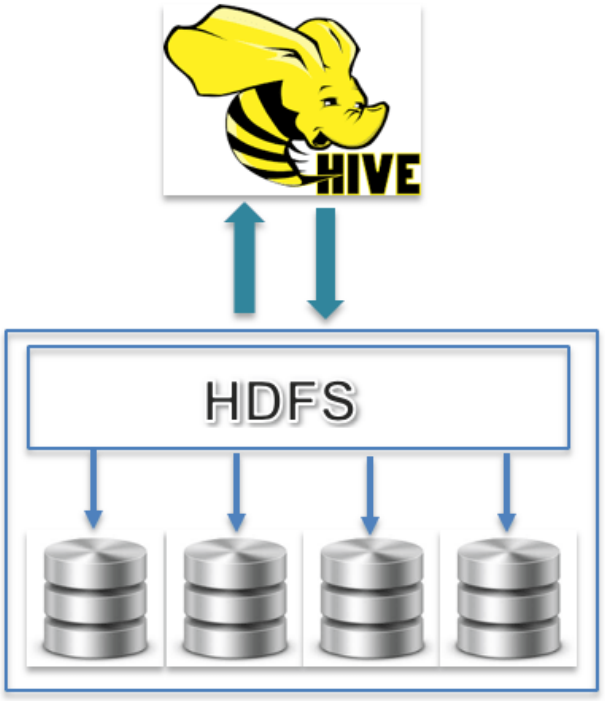
PRÁCTICA Nº 4.3
Sqoop. Jobs
y
otras opciones de configuración



- Nombre y apellidos: Alvaro Lucio-Villegas de Cea

Índice

Ejercicio 1- Réplica de tablas MySQL en HIVE

a) Investiga cómo utilizar la operación “sqoop create-hive-table” para crear una réplica de la tabla copiaempleados de la bbdd MySQL en la bbdd HIVE “mibbddhive” que hemos creado durante las clases. Desarrolla el comando que haga este trabajo.

```
sqoop create-hive-table
--connect jdbc:mysql://localhost/mibd
--username=cloudera
--password=cloudera
--table copiaempleados
--hive-database mibbddhive
--hive-table hivecopiaempleados
--create-hive-table
```

Crear la base de datos en Hive.

```
WARNING: Hive CLI is deprecated and migra
hive> CREATE DATABASE mibbddhive
    > ;
OK
Time taken: 1.559 seconds
hive>
```

Creamos la tabla con las columnas, pero vacía de registros.

```
[cloudera@quickstart ~]$ sqoop create-hive-table --connect jdbc:mysql://localhost/mibd --username=root --password=cloudera  --table copiaempleados  --hive-database mibbddhive --hive-table replica_copiaempleados
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/27 11:18:22 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
23/03/27 11:18:22 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/27 11:18:22 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
23/03/27 11:18:22 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
23/03/27 11:18:23 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/27 11:18:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/27 11:18:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/27 11:18:23 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.13.0.jar!/hive-log4j.properties
OK
Time taken: 1.589 seconds
[cloudera@quickstart ~]$
```

Ahora cargamos todos los registros de la tabla a hive.

```
hive> load data inpath "copiaempleados" into table hivecopiaempleados;
Loading data to table mibbddhive.hivecopiaempleados
Table mibbddhive.hivecopiaempleados stats: [numFiles=4, totalSize=154]
OK
Time taken: 0.434 seconds
hive> select*from hivecopiaempleados;
OK
8       Armando Bronca  24      21000
9       Dolores Fuertes 26      24000
10      Javier Cidoncha 28      25000
11      Lorena  35      28000
12      Miriam  42      30000
13      Pedro   43      25000
14      Juan    45      39000
Time taken: 0.059 seconds, Fetched: 7 row(s)
hive>
```

Ejercicio 2 – Exportación de Actualizaciones

En la máquina Cloudera, en la bbdd MySql que tenemos instalada hay 2 tablas: empleados y copiaempleados. La segunda tabla debe tener 7 registros de la actividad2.

a) Desarrolla una sentencia de exportación que actualice en la tabla copiaempleados posibles cambios en el directorio "/tablashdfs". Sólo registrará actualizaciones en los ficheros. Utiliza la aplicación HUE disponible a partir del explorador en cloudera, para acceder al directorio, y cambiar algunas filas concretas de alguno de los ficheros. Lanza el comando y verifica lo que ocurre.

```
sqoop export –connect jdbc:mysql://localhost/mibd – username=root –password=cloudera –table copiaempleados –export-dir /tablashdfs – update-key id –update-mode updateonly.
```

