19 |    |
| 12   | Alexander | United States of Ame | Omaha | NULL | 20 |    |
| 21   | David Webe | China | Peking | NULL | 20 |    |
| 46   | Alan Trinh | Japan | Tokyo | NULL | 20 |    |
| 58   | Cameron St | United States of Ame | Los Angele | NULL | 20 |    |
| 94   | Devon Mira | Colombia | Pereira | NULL | 20 |    |
| 98   | Jason Hund | United States of Ame | Los Angele | NULL | 20 |    |
| 106  | Ryan Barre | United States of Ame | Chicago | NULL | 20 |    |
| 121  | James Rice | Canada | Toronto | NULL | 20 |    |
| 150  | Abdul Jabb | Egypt | Cairo | NULL | 20 |    |
| 196  | Dylan Bell | United States of Ame | Tucson | NULL | 20 |    |
| 235  | Ryan Tyler | United States of Ame | Los Angele | NULL | 20 |    |
| 268  | Bryant Ron | Mexico | Aguascalie | NULL | 20 |    |
| 269  | Garrett He | United States of Ame | Mobile | NULL | 20 |    |
| 281  | Dylan Pai | United States of Ame | Sugar Land | NULL | 20 |    |
| 283  | Jesse Carb | Colombia | Bogotá | NULL | 20 |    |
| 288  | Weldon Hig | United States of Ame | New York | NULL | 20 |    |
| 302  | Austin Haa | United States of Ame | Columbus | NULL | 20 |    |
| 304  | Zachary Mu | United States of Ame | Los Angele | NULL | 20 |    |
| 14   | Kolbt Stru | United States of Ame | Mission Vl | NULL | 21 |    |
| 27   | Brandon Ba | United States of Ame | South Hill | NULL | 21 |    |
| 32   | Michael Be | United States of Ame | Phoenix | NULL | 21 |    |
| 33   | Sean Rozga | United States of Ame | Los Angele | NULL | 21 |    |
| 52   | Zachary Br | Spain | Madrid | NULL | 21 |    |
| 61   | Colin Lemo | China | Shanghai | NULL | 21 |    |
| 65   | Cameron Ha | Poland | Warsaw | NULL | 21 |    |
| 71   | Joseph Sml | United States of Ame | Sacramento | NULL | 21 |    |
| 75   | Joseph Sni | United States of Ame | Los Angele | NULL | 21 |    |
| 76   | Sourinthon | China | Wuhan | NULL | 21 |    |
| 86   | Myles Vaug | United States of Ame | Waukegan | NULL | 21 |    |
| 87   | Juan Guerr | Mexico | Mexico | NULL | 21 |    |
| 123  | Gareth New | United States of Ame | Los Angele | NULL | 21 |    |
| 124  | Austin Har | United States of Ame | Las Vegas | NULL | 21 |    |
| 130  | Casey Buhr | United States of Ame | Phoenix | NULL | 21 |    |
+-----+-----+-----+-----+-----+-----+-----+

```

Bucketing:

It is a method of dividing data within partitions into fixed-size buckets based on the hash value of a specified column. It helps distribute data evenly, Improving the query performance by reducing the data.

- Create a bucket table and mention how many buckets you need.
- Now insert main table values into bucket table using--

insert into table buckettablename select * from maintablename

```

hive> create table bk1(id int,name varchar(10),nationality varchar(20),city varchar(10),gender varchar(10),age int)clustered by(city) into 3 bucke
ts;
OK
Time taken: 0.152 seconds
hive> Insert into table bk1 select * from student3;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spa
rk, tez) or using Hive 1.X releases.
Query ID = ubh01_20240313193834_92da57ee-472d-4c3c-9f82-0ce2d3a770f3
Total Jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1710337172011_0002, Tracking URL = http://ubh01:8088/proxy/application_1710337172011_0002/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job -kill job_1710337172011_0002
Hadoop job Information for Stage-1: number of mappers: 1; number of reducers: 3
2024-03-13 19:38:44,882 Stage-1 map = 0%, reduce = 0%
2024-03-13 19:38:53,377 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.14 sec
2024-03-13 19:39:08,365 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 10.13 sec
2024-03-13 19:39:12,154 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 16.47 sec
2024-03-13 19:39:14,826 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 23.57 sec
MapReduce Total cumulative CPU time: 23 seconds 570 msec
Ended Job = job_1710337172011_0002
Loading data to table gcp.bk1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 3 Cumulative CPU: 23.57 sec  HDFS Read: 36661 HDFS Write: 14006 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 570 msec
OK
Time taken: 42.604 seconds
hive>

```

Check for bucket creation in hive directory of hdfs:

