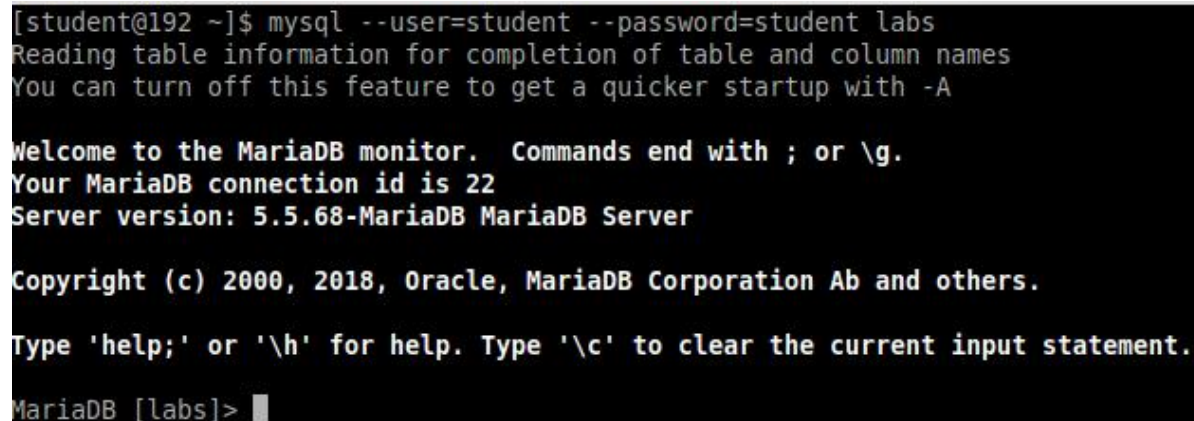


Lab 1: Data Ingestion with Sqoop for RDBMS (MariaDB)

1. In a terminal window, log in to MariaDB and select the database labs.

Database: labs

```
$ mysql --user=student --password=student labs
```



```
[student@192 ~]$ mysql --user=student --password=student labs
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 22
Server version: 5.5.68-MariaDB MariaDB Server

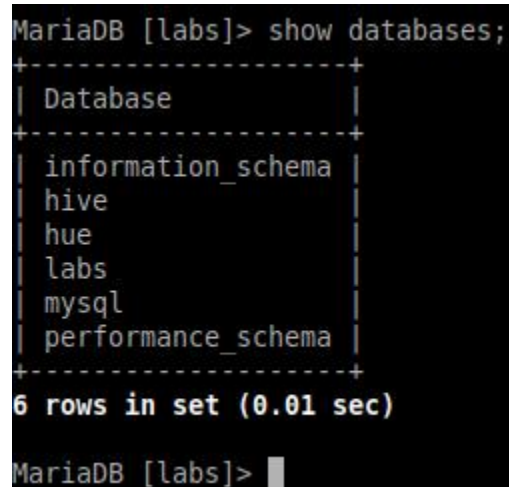
Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

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

MariaDB [labs]> █
```

2. If the login is successful, the "MariaDB [labs]>" prompt appears and a screen waiting for commands is displayed. Enter a command to check which database exists here.

MariaDB> show databases;



```
MariaDB [labs]> show databases;
+-----+
| Database                |
+-----+
| information_schema       |
| hive                     |
| hue                      |
| labs                     |
| mysql                    |
| performance_schema       |
+-----+
6 rows in set (0.01 sec)

MariaDB [labs]> █
```

3. Next, enter the command to review the table in labs.

```

Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 22
Server version: 5.5.68-MariaDB MariaDB Server

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and
Type 'help;' or '\h' for help. Type '\c' to clear the current
MariaDB [labs]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| hive          |
| hue          |
| labs         |
| mysql        |
| performance_schema |
+-----+
6 rows in set (0.01 sec)

MariaDB [labs]> show tables;
+-----+
| Tables_in_labs |
+-----+
| authors        |
| authors_export |
| posts         |
+-----+
3 rows in set (0.00 sec)

MariaDB [labs]> 

```

```

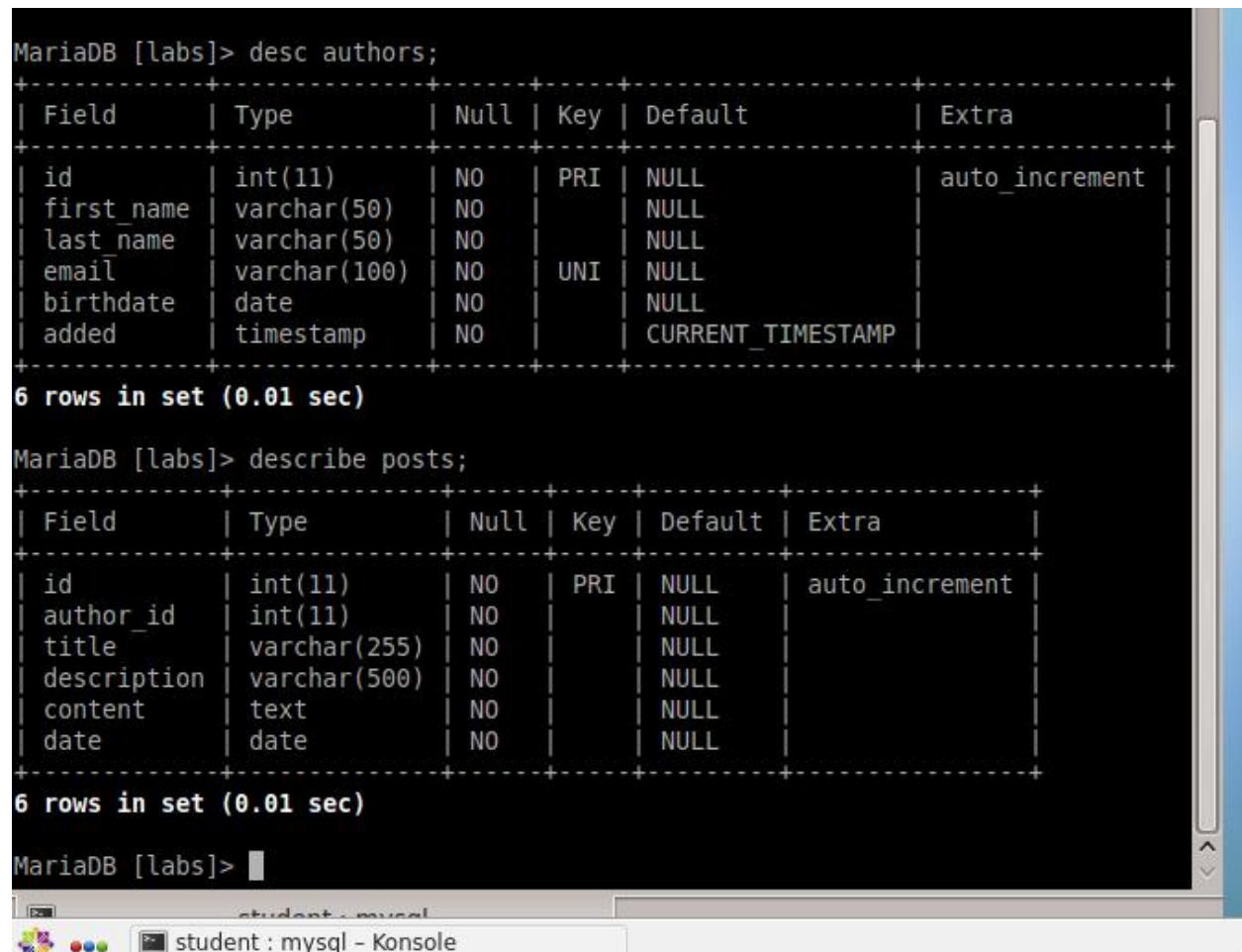
MariaDB> desc authors;
MariaDB> describe posts;