Salida del comando.

```
        at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/mibd --username=root --password=cloudera --table copiaempleados --export-dir /tablashdfs --update-key id --update-mode updateonly
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/27 12:12:16 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
23/03/27 12:12:16 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/27 12:12:16 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/27 12:12:16 INFO tool.CodeGenTool: Beginning code generation
23/03/27 12:12:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/27 12:12:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/27 12:12:16 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/28bd4519bc7a9d0115f442d5fd19fbf3/copiaempleados.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/27 12:12:17 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/28bd4519bc7a9d0115f442d5fd19fbf3/copiaempleados.jar
23/03/27 12:12:17 INFO mapreduce.ExportJobBase: Beginning export of copiaempleados
23/03/27 12:12:17 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/27 12:12:17 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/27 12:12:18 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
23/03/27 12:12:18 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/27 12:12:18 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/27 12:12:18 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/27 12:12:20 INFO input.FileInputFormat: Total input paths to process : 4
23/03/27 12:12:20 INFO input.FileInputFormat: Total input paths to process : 4
23/03/27 12:12:20 INFO mapreduce.JobSubmitter: number of splits:3
23/03/27 12:12:20 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/27 12:12:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678734946391_0022
23/03/27 12:12:20 INFO impl.YarnClientImpl: Submitted application application_1678734946391_0022
23/03/27 12:12:20 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1678734946391_0022/
23/03/27 12:12:20 INFO mapreduce.Job: Running job: job_1678734946391_0022
23/03/27 12:12:25 INFO mapreduce.Job: Job job_1678734946391_0022 running in uber mode : false
23/03/27 12:12:25 INFO mapreduce.Job:  map 0% reduce 0%
23/03/27 12:12:31 INFO mapreduce.Job:  map 33% reduce 0%
23/03/27 12:12:32 INFO mapreduce.Job:  map 67% reduce 0%
23/03/27 12:12:33 INFO mapreduce.Job:  map 100% reduce 0%
23/03/27 12:12:33 INFO mapreduce.Job: Job job_1678734946391_0022 completed successfully
23/03/27 12:12:33 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=513651
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=847
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
```

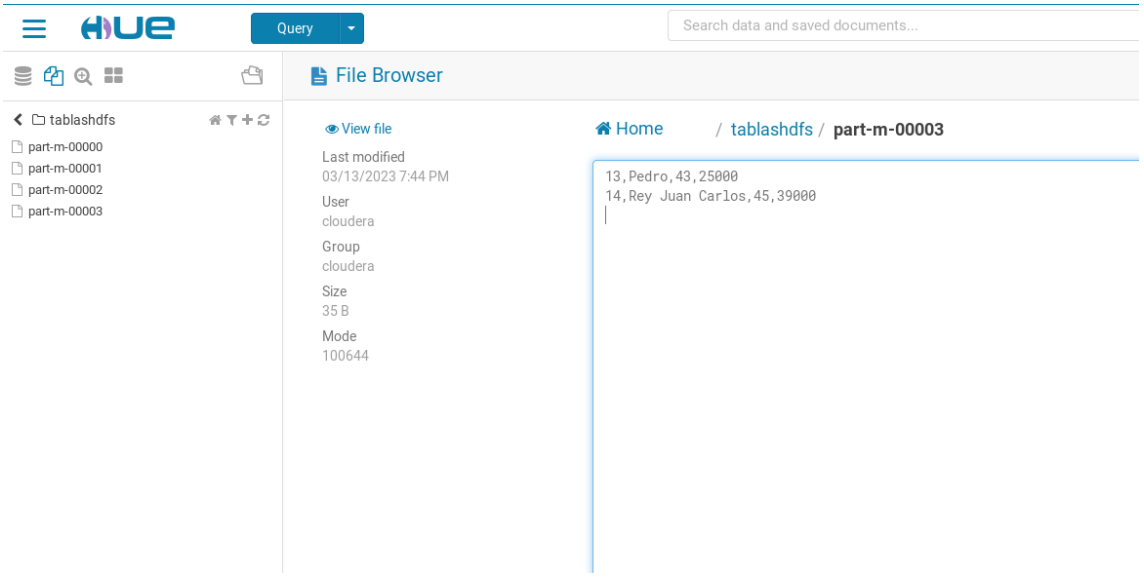
```
23/03/27 12:12:33 INFO mapreduce.Job:  map 100% reduce 0%
23/03/27 12:12:33 INFO mapreduce.Job: Job job_1678734946391_0022 completed successfully
23/03/27 12:12:33 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=513651
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=847
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=3
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=7225
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7225
                Total vcore-milliseconds taken by all map tasks=7225
                Total megabyte-milliseconds taken by all map tasks=7398400
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=624
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=152
                CPU time spent (ms)=1500
                Physical memory (bytes) snapshot=639688704
                Virtual memory (bytes) snapshot=4751138816
                Total committed heap usage (bytes)=819462144
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/27 12:12:33 INFO mapreduce.ExportJobBase: Transferred 847 bytes in 14.8488 seconds (57.0416 bytes/sec)
23/03/27 12:12:33 INFO mapreduce.ExportJobBase: Exported 7 records.
```

