

## SQOOP COMMAND

script to extract patient info using sqoop into hive table

1.Run the below command to sqoop the data from mysql to hive

```
sqoop import \  
--connect jdbc:mysql://upgraddetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--username student \  
--password STUDENT123 \  
--table patients_information \  
--direct \  
-m 1 \  
--hive-import \  
--create-hive-table \  
--hive-table patients_contacts_info;
```

## SNAPSHOT OF DATA COLLECTED

```
[hadoop@ip-172-31-70-56 ~]$ sqoop import \  
> --connect jdbc:mysql://upgraddetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
> --username student \  
> --password STUDENT123 \  
> --table patients_information \  
> --direct \  
> -m 1 \  
> --hive-import \  
> --create-hive-table \  
> --hive-table patients_contacts_info;  
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.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.Reload4jLoggerFactory]  
2023-07-09 17:15:07,134 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2023-07-09 17:15:07,199 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2023-07-09 17:15:07,210 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override  
2023-07-09 17:15:07,210 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.  
2023-07-09 17:15:07,333 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
2023-07-09 17:15:07,333 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.  
2023-07-09 17:15:07,818 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'patients_information' AS t LIMIT 1  
2023-07-09 17:15:07,852 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'patients_information' AS t LIMIT 1  
2023-07-09 17:15:07,882 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce  
2023-07-09 17:15:10,998 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/8becdb1ba6a42d1f20c624ac4cdec9/patients_information.jar  
2023-07-09 17:15:11,036 INFO manager.DirectMySQLManager: Beginning mysqldump fast path import  
2023-07-09 17:15:11,036 INFO mapreduce.ImportJobBase: Beginning import of patients_information  
2023-07-09 17:15:11,275 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar  
2023-07-09 17:15:12,213 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  
2023-07-09 17:15:13,106 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-70-56.ec2.internal/172.31.70.56:8032  
2023-07-09 17:15:13,371 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-70-56.ec2.internal/172.31.70.56:10200  
2023-07-09 17:15:13,878 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1688922864662_0001  
2023-07-09 17:15:26,223 INFO db.DBInputFormat: Using read committed transaction isolation  
2023-07-09 17:15:27,064 INFO mapreduce.JobSubmitter: number of splits:1  
2023-07-09 17:15:27,336 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1688922864662_0001  
2023-07-09 17:15:27,336 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2023-07-09 17:15:27,527 INFO conf.Configuration: resource-types.xml not found  
2023-07-09 17:15:27,528 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
```

```

2023-07-09 17:15:41,553 INFO mapreduce.Job: Counters: 33
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=299308
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=85
    HDFS: Number of bytes written=230
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=275616
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=2871
    Total vcore-milliseconds taken by all map tasks=2871
    Total megabyte-milliseconds taken by all map tasks=8819712
  Map-Reduce Framework
    Map input records=1
    Map output records=5
    Input split bytes=85
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=51
    CPU time spent (ms)=1170
    Physical memory (bytes) snapshot=328544256
    Virtual memory (bytes) snapshot=4409651200
    Total committed heap usage (bytes)=314572800
    Peak Map Physical memory (bytes)=328544256
    Peak Map Virtual memory (bytes)=4409651200
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=230
2023-07-09 17:15:41,559 INFO mapreduce.ImportJobBase: Transferred 230 bytes in 29.3128 seconds (7.8464 bytes/sec)
2023-07-09 17:15:41,566 INFO mapreduce.ImportJobBase: Retrieved 5 records.
2023-07-09 17:15:41,566 INFO mapreduce.ImportJobBase: Publishing Hive/Hcat import job data to Listeners for table patients_information
2023-07-09 17:15:41,596 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'patients_information' AS t LIMIT 1