```

```
MariaDB [labs]> desc authors;
+-----+-----+-----+-----+-----+-----+
| Field      | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| id         | int(11)   | NO   | PRI | NULL         | auto_increment |
| first_name | varchar(50)| NO   |     | NULL         |               |
| last_name  | varchar(50)| NO   |     | NULL         |               |
| email      | varchar(100)| NO   | UNI | NULL         |               |
| birthdate  | date      | NO   |     | NULL         |               |
| added      | timestamp | NO   |     | CURRENT_TIMESTAMP |               |
+-----+-----+-----+-----+-----+-----+
6 rows in set (0.01 sec)

MariaDB [labs]> describe posts;
+-----+-----+-----+-----+-----+-----+
| Field      | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| id         | int(11)   | NO   | PRI | NULL         | auto_increment |
| author_id  | int(11)   | NO   |     | NULL         |               |
| title      | varchar(255)| NO   |     | NULL         |               |
| description | varchar(500)| NO   |     | NULL         |               |
| content    | text      | NO   |     | NULL         |               |
| date       | date      | NO   |     | NULL         |               |
+-----+-----+-----+-----+-----+-----+
6 rows in set (0.01 sec)

MariaDB [labs]> 
```



4. Review the structure of the authors, posts tables and review some records.

```
MariaDB> SELECT id, first_name, last_name, email, added
FROM authors limit 5;
```

```
MariaDB [labs]> SELECT id, first_name, last_name, email, added
-> FROM authors limit 5;
+-----+-----+-----+-----+-----+
| id | first_name | last_name | email | added |
+-----+-----+-----+-----+-----+
| 1 | Walton | Adams | barmstrong@example.com | 1997-01-02 04:18: |
| 2 | Marietta | Walsh | hand.stella@example.net | 2010-08-26 18:20: |
| 3 | Lily | Wintheiser | darren.blanda@example.org | 1973-06-11 07:28: |
| 4 | Estevan | Gleason | shanahan.aliyah@example.net | 1995-01-29 16:08: |
| 5 | Thaddeus | Rowe | bednar.robin@example.net | 2017-01-05 04:13: |
+-----+-----+-----+-----+-----+
5 rows in set (0.01 sec)

MariaDB [labs]> 
```

5. Type quit to exit MariaDB and press Enter.

MariaDB> quit

```
MariaDB [labs]> quit
Bye
[student@192 ~]$ 
```

6. Run the following command to check the basic options of sqoop.

\$sqoop help

```
[student@192 ~]$ sqoop help
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 00:18:42,452 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
usage: sqoop COMMAND [ARGS]

Available commands:
  codegen          Generate code to interact with database records
  create-hive-table Import a table definition into Hive
  eval            Evaluate a SQL statement and display the results
  export          Export an HDFS directory to a database table
  help            List available commands
  import          Import a table from a database to HDFS
  import-all-tables Import tables from a database to HDFS
  import-mainframe Import datasets from a mainframe server to HDFS
  job             Work with saved jobs
  list-databases  List available databases on a server
  list-tables     List available tables in a database
  merge          Merge results of incremental imports
  metastore       Run a standalone Sqoop metastore
  version         Display version information

See 'sqoop help COMMAND' for information on a specific command.
[student@192 ~]$
```

7. To see detailed options for each sub-command, enter the desired subcommand after help. To see detailed options for import, run the command as follows.

`$sqoop help import`

```
[student@192 ~]$ sqoop help import
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 00:19:56,071 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
usage: sqoop import [GENERIC-ARGS] [TOOL-ARGS]

Common arguments:
  --connect <jdbc-uri>          Specify JDBC
                                connect
                                string
  --connection-manager <class-name> Specify
                                connection
                                manager
                                class name
  --connection-param-file <properties-file> Specify
                                connection
                                parameters
                                file
```

8. Run the list of databases in MariaDB and tables in database labs with the following command.

`$sqoop list-databases --connect jdbc:mysql://localhost --username student - password student`

`$sqoop list-tables --connect jdbc:mysql://localhost/labs --username student -P`


```

... / more
[student@192 ~]$ sqoop list-tables --connect jdbc:mysql://localhost/labs --username student -P
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 00:46:52,381 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2024-06-21 00:46:55,262 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
authors
authors_export
posts
[student@192 ~]$

```

9. Import all tables in labs database using the import-all-tables command.

`$sqoop import-all-tables --connect jdbc:mysql://localhost/labs \`
`--username student --password student`

```

[student@192 ~]$ sqoop import-all-tables --connect jdbc:mysql://localhost/labs \
> --username student --password student
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 00:49:19,972 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 00:49:20,087 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 00:49:20,302 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 00:49:20,898 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 00:49:20,919 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
MIT 1
2024-06-21 00:49:21,025 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
MIT 1
2024-06-21 00:49:21,042 INFO orm.CompilationManager: HADOOP MAPRED HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/838c383fe6b699a8ce6b5ebdfc78e4d6/authors.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-06-21 00:49:26,837 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/838c383fe6b699a8ce6b5ebdfc78e4d6/authors.jar
2024-06-21 00:49:26,862 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-06-21 00:49:26,862 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-06-21 00:49:26,862 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-06-21 00:49:26,863 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)

```

```

Total time spent by all maps in occupied slots (ms)=65924
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=65924
Total vcore-milliseconds taken by all map tasks=65924
Total megabyte-milliseconds taken by all map tasks=67506176
Map-Reduce Framework
  Map input records=10000
  Map output records=10000
  Input split bytes=415
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2764
  CPU time spent (ms)=23580
  Physical memory (bytes) snapshot=1086586880
  Virtual memory (bytes) snapshot=11257503744
  Total committed heap usage (bytes)=736100352
  Peak Map Physical memory (bytes)=274432000
  Peak Map Virtual memory (bytes)=2817224704
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=760827
2024-06-21 00:50:29,435 INFO mapreduce.ImportJobBase: Transferred 742.9951 KB in 60.8961 seconds (12.201 KB/sec)
2024-06-21 00:50:29,441 INFO mapreduce.ImportJobBase: Retrieved 10000 records.
2024-06-21 00:50:29,441 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 00:50:29,481 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors_export` AS t LIMIT 1
2024-06-21 00:50:29,520 INFO orm.CompilationManager: HADOOP MAPRED HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/838c383fe6b699a8ce6b5ebdfc78e4d6/authors_export.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-06-21 00:50:30,470 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/838c383fe6b699a8ce6b5ebdfc78e4d6/authors_export.jar
2024-06-21 00:50:30,479 ERROR tool.ImportAllTablesTool: Error during import: No primary key could be found for table authors_export. Please specify one with --split-by or perform a sequential import with '-m 1'.
[student@192 ~]$

```

10. Execute the command to fetch the posts table from the labs database using Sqoop and store it in HDFS.

```
$ sqoop import --connect jdbc:mysql://localhost/labs \
--username student --password student --table posts
```

```

[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs \
> --username student --password student --table posts
Warning: /usr/local/sqoop/sqoop-1.4.7/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 00:51:18,661 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 00:51:18,776 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 00:51:19,015 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 00:51:19,015 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 00:51:19,585 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `posts` AS t LIMIT 1
2024-06-21 00:51:19,739 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `posts` AS t LIMIT 1
2024-06-21 00:51:19,784 INFO orm.CompilationManager: HADOOP MAPRED HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/5bd922eabc2588c8db5c1b4f6b291dd4/posts.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-06-21 00:51:22,863 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/5bd922eabc2588c8db5c1b4f6b291dd4/posts.jar
2024-06-21 00:51:22,885 WARN manager.MySQLManager: It looks like you are importing from mysql.

