

## Task-17

Name - Jainam Shah  
ref ID- DIRSS2131

Q1. Compression algorithm while importing the data to hdfs work with deflate, bzip, lz4, gz4.

### Deflate

```
bash-4.1$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --target-dir /25aug_deflate1 -m 1 --compress --compression-codec deflate
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/08/25 00:23:16 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/25 00:23:17 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/25 00:23:17 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/08/25 00:23:17 INFO tool.CodeGenTool: Beginning code generation
22/08/25 00:23:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:23:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:23:17 INFO orm.CompilationManager: HADOOP_MAPRED_HOME = /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hdfs/compile/05cc016ce3398c0d2f8a2dbd9b4b145d/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/25 00:23:20 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hdfs/compile/05cc016ce3398c0d2f8a2dbd9b4b145d/orders.jar
22/08/25 00:23:20 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/25 00:23:20 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/25 00:23:20 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/25 00:23:20 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/25 00:23:20 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/25 00:23:20 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/25 00:23:21 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/25 00:23:22 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:23 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
```

---

```

        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:24 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:24 INFO db.DBInputFormat: Using read committed transaction isolation
22/08/25 00:23:24 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:23:24 INFO mapreduce.JobSubmitter: number of splits:1
22/08/25 00:23:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1661401342371_0009
22/08/25 00:23:25 INFO impl.YarnClientImpl: Submitted application application_1661401342371_0009
22/08/25 00:23:25 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1661401342371_0009/
22/08/25 00:23:25 INFO mapreduce.Job: Running job: job_1661401342371_0009
22/08/25 00:23:35 INFO mapreduce.Job: Job job_1661401342371_0009 running in uber mode : false
22/08/25 00:23:35 INFO mapreduce.Job: map 0% reduce 0%
22/08/25 00:23:52 INFO mapreduce.Job: map 100% reduce 0%
22/08/25 00:23:52 INFO mapreduce.Job: Job job_1661401342371_0009 completed successfully
22/08/25 00:23:53 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=171913
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=471094
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=7344128
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=14344
    Total time spent by all map tasks (ms)=14344
    Total time spent by all maps in occupied slots (ms)=7344128
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=14344
    Total vcore-milliseconds taken by all map tasks=14344
    Total megabyte-milliseconds taken by all map tasks=7344128
  Map-Reduce Framework
    Map input records=68883
    Map output records=68883
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=268
    CPU time spent (ms)=3990
    Physical memory (bytes) snapshot=99635200
    Virtual memory (bytes) snapshot=702337024
    Total committed heap usage (bytes)=29949952
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=471094
22/08/25 00:23:53 INFO mapreduce.ImportJobBase: Transferred 460.0527 KB in 31.3822 seconds (14.6597 KB/sec)
22/08/25 00:23:53 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
-bash-4.1$ hdfs dfs -ls /25aug_deflate1
Found 2 items
-rw-r--r-- 1 hdfs supergroup      0 2022-08-25 00:23 /25aug_deflate1/_SUCCESS
-rw-r--r-- 1 hdfs supergroup  471094 2022-08-25 00:23 /25aug_deflate1/part-m-00000.deflate
-bash-4.1$ █
```