```

hive> exit;
ubh01@ubh01:~$ hdfs dfs -ls /user/hive/warehouse/gcp.db
24/03/13 19:40:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 5 items
drwxrwxr-x  - ubh01 supergroup          0 2024-03-13 19:37 /user/hive/warehouse/gcp.db/bk
drwxrwxr-x  - ubh01 supergroup          0 2024-03-13 19:39 /user/hive/warehouse/gcp.db/bk1
drwxrwxr-x  - ubh01 supergroup          0 2024-03-13 19:30 /user/hive/warehouse/gcp.db/dp
drwxrwxr-x  - ubh01 supergroup          0 2024-03-13 07:34 /user/hive/warehouse/gcp.db/sp3
drwxrwxr-x  - ubh01 supergroup          0 2024-03-13 07:23 /user/hive/warehouse/gcp.db/student3

```

HBase:

HBase is an open-source, distributed, column-oriented database built on top of the HDFS. It provides real-time read/write access to large datasets, with horizontal scalability and fault tolerance.

- Start Hadoop service by running Hadoop start command
- Then start HBase service by running command – **hbase**

The screenshot shows a terminal window on an Ubuntu desktop environment. The terminal output is as follows:

```
Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Wed 20:49
ubh01@ubh01:~$ ./hadoop-2.7.1/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
24/03/13 20:35:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-namenode-ubh01.out
localhost: starting datanode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-datanode-ubh01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: startingsecondarynamenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-secondarynamenode-ubh01.out
24/03/13 20:36:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-resourcemanager-ubh01.out
localhost: starting nodemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-nodemanager-ubh01.out
ubh01@ubh01:~$ ./hbase-1.1.2/bin/start-hbase.sh
starting master, logging to /home/ubh01/hbase-1.1.2/logs/hbase-ubh01-master-ubh01.out
OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128M; support was removed in 8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128M; support was removed in 8.0
ubh01@ubh01:~$ jps
4016 DataNode
4257 SecondaryNameNode
4423 ResourceManager
3866 NameNode
4570 NodeManager
5035 HMaster
5261 Jps
ubh01@ubh01:~$ hbase
Usage: hbase [<options>] <command> [<args>]
Options:
  --config DTR  Configuration direction to use. Default: /conf
ubh01@ubh01:~$
```

- Now start hbase script by running command – **hbase shell**
- To verify list of tables in hbase, the command is – **list**
- To check status and version of hbase, the commands are – **status** and – **version**
- To verify current user, the command is –**whoami**

The screenshot shows a terminal window on an Ubuntu desktop environment. The terminal output is as follows:

```
Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Wed 20:49
ubh01@ubh01:~$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.slf4j.impl.Log4jLoggerFactory]
2024-03-13 20:37:39,835 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
ava classes where applicable
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.2, rcc2b70cf03e3378800661ec5cab11eb43fafefc, Wed Aug 26 20:11:27 PDT 2015
hbase(main):001:0> list
TABLE
0 row(s) in 0.3380 seconds
=> []
hbase(main):002:0> status
1 servers, 0 dead, 2.0000 average load
hbase(main):003:0> version
1.1.2, rcc2b70cf03e3378800661ec5cab11eb43fafefc, Wed Aug 26 20:11:27 PDT 2015
hbase(main):004:0> whoami
ubh01 (auth:SIMPLE)
    groups: ubh01, adm, cdrom, sudo, dip, plugdev, lpadmin, sambashare
hbase(main):005:0> create 'employee','personal','office'
0 row(s) in 2.3990 seconds
ubh01@ubh01:~$
```

- To create columnar table in HBase,
Syntax:
create 'tablename','columnarfamiliy_name1','columnarfamiliy_name2'
- To get details of a table, the command is –**describe ‘tablename’**
- To insert a value into table,
Syntax:
put ‘tablename’,‘rowkey’,‘columnarfamiliy_name:column_name’,‘value’