```

When this command is executed, the posts directory is created under the /user/student home directory of HDFS and data is stored as follows

```

2024-06-21 00:52:07,764 INFO mapreduce.ImportJobBase: Transferred 39.1174 MB in 43.8126 seconds (914.2613 KB/sec)
2024-06-21 00:52:07,775 INFO mapreduce.ImportJobBase: Retrieved 110000 records.
[student@192 ~]$ hdfs dfs -ls /user/student/posts
Found 5 items
-rw-r--r--  1 student student      0 2024-06-21 00:52 /user/student/posts/_SUCCESS
-rw-r--r--  1 student student 10247865 2024-06-21 00:51 /user/student/posts/part-m-00000
-rw-r--r--  1 student student 10260760 2024-06-21 00:52 /user/student/posts/part-m-00001
-rw-r--r--  1 student student 10257979 2024-06-21 00:52 /user/student/posts/part-m-00002
-rw-r--r--  1 student student 10250911 2024-06-21 00:52 /user/student/posts/part-m-00003
[student@192 ~]$

```

11. Create a target directory in HDFS to import table data into.

\$hdfs dfs -mkdir /mywarehouse

```

[student@192 ~]$ hdfs dfs -mkdir /mywarehouse
[student@192 ~]$

```

12. Import the authors table and save it to the HDFS directory we created above using ',' to delimit the fields.

```

$ sqoop import --connect jdbc:mysql://localhost/labs \
--username student --password student \
--table authors --fields-terminated-by ',' \
--target-dir /mywarehouse/authors

```



```
[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs \
> --username student --password student \
> --table authors --fields-terminated-by ',' \
> --target-dir /mywarehouse/authors
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER HOME to the root of your Zookeeper installation.
2024-06-21 00:56:31,074 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 00:56:31,211 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 00:56:31,567 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 00:56:31,074 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 00:56:32,225 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 00:56:32,343 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 00:56:32,365 INFO orm.CompilationManager: HADOOP MAPRED HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/a4f09faed19ee5e339bd86659ed2c4e0/authors.java uses or overrides a deprecated API.
Note: Reccompile with -Xlint:deprecation for details.
2024-06-21 00:56:35,437 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/a4f09faed19ee5e339bd86659ed2c4e0/authors.jar
2024-06-21 00:56:35,471 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-06-21 00:56:35,471 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
```

```
FILE: Number of write operations=0
HDFS: Number of bytes read=415
HDFS: Number of bytes written=760827
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=4
  Other local map tasks=4
  Total time spent by all maps in occupied slots (ms)=63597
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=63597
  Total vcore-milliseconds taken by all map tasks=63597
  Total megabyte-milliseconds taken by all map tasks=65123328
Map-Reduce Framework
  Map input records=10000
  Map output records=10000
  Input split bytes=415
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2504
  CPU time spent (ms)=25750
  Physical memory (bytes) snapshot=1095610368
  Virtual memory (bytes) snapshot=11270905856
  Total committed heap usage (bytes)=745537536
  Peak Map Physical memory (bytes)=276344832
  Peak Map Virtual memory (bytes)=2818957312
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=760827
2024-06-21 00:57:26,869 INFO mapreduce.ImportJobBase: Transferred 742.9951 KB in 49.7603 seconds (14.9315 KB/sec)
2024-06-21 00:57:26,881 INFO mapreduce.ImportJobBase: Retrieved 10000 records.
[student@192 ~]$
```

13. Review that the command worked with hdfs commands for target-dir.

```
$ hdfs dfs -ls /mywarehouse/authors
$ hdfs dfs -cat /mywarehouse/authors/part-m-00000
```

```
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors
Found 5 items
-rw-r--r-- 1 student supergroup          0 2024-06-21 00:57 /mywarehouse/authors/ SUCCESS
-rw-r--r-- 1 student supergroup    189526 2024-06-21 00:57 /mywarehouse/authors/part-m-00000
-rw-r--r-- 1 student supergroup    190441 2024-06-21 00:57 /mywarehouse/authors/part-m-00001
-rw-r--r-- 1 student supergroup    190481 2024-06-21 00:57 /mywarehouse/authors/part-m-00002
-rw-r--r-- 1 student supergroup    190379 2024-06-21 00:57 /mywarehouse/authors/part-m-00003
[student@192 ~]$
```

```
2464,Malika,Raynor,kaitlyn.effertz@example.net,2014-05-13,2011-02-26 06:01:50.0
2465,Alysha,Kris,mina42@example.com,1985-09-29,2017-10-23 23:37:25.0
2466,Jammie,D'Amore,hosinski@example.org,1982-08-16,1994-01-07 13:15:21.0
2467,Kaci,DuBuque,huels.augustine@example.com,1998-11-22,2002-12-31 14:16:03.0
2468,Chasity,Greenfelder,ervin61@example.org,1980-02-29,1993-06-01 02:22:58.0
2469,Talon,Murray,april.mann@example.net,1976-11-16,1982-07-16 08:25:01.0
2470,Frederique,Leuschke,jyundt@example.com,2012-11-11,1975-06-05 04:45:32.0
2471,Willie,Terry,elsie50@example.org,2014-05-03,1993-03-25 21:20:35.0
2472,Alexis,Lindgren,maudie84@example.org,2012-06-26,1980-07-29 09:30:09.0
2473,Nina,Fadel,mohr.lukas@example.com,1987-03-30,2007-03-03 22:57:06.0
2474,Mohamed,Mills,yundt.marisa@example.net,1979-12-27,1996-02-02 12:44:29.0
2475,Jammie,Wiza,tomas.konopelski@example.org,1970-09-19,1982-02-07 04:31:06.0
2476,Khalid,Brekke,stracke.ivah@example.net,1991-10-31,2004-01-06 20:07:46.0
2477,Abdullah,Olson,stephanie72@example.com,2004-09-26,1988-03-08 07:45:09.0
2478,Laurie,Hammes,nharber@example.org,1999-07-14,1976-10-25 05:45:55.0
2479,Mittie,Hoeger,jonatan.swift@example.org,1975-12-24,1980-05-25 10:54:08.0
2480,Travis,Smitham,peggie.dickinson@example.com,1999-05-26,1982-02-04 22:53:54.0
2481,Owen,Walsh,abdiel.frami@example.net,2015-03-30,2009-09-12 07:18:24.0
2482,Adam,Keeling,ottis45@example.org,1999-08-30,1971-07-24 09:36:25.0
2483,Gordon,Thompson,heidi.huel@example.org,1984-09-26,1973-09-15 06:25:53.0
2484,Dorthy,Hermann,cebert@example.com,1999-06-14,2017-10-31 07:50:47.0
2485,Dan,Green,moore.maxime@example.com,1987-05-09,1991-08-10 21:28:27.0
2486,Ross,Bergnaum,celine15@example.org,1998-12-12,2016-05-22 12:46:20.0
2487,Alexie,Goldner,leonie.hansen@example.org,1982-06-13,1970-03-24 18:30:10.0
2488,Raymundo,Hirthe,sigurd.marks@example.net,2017-10-14,1988-04-15 13:59:39.0
2489,Alta,Thiel,heidenreich.dshaun@example.com,1984-11-09,1975-11-08 05:58:36.0
2490,Alexys,Ferry,sam.grant@example.com,1978-07-16,1979-10-30 11:39:53.0
2491,Andreane,Lind,johanna52@example.org,1980-07-17,1994-09-24 23:00:06.0
2492,Hilario,Sipes,zbraun@example.org,2004-12-23,2002-01-18 13:10:49.0
2493,Holden,Reinger,nolan.christ@example.com,1992-11-29,2014-09-18 21:23:36.0
2494,Jarred,Skiles,millie.mayert@example.com,1990-05-29,1986-05-08 00:50:33.0
2495,Miller,Schumm,uschuppe@example.net,1976-06-16,1995-09-28 08:11:23.0
2496,Candida,Hamill,fvolkman@example.org,1990-08-17,1999-05-08 20:17:54.0
2497,Edmond,Johns,beatty.helga@example.com,2003-03-18,1990-04-17 07:46:01.0
2498,Prince,Bartoletti,oswald76@example.com,1972-07-01,2006-04-08 06:57:59.0
2499,Maya,Quitzon,eldon.flatley@example.org,2006-05-04,1974-11-11 21:57:16.0
2500,Alexandre,Ratke,turner.cornelius@example.com,2012-02-04,2012-03-24 20:01:04.0
[student@192 ~]$
```