```

```

LOAD DATA INPATH 'hdfs://ip-172-31-70-56.ec2.internal:8020/user/hadoop/patients_information' INTO TABLE 'patients_contacts_info'
2023-07-09 17:15:55,074 INFO ql.Driver: Starting task [Stage-0:MOVE] in serial mode
2023-07-09 17:15:55,077 INFO metastore.HiveMetaStoreClient: Closed a connection to metastore, current connections: 0
Loading data to table default.patients_contacts_info
2023-07-09 17:15:55,077 INFO exec.Task: Loading data to table default.patients_contacts_info from hdfs://ip-172-31-70-56.ec2.internal:8020/user/hadoop/patients_informat
ion
2023-07-09 17:15:55,081 INFO metastore.HiveMetaStoreClient: Trying to connect to metastore with URI thrift://ip-172-31-70-56.ec2.internal:9083
2023-07-09 17:15:55,081 INFO metastore.HiveMetaStoreClient: Opened a connection to metastore, current connections: 1
2023-07-09 17:15:55,083 INFO metastore.HiveMetaStoreClient: Connected to metastore.
2023-07-09 17:15:55,084 INFO metastore.RetryingMetaStoreClient: RetryingMetaStoreClient proxy=class org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient ugi=ha
doop (auth:SIMPLE) retries=1 delay=1 lifetime=0
2023-07-09 17:15:55,324 INFO ql.Driver: Starting task [Stage-1:STATS] in serial mode
2023-07-09 17:15:55,324 INFO metastore.HiveMetaStoreClient: Closed a connection to metastore, current connections: 0
2023-07-09 17:15:55,325 INFO stats.BasicStatsTask: Executing stats task
2023-07-09 17:15:55,334 INFO metastore.HiveMetaStoreClient: Trying to connect to metastore with URI thrift://ip-172-31-70-56.ec2.internal:9083
2023-07-09 17:15:55,335 INFO metastore.HiveMetaStoreClient: Opened a connection to metastore, current connections: 1
2023-07-09 17:15:55,336 INFO metastore.HiveMetaStoreClient: Connected to metastore.
2023-07-09 17:15:55,336 INFO metastore.RetryingMetaStoreClient: RetryingMetaStoreClient proxy=class org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient ugi=ha
doop (auth:SIMPLE) retries=1 delay=1 lifetime=0
2023-07-09 17:15:55,397 INFO stats.BasicStatsTask: Table default.patients_contacts_info stats: [numFiles=1, numRows=0, totalSize=230, rawDataSize=0]
2023-07-09 17:15:55,397 INFO ql.Driver: Completed executing command(queryId=hadoop_20230709171554_b1f71310-cf9e-4ab2-b8f1-f0509038223e); Time taken: 0.324 seconds
OK
2023-07-09 17:15:55,397 INFO ql.Driver: OK
2023-07-09 17:15:55,397 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
Time taken: 0.539 seconds
2023-07-09 17:15:55,397 INFO CliDriver: Time taken: 0.539 seconds
2023-07-09 17:15:55,397 INFO conf.HiveConf: Using the default value passed in for log id: ebae8afa-5064-4fd4-bc3e-d155e7539269
2023-07-09 17:15:55,398 INFO session.SessionState: Resetting thread name to main
2023-07-09 17:15:55,398 INFO conf.HiveConf: Using the default value passed in for log id: ebae8afa-5064-4fd4-bc3e-d155e7539269
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
2023-07-09 17:15:55,414 INFO tez.TezSessionPoolManager: Closing tez session if not default: sessionId=ebae8afa-5064-4fd4-bc3e-d155e7539269, queueName=null, user=hadoop,
doAs=true, isOpen=true, isDefault=false
2023-07-09 17:15:55,414 INFO tez.TezSessionState: Closing Tez Session
2023-07-09 17:15:55,415 INFO client.TezClient: Shutting down Tez Session, sessionName=HIVE-ebae8afa-5064-4fd4-bc3e-d155e7539269, applicationId=application_1688922864662
_0002
2023-07-09 17:15:55,438 INFO tez.TezSessionState: Attempting to clean up resources for ebae8afa-5064-4fd4-bc3e-d155e7539269: hdfs://ip-172-31-70-56.ec2.internal:8020/tmp
/hive/hadoop/tez_session_dir/ebae8afa-5064-4fd4-bc3e-d155e7539269-resources; 0 additional files, 0 localized resources
2023-07-09 17:15:55,443 INFO session.SessionState: Deleted directory: /tmp/hive/hadoop/ebae8afa-5064-4fd4-bc3e-d155e7539269 on fs with scheme hdfs
2023-07-09 17:15:55,444 INFO session.SessionState: Deleted directory: /tmp/hadoop/ebae8afa-5064-4fd4-bc3e-d155e7539269 on fs with scheme file
2023-07-09 17:15:55,444 INFO hive.HiveImport: Hive import complete.
2023-07-09 17:15:55,446 INFO hive.HiveClientCommon: Export directory is contains the _SUCCESS file only, removing the directory.
[hadoop@ip-172-31-70-56 ~]$

```

2.Run the below query in hive to check the data has been imported to hive table

Select \* from patients\_contacts\_info;

```
hive> Select * from patients_contacts_info;
OK
1      Alex S  XDC test Address      8982739282      1      23      NULL
2      Sammy A New Building Address  2382739282      2      45      NULL
3      Karan C Aws Address           8923739282      3      56      NULL
4      Dara M  India Address         2182739282      4      67      NULL
5      Pam     ABC test Address       4982739282      5      72      NULL
Time taken: 2.974 seconds, Fetched: 5 row(s)
hive> 
```