## Bzip

```
-bash-4.1$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --target-dir /25aug_bzip -m 1 --compress --compression-codec Bzip2
Warning: /usr/lib/sqoop/.accumulo does not exist! Accumulo imports will fail.
Please set ACCUMULO_HOME to the root of your Accumulo installation.
22/08/25 00:48:44 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/25 00:48:44 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/25 00:48:45 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/08/25 00:48:45 INFO tool.CodeGenTool: Beginning code generation
22/08/25 00:48:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:48:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:48:46 INFO orm.CompilationManager: HADOOP MAPRED HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hdfs/compile/3fe7135ffddeddfce53b3c292d2da2fd4/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/25 00:48:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hdfs/compile/3fe7135ffddeddfce53b3c292d2da2fd4/orders.jar
22/08/25 00:48:49 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/25 00:48:49 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/25 00:48:49 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/25 00:48:49 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/25 00:48:49 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/25 00:48:49 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/25 00:48:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/25 00:48:51 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/08/25 00:48:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:48:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)

22/08/25 00:49:08 INFO mapreduce.Job: Job job_1661401342371_0011 running in uber mode : false
22/08/25 00:49:08 INFO mapreduce.Job: map 0% reduce 0%
22/08/25 00:49:22 INFO mapreduce.Job: map 100% reduce 0%
22/08/25 00:49:22 INFO mapreduce.Job: Job job_1661401342371_0011 completed successfully
22/08/25 00:49:22 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=171901
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=387672
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=5499904
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=10742
    Total vcore-milliseconds taken by all map tasks=10742
    Total megabyte-milliseconds taken by all map tasks=5499904
  Map-Reduce Framework
    Map input records=68883
    Map output records=68883
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=135
    CPU time spent (ms)=3070
    Physical memory (bytes) snapshot=107065344
    Virtual memory (bytes) snapshot=702316544
    Total committed heap usage (bytes)=29949952
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=387672
22/08/25 00:49:22 INFO mapreduce.ImportJobBase: Transferred 378.5859 KB in 31.4152 seconds (12.0511 KB/sec)
22/08/25 00:49:22 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
-bash-4.1$ hdfs dfs -cat /25aug_bzip
^[cat: '/25aug_bzip': Is a directory
-bash-4.1$ hdfs dfs -ls /25aug_bzip
Found 2 items
-rw-r--r-- 1 hdfs supergroup          0 2022-08-25 00:49 /25aug_bzip/_SUCCESS
-rw-r--r-- 1 hdfs supergroup  387672 2022-08-25 00:49 /25aug_bzip/part-m-00000.b2
```

## Gzip

```
-bash-4.1$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --target-dir /25aug_gzip -m 1 --compress --compression-codec Gzip
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/08/25 01:10:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/25 01:10:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/25 01:10:46 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/08/25 01:10:46 INFO tool.CodeGenTool: Beginning code generation
22/08/25 01:10:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 01:10:47 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 01:10:47 INFO orm.CompilationManager: HADOOP_MAPPED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hdfs/compile/385c204704af5fcbb88d9fce078d464/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/25 01:10:49 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hdfs/compile/385c204704af5fcbb88d9fce078d464/orders.jar
22/08/25 01:10:49 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/25 01:10:49 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/25 01:10:49 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/25 01:10:49 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/25 01:10:49 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/25 01:10:50 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/25 01:10:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/25 01:10:51 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8083
22/08/25 01:10:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 01:10:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 01:10:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 01:10:52 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/25 01:11:06 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1661401342371_0013/
22/08/25 01:10:54 INFO mapreduce.Job: Running job: job_1661401342371_0013
22/08/25 01:11:06 INFO mapreduce.Job: Job job_1661401342371_0013 running in uber mode : false
22/08/25 01:11:06 INFO mapreduce.Job: map 0% reduce 0%
22/08/25 01:11:19 INFO mapreduce.Job: map 100% reduce 0%
22/08/25 01:11:20 INFO mapreduce.Job: Job job_1661401342371_0013 completed successfully
22/08/25 01:11:20 INFO mapreduce.Job: Counters: 30
File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=171899
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=471106
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=5779456
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=11288
    Total vcore-milliseconds taken by all map tasks=11288
    Total megabyte-milliseconds taken by all map tasks=5779456
Map-Reduce Framework
    Map input records=68883
    Map output records=68883
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=166
    CPU time spent (ms)=3590
    Physical memory (bytes) snapshot=112705536
    Virtual memory (bytes) snapshot=702337024
    Total committed heap usage (bytes)=29949952
File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=471106
22/08/25 01:11:20 INFO mapreduce.ImportJobBase: Transferred 460.0645 KB in 29.5234 seconds (15.583 KB/sec)
22/08/25 01:11:20 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
-bash-4.1$ hdfs dfs -ls /25aug_gzip
Found 2 items
-rw-r--r-- 1 hdfs supergroup          0 2022-08-25 01:11 /25aug_gzip/_SUCCESS
-rw-r--r-- 1 hdfs supergroup 471106 2022-08-25 01:11 /25aug_gzip/part-m-00000.gz
```