student : bash

student : bash - Konsole

14. Import the only specified columns with `--columns` for authors in hdfs home directory. The imported columns are `first_name`, `last_name`, `email`.

```
$ sqoop import --connect jdbc:mysql://localhost/labs --username student --password student --table authors --fields-terminated-by '\t' --columns "first_name, last_name,
```


email

```
[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs \
> --username student --password student \
> --table authors --fields-terminated-by '\t' \
> --columns "first_name,last_name,email"
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 01:02:27,105 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 01:02:27,243 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 01:02:27,581 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 01:02:27,582 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 01:02:28,346 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:02:28,469 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:02:28,509 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/03464fca705421e4670d210975578b72/authors.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
```

```
FILE: Number of write operations=0
HDFS: Number of bytes read=415
HDFS: Number of bytes written=381933
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
HDFS: Number of bytes read erasure-coded=0
Job Counters
  Killed map tasks=1
  Launched map tasks=4
  Other local map tasks=4
  Total time spent by all maps in occupied slots (ms)=60895
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=60895
  Total vcore-milliseconds taken by all map tasks=60895
  Total megabyte-milliseconds taken by all map tasks=62356480
Map-Reduce Framework
  Map input records=10000
  Map output records=10000
  Input split bytes=415
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2284
  CPU time spent (ms)=18510
  Physical memory (bytes) snapshot=1056452608
  Virtual memory (bytes) snapshot=11237179392
  Total committed heap usage (bytes)=750780416
  Peak Map Physical memory (bytes)=265682944
  Peak Map Virtual memory (bytes)=2812506112
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=381933
2024-06-21 01:08:35,838 INFO mapreduce.ImportJobBase: Transferred 372.9814 KB in 44.4735 seconds (8.3866 KB/sec)
2024-06-21 01:08:35,855 INFO mapreduce.ImportJobBase: Retrieved 10000 records.
[student@192 ~]$
```

Result in

```

2024-06-21 01:08:35,838 INFO mapreduce.ImportJobBase: Transferred 372.9
/sec)
2024-06-21 01:08:35,855 INFO mapreduce.ImportJobBase: Retrieved 10000 r
[student@192 ~]$ hdfs dfs -tail authors/part-m-00000
ed Mills yundt.marisa@example.net
Jammie Wiza tomas.konopelski@example.org
Khalid Brekke stracke.ivah@example.net
Abdullah Olson stephanie72@example.com
Laurie Hammes nharber@example.org
Mittie Hoeger jonatan.swift@example.org
Travis Smitham peggie.dickinson@example.com
Owen Walsh abdiel.frami@example.net
Adam Keeling ottis45@example.org
Gordon Thompson heidi.huel@example.org
Dorthy Hermann cebert@example.com
Dan Green moore.maxime@example.com
Ross Bergnaum celine15@example.org
Alexie Goldner leonie.hansen@example.org
Raymundo Hirthe sigurd.marks@example.net
Alta Thiel heidenreich.deshaun@example.com
Alexys Ferry sam.grant@example.com
Andreane Lind johanna52@example.org
Hilario Sipes zbraun@example.org
Holden Reinger nolan.christ@example.com
Jarred Skiles millie.mayert@example.com
Miller Schumm uschuppe@example.net
Candida Hamill fvolkman@example.org
Edmond Johns beatty.helga@example.com
Prince Bartoletti oswald76@example.com
Maya Quitzon eldon.flatley@example.org
Alexandre Ratke turner.cornelius@example.com
[student@192 ~]$ █

```

15. Import the only matching row with `--where` statement. The imported rows are the first named 'Dorthy' in the authors table.

```
$ sqoop import --connect jdbc:mysql://localhost/test --username student --password student --table authors --fields-terminated-by '\t' --where "first_name='Dorthy'" --target-dir authors_Dorthy
```

16. Import a table using an alternate file format instead of text format. Import the authors table to Parquet format.

```
$sqoop import --connect jdbc:mysql://localhost/labs --username student --password student --table authors --target-dir /mywarehouse/authors_parquet --as-parquetfile
```



```
[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs --username student --password student -
-table authors --target-dir /mywarehouse/authors_parquet --as-parquetfile
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER HOME to the root of your Zookeeper installation.
2024-06-21 01:14:29,385 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 01:14:29,557 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Con
sider using -P instead.
2024-06-21 01:14:29,888 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 01:14:29,888 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 01:14:29,888 INFO tool.CodeGenTool: Will generate java class as codegen authors
2024-06-21 01:14:30,575 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LI
MIT 1
2024-06-21 01:14:30,646 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LI
MIT 1
2024-06-21 01:14:30,666 INFO orm.CompilationManager: HADOOP MAPRED HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/176a40fddd1eb38870892e574510436a/codegen_authors.java uses or overrides a
deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-06-21 01:14:33,818 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/176a40fdd
d1eb38870892e574510436a/codegen_authors.jar
2024-06-21 01:14:33,875 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-06-21 01:14:33,875 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-06-21 01:14:33,875 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-06-21 01:14:33,876 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-06-21 01:14:33,911 INFO mapreduce.ImportJobBase: Beginning import of authors
2024-06-21 01:14:33,913 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapre
duce.jobtracker.address
2024-06-21 01:14:34,286 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.jo
b.jar
2024-06-21 01:14:35,355 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LI
MIT 1
2024-06-21 01:14:35,359 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LI
MIT 1
```