b) Crea un JOB que contenga este comando que acabas de desarrollar.

```
[cloudera@quickstart ~]$ sqoop job --create ejercicio2V2 -- export --connect jdbc:mysql://localhost/mibd --username=root --password=cloudera --table copiaempleados --export-dir /tablashdfs --update-key id --update-mode updateonly
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/29 11:49:03 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
23/03/29 11:49:04 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
```

c) Modifica otros registros distintos y lanza el job. Verifica lo que ocurre.

Modificamos desde la interfaz las tablas desde el fichero en HDFS, en este caso cambiamos el registro “Juan” por “Rey Juan Carlos”.



Ejecutamos el job que he creado anteriormente.

```
[cloudera@quickstart ~]$ sqoop job --exec ejercicio2V2
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/29 11:49:17 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
Enter password:
23/03/29 11:49:20 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/29 11:49:20 INFO tool.CodeGenTool: Beginning code generation
23/03/29 11:49:20 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 11:49:20 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 11:49:20 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/bf7313bdaee6288901a9bd98c404244b/copiaempleados.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/29 11:49:21 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/bf7313bdaee6288901a9bd98c404244b/copiaempleados.jar
23/03/29 11:49:21 INFO mapreduce.ExportJobBase: Beginning export of copiaempleados
23/03/29 11:49:21 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/29 11:49:21 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/29 11:49:22 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
23/03/29 11:49:22 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/29 11:49:22 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/29 11:49:22 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/29 11:49:24 INFO input.FileInputFormat: Total input paths to process : 4
23/03/29 11:49:24 INFO input.FileInputFormat: Total input paths to process : 4
23/03/29 11:49:24 INFO mapreduce.JobSubmitter: number of splits:3
23/03/29 11:49:24 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/29 11:49:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678734946391_0035
23/03/29 11:49:24 INFO impl.YarnClientImpl: Submitted application application_1678734946391_0035
23/03/29 11:49:24 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1678734946391_0035/
23/03/29 11:49:24 INFO mapreduce.Job: Running job: job_1678734946391_0035
23/03/29 11:49:30 INFO mapreduce.Job: Job job_1678734946391_0035 running in uber mode : false
23/03/29 11:49:30 INFO mapreduce.Job:  map 0% reduce 0%
23/03/29 11:49:34 INFO mapreduce.Job:  map 33% reduce 0%
23/03/29 11:49:35 INFO mapreduce.Job:  map 67% reduce 0%
23/03/29 11:49:36 INFO mapreduce.Job:  map 100% reduce 0%
23/03/29 11:49:36 INFO mapreduce.Job: Job job_1678734946391_0035 completed successfully
23/03/29 11:49:36 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=514938
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=847
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=21
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=3
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=6680
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6680
                Total vcore-milliseconds taken by all map tasks=6680
                Total megabyte-milliseconds taken by all map tasks=6840320
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=624
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=132
                CPU time spent (ms)=1510
                Physical memory (bytes) snapshot=645779456
                Virtual memory (bytes) snapshot=4732059648
                Total committed heap usage (bytes)=819462144
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/29 11:49:36 INFO mapreduce.ExportJobBase: Transferred 847 bytes in 14.311 seconds (59.1854 bytes/sec)
23/03/29 11:49:36 INFO mapreduce.ExportJobBase: Exported 7 records.
[cloudera@quickstart ~]$
```