## lz4

```
-bash-4.1$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --target-dir /25aug_lz4 -m 1 --compress --compression-codec lz4
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/08/25 00:40:07 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/25 00:40:07 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/25 00:40:07 INFO tool.CodeGenTool: Beginning code generation
22/08/25 00:40:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:40:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/25 00:40:08 INFO orm.CompilationManager: HADOOP MAPRED HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hdfs/compile/5b9f240570146444324ab8b169f80f6e/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/25 00:40:10 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hdfs/compile/5b9f240570146444324ab8b169f80f6e/orders.jar
22/08/25 00:40:10 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/25 00:40:10 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/25 00:40:10 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/25 00:40:10 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/25 00:40:10 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/25 00:40:11 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/25 00:40:12 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/25 00:40:12 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/08/25 00:40:13 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:40:13 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:894)
22/08/25 00:40:13 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
22/08/25 00:40:42 INFO mapreduce.ImportJobBase: Transferred 846.4258 KB in 30.6381 seconds (27.6266 KB/sec)
22/08/25 00:40:42 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
```

```
-bash-4.1$ hdfs dfs -ls /25aug_lz4
Found 2 items
-rw-r--r-- 1 hdfs supergroup      0 2022-08-25 00:40 /25aug_lz4/_SUCCESS
-rw-r--r-- 1 hdfs supergroup 866740 2022-08-25 00:40 /25aug_lz4/part-m-00000.lz4
```

Q2. Try to import all the table from the retail\_db database to the hadoop hdfs location /allTable

- Remember in case the same folder is present so you should still run the same sqoop command & should get the output.

```
[cloudera@quickstart ~]$ sqoop import-all-tables --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --warehouse-dir /27aug12 -m 1

at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 03:19:37 INFO mapreduce.JobSubmitter: number of splits:1
22/08/26 03:19:37 INFO mapreduce.Job: Submitting tokens for job: job_1661493687528_0013
22/08/26 03:19:37 INFO impl.YarnClientImpl: Submitted application application_1661493687528_0013
22/08/26 03:19:37 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1661493687528_0013/
22/08/26 03:19:37 INFO mapreduce.Job: Running job: job_1661493687528_0013
22/08/26 03:19:48 INFO mapreduce.Job: Job: job_1661493687528_0013 running in uber mode : false
22/08/26 03:19:57 INFO mapreduce.Job: map 0% reduce 0%
22/08/26 03:19:57 INFO mapreduce.Job: Job: job_1661493687528_0013 completed successfully
22/08/26 03:19:57 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172104
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=173993
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=2756608
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=5384
    Total vcore-milliseconds taken by all map tasks=5384
    Total megabyte-milliseconds taken by all map tasks=2756608
  Map-Reduce Framework
    Map input records=1345
    Map output records=1345
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=104
    CPU time spent (ms)=790
    Physical memory (bytes) snapshot=110809088
    Virtual memory (bytes) snapshot=699101184
    Total committed heap usage (bytes)=29949952
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=173993
22/08/26 03:19:57 INFO mapreduce.ImportJobBase: Transferred 169.915 KB in 22.1723 seconds (7.6634 KB/sec)
22/08/26 03:19:57 INFO mapreduce.ImportJobBase: Retrieved 1345 records.
```

```
-bash-4.1$ hdfs dfs -ls /27aug12
Found 6 items
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:13 /27aug12/categories
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:13 /27aug12/customers
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:14 /27aug12/departments
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:19 /27aug12/order_items
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:19 /27aug12/orders
drwxr-xr-x  - cloudera supergroup      0 2022-08-26 03:19 /27aug12/products
```

Q3. Try to find out the difference between the option => --warehouse-dir & --target-dir

--warehouse-dir points to the Hive folder to import data into (I've used it when importing tables wholesale). While --target-dir is needed when importing into hive via query (sqoop errs asking for it). In the latter scenario, it is used as a temporary area for the mappers to be followed by LOAD INPATH.

- . Target-dir will work only when you import a single table. That implies this won't work when you use "Sqoop import-all-tables"
- . warehouse-dir creates the parent directory in which all your tables will be stored in the folders which are named after the table name.
- . If you are importing table by table, each time you need to provide the distinctive target-directory location as target-directory location can't be same in each import.