17. view the results of the import commands by listing the contents in HDFS (target-dir).

```
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors_parquet
Found 6 items
drwxr-xr-x - student supergroup 0 2024-06-21 01:14 /mywarehouse/authors_parquet/.metadata
drwxr-xr-x - student supergroup 0 2024-06-21 01:15 /mywarehouse/authors_parquet/.signals
-rw-r--r-- 1 student supergroup 97582 2024-06-21 01:15 /mywarehouse/authors_parquet/0e35e37b-cc97-45
49-8a5b-7e9214c2ba59.parquet
-rw-r--r-- 1 student supergroup 97715 2024-06-21 01:15 /mywarehouse/authors_parquet/12ad3516-3dc5-40
79-8c16-e4e999cdeda2.parquet
-rw-r--r-- 1 student supergroup 97492 2024-06-21 01:15 /mywarehouse/authors_parquet/358933bc-5120-4d
f4-8aa7-6c821b28fe0e.parquet
-rw-r--r-- 1 student supergroup 97397 2024-06-21 01:15 /mywarehouse/authors_parquet/b14269d8-3ed2-43
35-a55e-d17172812418.parquet
```

```
$ hdfs dfs -get /mywarehouse/authors_parquet/ e35e37b-cc97-4549-8a5b-
7e9214c2ba59.parquet
```

```
$ parquet-tools show e35e37b-cc97-4549-8a5b-7e9214c2ba59.parquet
```

```

7468 | Adelle | Tremblay | verda88@example.net | 506444400000 | 1320537070000
7469 | Russell | Simonis | erich.romaguera@example.com | 1304953200000 | 667798179000
7470 | Jamie | Brown | clint.reilly@example.org | 109177200000 | 767473544000
7471 | Adam | Jacobs | cecelia75@example.net | 1317999600000 | 432076835000
7472 | Dannie | Schultz | stroman.candace@example.org | 613666800000 | 740218105000
7473 | Leo | Streich | hmaggio@example.com | 1533394800000 | 1504424400000
7474 | Wilfrid | Wintheiser | ward.cora@example.net | 631292400000 | 1191941891000
7475 | Jensen | Weimann | iliana.hahn@example.org | 940690800000 | 572446042000
7476 | Annamarie | Pagac | luella22@example.net | 1127746800000 | 1100084763000
7477 | Marcella | Durgan | jaunita28@example.net | 1313938800000 | 1007912184000
7478 | Nettie | Bradtke | georgiana.kertzmnn@example.net | 344444400000 | 1258366581000
7479 | Hayley | Hoeger | kmurphy@example.com | 762361200000 | 1091308428000
7480 | Javonte | Mante | carter.santino@example.org | 624207600000 | 1079870372000
7481 | Celestino | Towne | willard79@example.com | 1085583600000 | 591907341000
7482 | Whitney | Jakubowski | stacey.schamberger@example.net | 1029510000000 | 704321406000
7483 | Christian | Davis | angelina67@example.org | 1277823600000 | 463131466000
7484 | Guisepppe | Nader | qhintz@example.net | 711385200000 | 1426518215000
7485 | Geovanny | Feest | xfritsch@example.net | 507740400000 | 821335346000
7486 | Eriberto | Fay | iquitzon@example.org | 174841200000 | 1193951228000
7487 | Colten | Oberbrunner | reagan.wilderman@example.com | 122223600000 | 1215787211000
7488 | Laurie | Schiller | gerardo.klein@example.net | 655052400000 | 538263933000
7489 | Lance | Gislason | kgoodwin@example.org | 840380400000 | 281002480000
7490 | Federico | Reichert | nicolas.kiera@example.com | 1318950000000 | 728182854000
7491 | Selena | Moen | helga.herzog@example.org | 1502377200000 | 1019626669000
7492 | Leila | White | hans.shields@example.net | 682095600000 | 1085232250000
7493 | Jamaal | DuBuque | moore.esmeralda@example.org | 811868400000 | 916063705000
7494 | Alexie | Reilly | blanca.d'amore@example.org | 980866800000 | 864477261000
7495 | Casimir | Deckow | liam81@example.org | 240505200000 | 253415494000
7496 | Tara | Spencer | schmitt.alanis@example.org | 365180400000 | 1458672615000
7497 | Candida | Herzog | chadrick.ortiz@example.com | 828802800000 | 577087626000
7498 | Ari | Rau | lucie.brekke@example.com | 483116400000 | 276924294000
7499 | Crawford | Okuneva | bayer.opheia@example.net | 687193200000 | 1126471214000
7500 | Guido | Gibson | yankunding@example.org | 103561200000 | 1211873116000
-----+-----+-----+-----+-----+-----+
[student@192 ~]$
::1 localhost4 localhost6 localhost.localdomain
localhost localhost4.localdomain4 localhost6.localdomain6
[student@192 ~]$ ss

```

18. Import a table using a compression option `--compress` or `-z` for authors table.
`$ sqoop import --connect jdbc:mysql://localhost/labs --username student --password student --table authors --target-dir /mywarehouse/authors_compressed --compress`


```
[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs \
> --username student --password student \
> --table authors --target-dir /mywarehouse/authors_compressed \
> --compress
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 01:21:52,646 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 01:21:52,714 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 01:21:53,019 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 01:21:53,019 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 01:21:53,617 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:21:53,682 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:21:53,715 INFO orm.CompilationManager: HADOOP MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/38c350524db5f90a078185d3e8bf4e54/authors.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-06-21 01:21:56,620 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/38c350524db5f90a078185d3e8bf4e54/authors.jar
2024-06-21 01:21:56,644 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-06-21 01:21:56,644 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-06-21 01:21:56,645 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-06-21 01:21:56,645 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-06-21 01:21:56,658 INFO mapreduce.ImportJobBase: Beginning import of authors
2024-06-21 01:21:56,660 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker.address
2024-06-21 01:21:56,998 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```

```
[student@192 ~]$ hdfs dfs -ls /mywarehouse/authors_compressed
Found 5 items
-rw-r--r-- 1 student supergroup 0 2024-06-21 01:22 /mywarehouse/authors_compressed/_SUCCESS
-rw-r--r-- 1 student supergroup 72776 2024-06-21 01:22 /mywarehouse/authors_compressed/part-m-00000.gz
-rw-r--r-- 1 student supergroup 72745 2024-06-21 01:22 /mywarehouse/authors_compressed/part-m-00001.gz
-rw-r--r-- 1 student supergroup 72689 2024-06-21 01:22 /mywarehouse/authors_compressed/part-m-00002.gz
-rw-r--r-- 1 student supergroup 72760 2024-06-21 01:22 /mywarehouse/authors_compressed/part-m-00003.gz
[student@192 ~]$
```

19. First, import the rows whose first name is "Dorthy" performed in step 15, and save it as dorthy folder in the hdfs home directory

```
$ sqoop import --connect jdbc:mysql://localhost/labs --username student --password student --table authors --fields-terminated-by '\t' --where "first_name='Dorthy'" --target-dir dorthy
```

```
[student@192 ~]$ sqoop import --connect jdbc:mysql://localhost/labs \
> --username student --password student \
> --table authors --fields-terminated-by '\t' \
> --where "first_name='Dorthy'" --target-dir /mywarehouse/dorthy
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 01:29:29,513 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 01:29:29,629 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 01:29:29,943 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 01:29:29,943 INFO tool.CodeGenTool: Beginning code generation
2024-06-21 01:29:30,643 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:29:30,709 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `authors` AS t LIMIT 1
2024-06-21 01:29:30,730 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
```

Export the saved dorthy folder as a table to the labs DB of RDBMS.

`$sqoop export --connect jdbc:mysql://localhost/labs --username student --password student --table authors_export --fields-terminated-by '\t' --export-dir dorthy`

```
[student@192 ~]$ sqoop export --connect jdbc:mysql://localhost/labs \
> --username student --password student \
> --table authors_export --fields-terminated-by '\t' \
> --export-dir /mywarehouse/dorthy
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-06-21 01:31:17,841 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-06-21 01:31:18,004 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-06-21 01:31:18,239 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-06-21 01:31:18,246 INFO tool.CodeGenTool: Beginning code generation
```

20. Review the contents of the exported records in MariaDB.

```
MariaDB [labs]> select * from authors_export;
+-----+-----+-----+-----+-----+-----+
| id   | first_name | last_name | email                | birthdate | added                |
+-----+-----+-----+-----+-----+-----+
| 1298 | Dorthy    | Dietrich | ibrekke@example.com  | 1983-02-12 | 1999-08-07 10:35:08 |
| 2484 | Dorthy    | Hermann  | cebert@example.com   | 1999-06-14 | 2017-10-31 07:50:47 |
| 3377 | Dorthy    | West     | mayer.braden@example.com | 2001-02-08 | 2008-05-10 20:54:13 |
+-----+-----+-----+-----+-----+-----+
3 rows in set (0.01 sec)

MariaDB [labs]>
```

student : mysql

student : mysql - Konsole

01:32 AM

Lab 2: Data Ingestion with Apache Flume

1. Simple Data Transfer

This Agent allows the user to generate events and subsequently log them to the console. This configuration defines a single agent named agent1.

1.1. Create configuration file

```
mkdir flume
cd flume
vi transfer.conf
```

1.2. Agent1 configuration file

The agent1 has a source that listens for data on port 3333, a channel that buffers event data in memory, and a sink that logs event data to the console.