Ahora nos dirigimos a SQL para comprobar si se han actualizado bien los datos.

```
mysql> select * from copiaempleados;
+----+-----------------+------+---------+
| id | nombre          | edad | salario |
+----+-----------------+------+---------+
|  8 | Armando Bronca  |   24 |   21000 |
|  9 | Dolores Fuertes |   26 |   24000 |
| 10 | Javier Cidoncha |   28 |   25000 |
| 11 | Lorena          |   35 |   28000 |
| 12 | Miriam          |   42 |   30000 |
| 13 | Pedro           |   43 |   25000 |
| 14 | Rey Juan Carlos |   45 |   39000 |
| 21 | Alvaro Lucio    |   13 |    5000 |
+----+-----------------+------+---------+
8 rows in set (0.00 sec)

mysql>
```

Ejercicio 3 – Exportación de Actualizaciones e Inserciones

Después de nuestras actividades de estos días, el directorio "/tablashdfs" debería tener más filas que las 7 que están en la tabla copiaempleados.

a) Desarrolla una sentencia de exportación que actualice en la tabla copiaempleados posibles cambios en el directorio "/tablashdfs", pero teniendo en cuenta actualizaciones en los datos disponibles y nuevas inserciones. Utiliza la aplicación HUE disponible a partir del explorador en cloudera, para acceder al directorio, y cambiar algunas filas concretas de las que ya existían en las 7 filas de copiaempleados. Lanza el comando y verifica que se han actualizado las filas ya existentes y que se han cargado las nuevas filas.

```
sqoop export –connect jdbc:mysql://localhost/mibd – username=root
–password=cloudera –table copiaempleados –export-dir /tablashdfs – update-key id
–update-mode allowinsert.
```

b) Crea un JOB que contenga este comando que acabas de desarrollar

```
[cloudera@quickstart ~]$ sqoop job --create ejercicio3 -- export --connect jdbc:mysql://localhost/mibd --username=root --password=cloudera --table copiaempleados --export-dir /tablashdfs --update-key id --update-mode allowinsert
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/29 11:57:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
23/03/29 11:57:33 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
[cloudera@quickstart ~]$
```

c) Realiza nuevas modificaciones en los registros de los ficheros del directorio "/tablashdfs".Añade nuevos registros a la tabla empleados y cargarlos en el directorio "/tablashdfs" mediante el job que ye debes tener creado durante las clases. Lanza el job creado en b) y verifique lo que ocurre.

Vemos el estado original de la tabla en SQL para después compararlo con el resultado final.

```
mysql> select * from copiaempleados;
+----+-----------------+------+---------+
| id | nombre          | edad | salario |
+----+-----------------+------+---------+
|  8 | Armando Bronca  |   24 |   21000 |
|  9 | Dolores Fuertes |   26 |   24000 |
| 10 | Javier Cidoncha |   28 |   25000 |
| 11 | Lorena          |   35 |   28000 |
| 12 | Miriam          |   42 |   30000 |
| 13 | Pedro           |   43 |   25000 |
| 14 | Rey Juan Carlos |   45 |   39000 |
+----+-----------------+------+---------+
7 rows in set (0.00 sec)

mysql>
```

Modificamos la tabla desde la interfaz de HDFS y agregamos un registro nuevo a la tabla.