```

Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Wed 20:50
ubh01@ubh01: ~
File Edit View Search Terminal Help
hbase(main):005:0> create 'employee','personal','office'
0 row(s) in 2.3990 seconds
=> Hbase::Table - employee
hbase(main):006:0> describe employee
NameError: undefined local variable or method `employee` for #<Object:0x7a3b7122>
hbase(main):007:0> describe 'employee'
Table employee is ENABLED
employee
COLUMN FAMILIES DESCRIPTION
{NAME => 'office', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'personal', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
E => '0'
2 row(s) in 0.1260 seconds
hbase(main):008:0> put 'employee','00001','personal:name','abhi'
0 row(s) in 0.2050 seconds
hbase(main):009:0> put 'employee','00001','personal:rphone','12345'
0 row(s) in 0.1340 seconds
hbase(main):010:0> put 'employee','00001','office:address','land mark 23'
0 row(s) in 0.0340 seconds
hbase(main):011:0> put 'employee','00001','office:phone','78923'
0 row(s) in 0.0380 seconds

```

- After inserting all values into table, to show all values in the table,
The command is – **scan ‘tablename’**

```

Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Wed 20:50
ubh01@ubh01: ~
File Edit View Search Terminal Help
{NAME => 'office', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'personal', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
2 row(s) in 0.1260 seconds
hbase(main):008:0> put 'employee','00001','personal:name','abhi'
0 row(s) in 0.2050 seconds
hbase(main):009:0> put 'employee','00001','personal:rphone','12345'
0 row(s) in 0.1340 seconds
hbase(main):010:0> put 'employee','00001','office:address','land mark 23'
0 row(s) in 0.0340 seconds
hbase(main):011:0> put 'employee','00001','office:phone','78923'
0 row(s) in 0.0380 seconds
hbase(main):012:0> scan 'employee'
ROW
 00001          COLUMN+CELL
                column=office:address, timestamp=1710343013002, value=land mark 23
                column=office:phone, timestamp=1710343049863, value=78923
 00001          COLUMN+CELL
                column=personal:name, timestamp=1710342825139, value=abhi
                column=personal:rphone, timestamp=1710342897227, value=12345
1 row(s) in 0.0880 seconds
hbase(main):013:0>

```

- To retrieve a particular row of a columnar table,
The command is **-get ‘tablename’,’rowkey’**

Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Activities Terminal

File Edit View Search Terminal Help

```
hbase(main):010:0> put 'employee','00001','office:address','land mark 23'
0 row(s) in 0.0340 seconds

hbase(main):011:0> put 'employee','00001','office:phone','78923'
0 row(s) in 0.0380 seconds

hbase(main):012:0> scan 'employee'
ROW                                     COLUMN+CELL
00001          column=office:address, timestamp=1710343013002, value=land mark 23
00001          column=office:phone, timestamp=1710343049863, value=78923
00001          column=personal:name, timestamp=1710342825139, value=abhi
00001          column=personal:rphone, timestamp=1710342897227, value=12345
1 row(s) in 0.0880 seconds

hbase(main):013:0> list
TABLE
employee
1 row(s) in 0.0190 seconds

=> ["employee"]
hbase(main):014:0> get 'employee','00001'
COLUMN          CELL
office:address  timestamp=1710343013002, value=land mark 23
office:phone    timestamp=1710343049863, value=78923
personal:name   timestamp=1710342825139, value=abhi
personal:rphone timestamp=1710342897227, value=12345
4 row(s) in 0.0820 seconds

hbase(main):015:0> set 'employee','office'
```