```
agent1.sources = netcatSrc
agent1.channels = memChannel
agent1.sinks = log
agent1.sources.netcatSrc.channels = memChannel
agent1.sinks.log.channel = memChannel
agent1.sources.netcatSrc.type = netcat
agent1.sources.netcatSrc.bind = 0.0.0.0
agent1.sources.netcatSrc.port = 3333
agent1.sinks.log.type = logger
agent1.channels.memChannel.type = memory
agent1.channels.memChannel.capacity = 100
```

```
File Edit View Bookmarks Settings Help
agent1.sources = netcatSrc
agent1.channels = memChannel
agent1.sinks = log

agent1.sources.netcatSrc.channels = memChannel
agent1.sinks.log.channel = memChannel
agent1.sources.netcatSrc.type = netcat
agent1.sources.netcatSrc.bind = 0.0.0.0
agent1.sources.netcatSrc.port = 3333

agent1.sinks.log.type = logger
agent1.channels.memChannel.type = memory
agent1.channels.memChannel.capacity = 100
~
~
~
~
~
~
~
```

1.3. Flume agent1 execution

flume-ng agent -name agent1 -conf-file transfer.conf

```

[student@192 flume]$ flume-ng agent -name agent1 -conf-file transfer.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/hadoop/hadoop/bin/hadoop) for
HDFS access
Info: Including HBASE libraries found via (/usr/local/hbase/hbase-2.3.5/bin/hbas
e) for HBASE access
Info: Including Hive libraries found via (/usr/local/hive/hive-3.1.2) for Hive a
ccess
+ exec /opt/jdk1.8.0_291/bin/java -Xmx20m -cp '/usr/local/flume/flume-1.9.0/lib/
*/home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/*:/
home/hadoop/hadoop/share/hadoop/common/*:/home/hadoop/hadoop/share/hadoop/hdfs:/
home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop/share/hadoop/hdfs
/*:/home/hadoop/hadoop/share/hadoop/mapreduce/*:/home/hadoop/hadoop/share/hadoop
/yarn:/home/hadoop/hadoop/share/hadoop/yarn/lib/*:/home/hadoop/hadoop/share/hado
op/yarn/*:/bin:/boot:/copyright:/dev:/etc:/home:/lib:/lib64:/media:/mnt:/opt:/pr
oc:/root:/run:/sbin:/srv:/sys:/tmp:/usr:/var:/lib/alsa:/lib/binfmt.d:/lib/cpp:/l
ib/crda:/lib/cups:/lib/debug:/lib/dkms:/lib/dracut:/lib/firewalld:/lib/firmware:/
lib/fontconfig:/lib/games:/lib/gcc:/lib/gems:/lib/grub:/lib/java:/lib/java-1.5.
0:/lib/java-1.6.0:/lib/java-1.7.0:/lib/java-1.8.0:/lib/java-ext:/lib/jvm:/lib/jv
m-common:/lib/jvm-exports:/lib/jvm-private:/lib/kbd:/lib/kde3:/lib/kde4:/lib/kd
ump:/lib/kernel:/lib/locale:/lib/modprobe.d:/lib/modules:/lib/modules-load.d:/li
b/mozilla:/lib/NetworkManager:/lib/node_modules:/lib/os-release:/lib/polkit-1:/l
ib/python2.7:/lib/rpm:/lib/sendmail:/lib/sendmail.postfix:/lib/sse2:/lib/sysctl.
d:/lib/systemd:/lib/tmpfiles.d:/lib/tuned:/lib/udev:/lib/yum-plugins:/usr/local/
hbase/hbase-2.3.5/conf:/opt/jdk1.8.0_291/lib/tools.jar:/usr/local/hbase/hbase-2.
3.5:/usr/local/hbase/hbase-2.3.5/lib/shaded-clients/hbase-shaded-client-byo-hado
op-2.3.5.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/audience-
annotations-0.5.0.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/
commons-logging-1.2.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdpart
y/htrace-core4-4.2.0-incubating.jar:/usr/local/hbase/hbase-2.3.5/lib/client-faci
ng-thirdparty/log4j-1.2.17.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-th
irdparty/slf4j-api-1.7.30.jar:/home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop
/share/hadoop/common/lib/*:/home/hadoop/hadoop/share/hadoop/common/*:/home/hadoo
p/hadoop/share/hadoop/hdfs:/home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/had
oop/hadoop/share/hadoop/hdfs/*:/home/hadoop/hadoop/share/hadoop/mapreduce/*:/hom
e/hadoop/hadoop/share/hadoop/yarn:/home/hadoop/hadoop/share/hadoop/yarn/lib/*:/h

```

```

starting
2024-06-21 18:51:05,108 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration
file:transfer.conf
2024-06-21 18:51:05,115 INFO conf.FlumeConfiguration: Processing:log
2024-06-21 18:51:05,121 INFO conf.FlumeConfiguration: Processing:netcatSrc
2024-06-21 18:51:05,121 INFO conf.FlumeConfiguration: Processing:netcatSrc
2024-06-21 18:51:05,121 INFO conf.FlumeConfiguration: Processing:memChannel
2024-06-21 18:51:05,121 INFO conf.FlumeConfiguration: Processing:memChannel
2024-06-21 18:51:05,127 INFO conf.FlumeConfiguration: Processing:netcatSrc
2024-06-21 18:51:05,127 INFO conf.FlumeConfiguration: Processing:log
2024-06-21 18:51:05,129 INFO conf.FlumeConfiguration: Added sinks: log Agent: agent1
2024-06-21 18:51:05,129 INFO conf.FlumeConfiguration: Processing:netcatSrc
2024-06-21 18:51:05,129 WARN conf.FlumeConfiguration: Agent configuration for 'agent1' has no config
filters.
2024-06-21 18:51:05,204 INFO conf.FlumeConfiguration: Post-validation flume configuration contains co
nfiguration for agents: [agent1]
2024-06-21 18:51:05,204 INFO node.AbstractConfigurationProvider: Creating channels
2024-06-21 18:51:05,216 INFO channel.DefaultChannelFactory: Creating instance of channel memChannel t
ype memory
2024-06-21 18:51:05,223 INFO node.AbstractConfigurationProvider: Created channel memChannel
2024-06-21 18:51:05,229 INFO source.DefaultSourceFactory: Creating instance of source netcatSrc, type
netcat
2024-06-21 18:51:05,241 INFO sink.DefaultSinkFactory: Creating instance of sink: log, type: logger
2024-06-21 18:51:05,246 INFO node.AbstractConfigurationProvider: Channel memChannel connected to [net
catSrc, log]
2024-06-21 18:51:05,257 INFO node.Application: Starting new configuration:{ sourceRunners:{netcatSrc=
EventDrivenSourceRunner: { source:org.apache.flume.source.NetcatSource{name:netcatSrc,state:IDLE} }}
sinkRunners:{log=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@50024546 counterGrou
p:{ name:null counters:{} } }} channels:{memChannel=org.apache.flume.channel.MemoryChannel{name: memC
hannel}} }
2024-06-21 18:51:05,272 INFO node.Application: Starting Channel memChannel
2024-06-21 18:51:05,591 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type:
CHANNEL, name: memChannel: Successfully registered new MBean.
2024-06-21 18:51:05,591 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: me
mChannel started
2024-06-21 18:51:05,591 INFO node.Application: Starting Sink log
2024-06-21 18:51:05,598 INFO node.Application: Starting Source netcatSrc
2024-06-21 18:51:05,602 INFO source.NetcatSource: Source starting
2024-06-21 18:51:05,648 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannel
Impl[/0:0:0:0:0:0:0:3333]

```

1.4. Open another terminal window and execute the telnet command.

telnet localhost 3333

Typing whatever you want...

Hadoop

```

[most-name [port]]
[student@192 ~]$ telnet localhost 3333
Trying ::1...
Connected to localhost.
Escape character is '^]'.

```



```
[host+name [port]]
[student@192 ~]$ telnet localhost 3333
Trying ::1...
Connected to localhost.
Escape character is '^]'.
hadoop
OK
speka
OK
█
```

1.5. Check that the message sent to telnet in step 4 is output from the terminal where the flume agent was executed in step 3

```
Impl[/0:0:0:0:0:0:0:3333]
2024-06-21 18:53:19,711 INFO sink.LoggerSink: Event: { headers:{} body: 68 61 64 6F 6F 70 0D
      hadoop. }
2024-06-21 18:53:23,713 INFO sink.LoggerSink: Event: { headers:{} body: 73 70 65 6B 61 0D
      speka. }
█
```

flume : java

1.6. telnet close with command after ctrl + quit.