Ahora ejecutamos el Job que creamos anteriormente, y el resultado esperado sería que se agregase el nuevo registro a la tabla en SQL.

```
[cloudera@quickstart ~]$ sqoop job --exec ejercicio3
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/29 12:04:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
Enter password:
23/03/29 12:04:49 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/29 12:04:49 INFO tool.CodeGenTool: Beginning code generation
23/03/29 12:04:50 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 12:04:50 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 12:04:50 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/d9e07c7cb78a72c7ed46558554722577/copiaempleados.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/29 12:04:50 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/d9e07c7cb78a72c7ed46558554722577/copiaempleados.jar
23/03/29 12:04:50 WARN manager.MySQLManager: MySQL Connector upsert functionality is using INSERT ON
23/03/29 12:04:50 WARN manager.MySQLManager: DUPLICATE KEY UPDATE clause that relies on table's unique key.
23/03/29 12:04:50 WARN manager.MySQLManager: Insert/update distinction is therefore independent on column
23/03/29 12:04:50 WARN manager.MySQLManager: names specified in --update-key parameter. Please see MySQL
23/03/29 12:04:50 WARN manager.MySQLManager: documentation for additional limitations.
23/03/29 12:04:50 INFO mapreduce.ExportJobBase: Beginning export of copiaempleados
23/03/29 12:04:50 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/29 12:04:51 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/29 12:04:51 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
23/03/29 12:04:51 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/29 12:04:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/29 12:04:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/29 12:04:53 INFO input.FileInputFormat: Total input paths to process : 4
23/03/29 12:04:53 INFO input.FileInputFormat: Total input paths to process : 4
23/03/29 12:04:53 INFO mapreduce.JobSubmitter: number of splits:4
23/03/29 12:04:53 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
23/03/29 12:04:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678734946391_0037
23/03/29 12:04:54 INFO impl.YarnClientImpl: Submitted application application_1678734946391_0037
23/03/29 12:04:54 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1678734946391_0037/
23/03/29 12:04:54 INFO mapreduce.Job: Running job: job_1678734946391_0037
23/03/29 12:05:00 INFO mapreduce.Job: Job job_1678734946391_0037 running in uber mode : false
23/03/29 12:05:00 INFO mapreduce.Job:  map 0% reduce 0%
23/03/29 12:05:04 INFO mapreduce.Job:  map 25% reduce 0%
23/03/29 12:05:05 INFO mapreduce.Job:  map 50% reduce 0%
23/03/29 12:05:06 INFO mapreduce.Job:  map 75% reduce 0%
23/03/29 12:05:07 INFO mapreduce.Job:  map 100% reduce 0%
23/03/29 12:05:07 INFO mapreduce.Job: Job job_1678734946391_0037 completed successfully
23/03/29 12:05:07 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=686652
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=959
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=22
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Other local map tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=9520
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=9520
                Total vcore-milliseconds taken by all map tasks=9520
                Total megabyte-milliseconds taken by all map tasks=9748480
        Map-Reduce Framework
                Map input records=8
                Map output records=8
                Input split bytes=688
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=146
                CPU time spent (ms)=2290
                Physical memory (bytes) snapshot=856178688
                Virtual memory (bytes) snapshot=6331265024
                Total committed heap usage (bytes)=1011875840
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/29 12:05:07 INFO mapreduce.ExportJobBase: Transferred 959 bytes in 15.8292 seconds (60.5841 bytes/sec)
23/03/29 12:05:07 INFO mapreduce.ExportJobBase: Exported 8 records.
[cloudera@quickstart ~]$
```

part-m-00000 (~/Downloads) - gedit

Una vez finalizado el comando nos dirigimos a la tabla en SQL para verificar si se han realizado los cambios correspondientes.

```
mysql> select * from copiaempleados;
+----+-----------------+------+---------+
| id | nombre          | edad | salario |
+----+-----------------+------+---------+
|  8 | Armando Bronca  |   24 |   21000 |
|  9 | Dolores Fuertes |   26 |   24000 |
| 10 | Javier Cidoncha |   28 |   25000 |
| 11 | Lorena          |   35 |   28000 |
| 12 | Miriam          |   42 |   30000 |
| 13 | Pedro           |   43 |   25000 |
| 14 | Rey Juan Carlos |   45 |   39000 |
| 21 | Alvaro Lucio    |   13 |   42000 |
+----+-----------------+------+---------+
8 rows in set (0.00 sec)

mysql>
```

Ejercicio 4 – Cargar datos en tabla HIVE

a) Desarrolla un comando Sqoop para cargar en la tabla copiaempleados de HIVE las filas que ya tienes en la tabla copiaempleados de MySQL. Ten en cuenta que, en este comando, no debe crearse la tabla HIVE, como hemos hecho en teoría, porque la tabla ya existe.

```
sqoop import
--connect jdbc:mysql://localhost/mibd

--username=cloudera

--password=cloudera

--table copiaempleados

--hive-database mibbddhive

--hive-import
--hive-table hivecopiaempleados
--hive-overwrite
--hive-overwrite
```

Ejecutamos el comando y observamos la salida.