```
^] (ctrl+)]
telnet> close
```

```
OK
^]
telnet> close█
```

2. Basic Data Transfer with spool directory

This agent 2 is to save the files coming into the spool directory to the local directory.

2.1. Create configuration file

```
vi transfer_spool.conf
```

2.2. Agent2 configuration file

```
agent2.sources = dirSrc
agent2.channels = memChannel
agent2.sinks = fileSink
agent2.sources.dirSrc.channels = memChannel
agent2.sinks.fileSink.channel = memChannel
agent2.sources.dirSrc.type = spoolDir
agent2.sources.dirSrc.spoolDir = /home/student/data/spool
agent2.sinks.fileSink.type = file_roll
agent2.sinks.fileSink.sink.directory = /home/student/data/output
agent2.sinks.fileSink.sink.rollInterval = 0
agent2.channels.memChannel.type = memory
agent2.channels.memChannel.capacity = 100
```

```
File Edit View Bookmarks Settings Help
agent2.sources = dirSrc
agent2.channels = memChannel
agent2.sinks = fileSink
agent2.sources.dirSrc.channels = memChannel
agent2.sinks.fileSink.channel = memChannel
agent2.sources.dirSrc.type = spoolDir
agent2.sources.dirSrc.spoolDir = /home/student/data/spool
agent2.sinks.fileSink.type = file_roll
agent2.sinks.fileSink.sink.directory = /home/student/data/output
agent2.sinks.fileSink.sink.rollInterval = 0
agent2.channels.memChannel.type = memory
agent2.channels.memChannel.capacity = 100
~
~
~
~
```

File spool ko có

2.3. Flume agent2 execution

flume-ng agent -name agent2 -conf-file transfer_spool.conf

2.4. Open another terminal window and copy two sql files to spool directory.

mkdir -p flume/incoming flume/output

cd /home/student/flume/incoming

cp ~/Data/*.txt .

vi hello.txt

This is test file for Flume.

```
[student@192 flume]$ mkdir incoming
[student@192 flume]$ mkdir ouput
[student@192 flume]$ ls
incoming ouput transfer.conf transfer_spool.conf
[student@192 flume]$
```

```
[student@192 flume]$ cp ~/Data/*.txt .
[student@192 flume]$
```

```
This is test file for Flume
~
~
~
```

```
[student@192 incoming]$ vi hello.txt
[student@192 incoming]$ ls -l
total 200
-rwxr-x---. 1 student student 170549 Jun 21 19:07 alice_in_wonderland.txt
-rw-rw-r--. 1 student student   29 Jun 21 19:07 hello.txt
-rw-r--r--. 1 student student  5812 Jun 21 19:07 kv1.txt
-rw-r--r--. 1 student student   32 Jun 21 19:07 people.txt
-rwxr-x---. 1 student student  4987 Jun 21 19:07 pig_data1.txt
-rwxr-x---. 1 student student  5240 Jun 21 19:07 pig_data2.txt
[student@192 incoming]$
```

2.5. You can check the message that pig_data1.txt, pig_data2.txt, alice_in_wonderland.txt, and hello.txt copied to the spool directory in step 4 are transmitted to the terminal where Agent 2 is running.

Không có file spool trong máy

2.6. The transferred files are stored as files in the OUTPUT directory

Không có file spool trong máy

3. Using Interceptor

3.1. Create configuration file

vi interceptor.conf

3.2. Agent3 configuration file

```
agent3.sources = netcatSrc
agent3.channels = memChannel
agent3.sinks = log
agent3.sources.netcatSrc.channels = memChannel
agent1.sinks.log.channel = memChannel
agent1.sources.netcatSrc.type = netcat
agent1.sources.netcatSrc.bind = 0.0.0.0
agent1.sources.netcatSrc.port = 3333
agent1.sinks.log.type = logger
agent1.channels.memChannel.type = memory
agent1.channels.memChannel.capacity = 100
agent03.sources.netcatSrc.interceptors = i1
agent03.sources.netcatSrc.interceptors.i1.type = host
agent03.sources.netcatSrc.interceptors.i1.hostHeader = hostname
```

```

File Edit View Bookmarks Settings Help
agent3.sources = netcatSrc
agent3.channels = memChannel
agent3.sinks = log
agent3.sources.netcatSrc.channels = memChannel
agent1.sinks.log.channel = memChannel
agent1.sources.netcatSrc.type = netcat
agent1.sources.netcatSrc.bind = 0.0.0.0
agent1.sources.netcatSrc.port = 3333
agent1.sinks.log.type = logger
agent1.channels.memChannel.type = memory
agent1.channels.memChannel.capacity = 100
agent03.sources.netcatSrc.interceptors = i1
agent03.sources.netcatSrc.interceptors.i1.type = host
agent03.sources.netcatSrc.interceptors.i1.hostHeader = hostname
~
~
~
~

```

```

[student@192 flume]$ flume-ng agent -name agent3 -conf-file interceptor.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/hadoop/hadoop/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/usr/local/hbase/hbase-2.3.5/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/usr/local/hive/hive-3.1.2) for Hive access
+ exec /opt/jdk1.8.0_291/bin/java -Xmx20m -cp '/usr/local/flume/flume-1.9.0/lib/*:/home/hadoop/hadoop/etc/had
oop:/home/hadoop/hadoop/share/hadoop/common/lib/*:/home/hadoop/hadoop/share/hadoop/common/*:/home/hadoop/hado
op/share/hadoop/hdfs:/home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop/share/hadoop/hdfs/*:/hom
e/hadoop/hadoop/share/hadoop/mapreduce/*:/home/hadoop/hadoop/share/hadoop/yarn:/home/hadoop/hadoop/share/hado
op/yarn/lib/*:/home/hadoop/hadoop/share/hadoop/yarn/*:/bin:/boot:/copyright:/dev:/etc:/home:/lib:/lib64:/medi
a:/mnt:/opt:/proc:/root:/run:/sbin:/srv:/sys:/tmp:/usr:/var:/lib/alsa:/lib/binfmt.d:/lib/cpp:/lib/crda:/lib/c
ups:/lib/debug:/lib/dkms:/lib/dracut:/lib/firewalld:/lib/firmware:/lib/fontconfig:/lib/games:/lib/gcc:/lib/ge
ms:/lib/grub:/lib/java:/lib/java-1.5.0:/lib/java-1.6.0:/lib/java-1.7.0:/lib/java-1.8.0:/lib/java-ext:/lib/jvm
:/lib/jvm-common:/lib/jvm-exports:/lib/jvm-private:/lib/kbd:/lib/kde3:/lib/kde4:/lib/kdump:/lib/kernel:/lib/
locale:/lib/modprobe.d:/lib/modules:/lib/modules-load.d:/lib/mozilla:/lib/NetworkManager:/lib/node_modules:/l
ib/os-release:/lib/polkit-1:/lib/python2.7:/lib/rpm:/lib/sendmail:/lib/sendmail.postfix:/lib/sse2:/lib/sysctl
.d:/lib/systemd:/lib/tmpfiles.d:/lib/tuned:/lib/udev:/lib/yum-plugins:/usr/local/hbase/hbase-2.3.5/conf:/opt/
jdk1.8.0_291/lib/tools.jar:/usr/local/hbase/hbase-2.3.5:/usr/local/hbase/hbase-2.3.5/lib/shaded-clients/hbase
-shaded-client-byo-hadoop-2.3.5.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/audience-annota
tions-0.5.0.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/commons-logging-1.2.jar:/usr/local/
hbase/hbase-2.3.5/lib/client-facing-thirdparty/hbase-protocol-2.0.jar:/usr/local/hbase/hbase-2.3.5

```



```

2024-06-22 14:05:53,733 INFO node.Application: Starting new configuration: { sourceRunners:{netcatSrc=EventDrivenSourceRunner: { source:org.apache.flume.source.NetcatSource{name:netcatSrc,state:IDLE} }} sinkRunners:{log=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@41d37f1a counterGroup:{ name:null counters:{} } }} channels:{memChannel=org.apache.flume.channel.MemoryChannel{name: memChannel}} }
2024-06-22 14:05:53,737 INFO node.Application: Starting Channel memChannel
2024-06-22 14:05:53,993 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: memChannel: Successfully registered new MBean.
2024-06-22 14:05:53,993 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: memChannel started
2024-06-22 14:05:53,994 INFO node.Application: Starting Sink log
2024-06-22 14:05:54,000 INFO node.Application: Starting Source netcatSrc
2024-06-22 14:05:54,006 INFO source.NetcatSource: Source starting
2024-06-22 14:05:54,021 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/0:0:0:0:0:0:0:0:3333]

```

flume : java

telnet localhost 3333

This is testing Flume with interceptor.

Hadoop

Spark

```

[student@192 ~]$ telnet localhost 3333
Trying ::1...
Connected to localhost.
Escape character is '^'.

```

3.5. The message sent to telnet in step 4 is output from the terminal where the flume agent was executed in step 3, and it is confirmed that the IP address where the agent is currently running is inserted into the event header and transmitted

```

2024-06-22 14:07:04,060 INFO sink.LoggerSink: Event: { headers:{hostname=192.168.1.12} body: 74 68 69 73 20 69 73 20 74 65 73 74 69 6E 67 20 this is testing }
2024-06-22 14:07:04,158 INFO sink.LoggerSink: Event: { headers:{hostname=192.168.1.12} body: 68 61 64 6F 6F 70 0D hadoop. }
2024-06-22 14:07:05,812 INFO sink.LoggerSink: Event: { headers:{hostname=192.168.1.12} body: 73 70 61 72 6B 0D spark. }

```

flume : java

3.6. Delete the temporary directory used for the flume operations.

\$cd ~/flume

\$rm -rf incoming output

```

student@192 ~]$ cd ~/flume
student@192 flume]$ rm -rf incoming output
student@192 flume]$ █

```

4. Create a new Flume dataflow from scratch

4.1. Create a new flume configuration file with the following:

Source	
Type	Netcat
Bind	localhost

18

Port	11111
Channel	
Type	Disk
Capacity	1000
transactionCapacity	100
Sink	
Type	logger

```

agent.sources = netcatSource
agent.sources.netcatSource.type = netcat
agent.sources.netcatSource.bind = localhost
agent.sources.netcatSource.port = 11111

# Define the channel
agent.channels = memoryChannel
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 1000
agent.channels.memoryChannel.transactionCapacity = 100

# Define the sink
agent.sinks = loggerSink
agent.sinks.loggerSink.type = logger

# Connect the source and sink to the channel
agent.sources.netcatSource.channels = memoryChannel
agent.sinks.loggerSink.channel = memoryChannel

```

4.2. Start the agent

```

[student@192 flume]$ flume-ng agent --conf ./conf --conf-file flume.conf --name agent -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/home/hadoop/hadoop/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/usr/local/hbase/hbase-2.3.5/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/usr/local/hive/hive-3.1.2) for Hive access
+ exec /opt/jdk1.8.0_291/bin/java -Xmx20m -Dflume.root.logger=INFO,console -cp './conf:/usr/local/flume/flume-1.9.0/lib/*:/home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/*:/home/hadoop/hadoop/share/hadoop/common/*:/home/hadoop/hadoop/share/hadoop/hdfs:/home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop/share/hadoop/hdfs/*:/home/hadoop/hadoop/share/hadoop/mapreduce/*:/home/hadoop/hadoop/share/hadoop/yarn:/home/hadoop/hadoop/share/hadoop/yarn/lib/*:/home/hadoop/hadoop/share/hadoop/yarn/*:/bin:/boot:/copyright:/dev:/etc:/home:/lib:/lib64:/media:/mnt:/opt:/proc:/root:/run:/sbin:/srv:/sys:/tmp:/usr:/var:/lib/alsa:/lib/binfmt.d:/lib/cpp:/lib/crda:/lib/cups:/lib/debug:/lib/dkms:/lib/dracut:/lib/firewalld:/lib/firmware:/lib/fontconfig:/lib/games:/lib/gcc:/lib/gems:/lib/grub:/lib/java:/lib/java-1.5.0:/lib/java-1.6.0:/lib/java-1.7.0:/lib/java-1.8.0:/lib/java-ext:/lib/jvm:/lib/jvm-common:/lib/jvm-exports:/lib/jvm-private:/lib/kbd:/lib/kde3:/lib/kde4:/lib/kdump:/lib/kernel:/lib/locale:/lib/modprobe.d:/lib/modules:/lib/modules-load.d:/lib/mozilla:/lib/NetworkManager:/lib/node_modules:/lib/os-release:/lib/polkit-1:/lib/python2.7:/lib/rpm:/lib/sendmail:/lib/sendmail.postfix:/lib/sse2:/lib/sysctl.d:/lib/systemd:/lib/tmpfiles.d:/lib/tuned:/lib/udev:/lib/yum-plugins:/usr/local/hbase/hbase-2.3.5/conf:/opt/jdk1.8.0_291/lib/tools.jar:/usr/local/hbase/hbase-2.3.5:/usr/local/hbase/hbase-2.3.5/lib/shaded-clients/hbase-shaded-client-byo-hadoop-2.3.5.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/audience-annotations-0.5.0.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/commons-logging-1.2.jar:/usr/local/hbase/hbase-2.3.5/lib/client-facing-thirdparty/htrace-core4-4.2.0-incubating.jar:/usr/local/hbase/hbase-2.3.5'

```



```

2024-06-22 14:18:47,186 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name:
memoryChannel started
2024-06-22 14:18:47,187 INFO node.Application: Starting Sink loggerSink
2024-06-22 14:18:47,191 INFO node.Application: Starting Source netcatSource
2024-06-22 14:18:47,200 INFO source.NetcatSource: Source starting
2024-06-22 14:18:47,248 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChann
elImpl[/127.0.0.1:11111]

```



4.3. From another terminal start telnet and connect to port 44444. Start typing and you should see the result from the other terminal.

```
$telnet localhost 11111
```

```
...
```

```
Hello world! <ENTER>
```

```
OK
```

```

[student@192 ~]$ telnet localhost 11111
Trying ::1...
telnet: connect to address ::1: Connection refused
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
helloworld!
OK
hadoop
OK
spark
OK

```

```

memoryChannel started
2024-06-22 14:18:47,187 INFO node.Application: Starting Sink loggerSink
2024-06-22 14:18:47,191 INFO node.Application: Starting Source netcatSource
2024-06-22 14:18:47,200 INFO source.NetcatSource: Source starting
2024-06-22 14:18:47,248 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChann
elImpl[/127.0.0.1:11111]
2024-06-22 14:19:41,257 INFO sink.LoggerSink: Event: { headers:{} body: 68 65 6C 6C 6F 77 6F 72 6C
64 21 0D helloworld!. }
2024-06-22 14:19:41,257 INFO sink.LoggerSink: Event: { headers:{} body: 68 61 64 6F 6F 70 0D
hadoop. }
2024-06-22 14:19:42,303 INFO sink.LoggerSink: Event: { headers:{} body: 73 70 61 72 6B 0D
spark. }

```