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/mibd --username=root --password=cloudera --table copiaempleados --hive-database mibbddhive --hive-import --hive-table hivecopiaempleados --hive-overwrite
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
23/03/29 12:23:35 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
23/03/29 12:23:35 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/29 12:23:35 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
23/03/29 12:23:35 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
23/03/29 12:23:35 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/29 12:23:35 INFO tool.CodeGenTool: Beginning code generation
23/03/29 12:23:35 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 12:23:35 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 12:23:35 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/51ec1591441e30321472f385f8ed4cdf/copiaempleados.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/29 12:23:36 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/51ec1591441e30321472f385f8ed4cdf/copiaempleados.jar
23/03/29 12:23:36 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/03/29 12:23:36 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/03/29 12:23:36 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/03/29 12:23:36 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/03/29 12:23:36 INFO mapreduce.ImportJobBase: Beginning import of copiaempleados
23/03/29 12:23:36 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/03/29 12:23:36 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/03/29 12:23:37 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/03/29 12:23:37 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/29 12:23:40 INFO db.DBInputFormat: Using read commited transaction isolation
23/03/29 12:23:40 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`id`), MAX(`id`) FROM `copiaempleados`
23/03/29 12:23:40 INFO db.IntegerSplitter: Split size: 3; Num splits: 4 from: 8 to: 21
23/03/29 12:23:40 INFO mapreduce.JobSubmitter: number of splits:4
23/03/29 12:23:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678734946391_0039
23/03/29 12:23:40 INFO impl.YarnClientImpl: Submitted application application_1678734946391_0039
23/03/29 12:23:40 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1678734946391_0039/
23/03/29 12:23:40 INFO mapreduce.Job: Running job: job_1678734946391_0039
23/03/29 12:23:46 INFO mapreduce.Job: Job job_1678734946391_0039 running in uber mode : false
23/03/29 12:23:46 INFO mapreduce.Job:  map 0% reduce 0%
23/03/29 12:23:50 INFO mapreduce.Job:  map 25% reduce 0%
23/03/29 12:23:51 INFO mapreduce.Job:  map 50% reduce 0%
23/03/29 12:23:52 INFO mapreduce.Job:  map 75% reduce 0%
23/03/29 12:23:53 INFO mapreduce.Job:  map 100% reduce 0%
23/03/29 12:23:54 INFO mapreduce.Job: Job job_1678734946391_0039 completed successfully
23/03/29 12:23:55 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=685940
                FILE: Number of read operations=0
                FILE: 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)=10040
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=10040
        Total vcore-milliseconds taken by all map tasks=10040
        Total megabyte-milliseconds taken by all map tasks=10280960
    Map-Reduce Framework
        Map input records=8
        Map output records=8
        Input split bytes=400
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=196
        CPU time spent (ms)=2700
        Physical memory (bytes) snapshot=874487808
        Virtual memory (bytes) snapshot=6316122112
        Total committed heap usage (bytes)=1092616192
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=190
23/03/29 12:23:55 INFO mapreduce.ImportJobBase: Transferred 190 bytes in 17.7811 seconds (10.6855 bytes/sec)
23/03/29 12:23:55 INFO mapreduce.ImportJobBase: Retrieved 8 records.
23/03/29 12:23:55 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `copiaempleados` AS t LIMIT 1
23/03/29 12:23:55 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.13.0.jar!/hive-log4j.properties
OK
Time taken: 0.859 seconds
Loading data to table mibbddhive.hivecopiaempleados
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/mibbddhive.db/hivecopiaempleados': User does not belong to supergroup
Table mibbddhive.hivecopiaempleados stats: [numFiles=4, numRows=0, totalSize=190, rawDataSize=0]
OK
Time taken: 0.626 seconds
[cloudera@quickstart ~]$
```

Para comprobar que se ha realizado correctamente la importación sin creación, nos dirigimos a Hive y visualizamos la tabla antes de lanzar el job y después de ejecutarlo .

```
Time taken: 0.058 seconds, Fetched: 7 row(s)
hive> select * from hivecopiaempleados;
OK
8       Armando Bronca  24      21000
9       Dolores Fuertes 26      24000
10      Javier Cidoncha 28      25000
11      Lorena  35      28000
12      Miriam  42      30000
13      Pedro   43      25000
14      Juan    45      39000
Time taken: 0.073 seconds, Fetched: 7 row(s)
hive> select * from hivecopiaempleados;
OK
8       Armando Bronca  24      21000
9       Dolores Fuertes 26      24000
10      Javier Cidoncha 28      25000
11      Lorena  35      28000
12      Miriam  42      30000
13      Pedro   43      25000
14      Rey Juan Carlos 45      39000
21      Alvaro Lucio    13      42000
Time taken: 0.066 seconds, Fetched: 8 row(s)
hive>
```