#### Q4. Try to import the data from the orders table of mysql

-only extract 3 columns such as => order\_id, order\_date, order\_customer\_id (in the hdfs table)

- Get all the above 3 columns for only for those order\_customer\_id whose customer\_id is greater than 5000

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --columns order_id,order_date,order_customer_id --where "order_customer_id > 5000" --target-dir /26aug_s
--m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/08/26 00:44:22 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/26 00:44:22 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/26 00:44:22 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/08/26 00:44:22 INFO tool.CodeGenTool: Beginning code generation
22/08/26 00:44:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/26 00:44:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/26 00:44:23 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/fbe22ea153109417bbd56c5bd9be2f74/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/26 00:44:26 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/fbe22ea153109417bbd56c5bd9be2f74/orders.jar
22/08/26 00:44:26 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/26 00:44:26 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/26 00:44:26 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/26 00:44:26 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/26 00:44:26 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/26 00:44:26 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/26 00:44:27 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/26 00:44:28 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/08/26 00:44:29 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 00:44:29 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 00:44:29 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 00:44:29 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
```

```

at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 08:44:30 INFO mapreduce.JobSubmitter: number of splits:1
22/08/26 08:44:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1661493687528_0005
22/08/26 08:44:31 INFO impl.YarnClientImpl: Submitted application application_1661493687528_0005
22/08/26 08:44:31 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1661493687528_0005/
22/08/26 08:44:48 INFO mapreduce.Job: Job job_1661493687528_0005 running in uber mode : false
22/08/26 08:44:48 INFO mapreduce.Job: map 0% reduce 0%
22/08/26 08:51:37 INFO mapreduce.Job: map 100% reduce 0%
22/08/26 08:51:38 INFO mapreduce.Job: Job job_1661493687528_0005 completed successfully
22/08/26 08:51:38 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172727
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=1366720
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=207502336
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=405278
    Total vcore-milliseconds taken by all map tasks=405278
    Total megabyte-milliseconds taken by all map tasks=207502336
  Map-Reduce Framework
    Map input records=41208
    Map output records=41208
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=981
    CPU time spent (ms)=2920
    Physical memory (bytes) snapshot=105828352
    Virtual memory (bytes) snapshot=699215872
    Total committed heap usage (bytes)=29949952
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=1366720
22/08/26 08:51:38 INFO mapreduce.ImportJobBase: Transferred 1.3084 MB in 430.7436 seconds (3.0986 KB/sec)
22/08/26 08:51:38 INFO mapreduce.ImportJobBase: Retrieved 41208 records.

```

```

-bash-4.1$ hdfs dfs -ls /26aug_s
Found 2 items
-rw-r--r-- 1 cloudera supergroup          0 2022-08-26 00:51 /26aug_s/_SUCCESS
-rw-r--r-- 1 cloudera supergroup 1366720 2022-08-26 00:51 /26aug_s/part-m-00000

```

```
-bash-4.1$ hdfs dfs -cat /26aug_s/part-m-00000
```

```

68794,2014-01-31 00:00:00.0,6873
68795,2014-02-09 00:00:00.0,6950
68796,2014-02-10 00:00:00.0,12377
68797,2014-02-12 00:00:00.0,5932
68799,2014-02-14 00:00:00.0,11190
68800,2014-02-17 00:00:00.0,10037
68802,2014-02-19 00:00:00.0,10670
68803,2014-02-23 00:00:00.0,9397
68805,2014-03-01 00:00:00.0,6599
68806,2014-03-02 00:00:00.0,10351
68809,2014-03-12 00:00:00.0,5946
68810,2014-03-15 00:00:00.0,9394
68814,2014-03-23 00:00:00.0,11408
68816,2014-03-26 00:00:00.0,8769
68817,2014-03-27 00:00:00.0,6704
68818,2014-03-31 00:00:00.0,12393
68820,2014-04-04 00:00:00.0,6358
68823,2014-04-07 00:00:00.0,7864
68825,2014-04-10 00:00:00.0,10201
68826,2014-04-14 00:00:00.0,7958
68827,2014-04-16 00:00:00.0,8814
68828,2014-04-20 00:00:00.0,5263
68829,2014-04-22 00:00:00.0,5981
68831,2014-04-24 00:00:00.0,8763
68832,2014-04-26 00:00:00.0,112093
68833,2014-04-30 00:00:00.0,12428
68834,2014-05-01 00:00:00.0,6938
68835,2014-05-03 00:00:00.0,8009
68839,2014-05-10 00:00:00.0,10090
68841,2014-05-14 00:00:00.0,12258
68842,2014-05-15 00:00:00.0,11268
68843,2014-05-16 00:00:00.0,10347
68844,2014-05-18 00:00:00.0,6584
68848,2014-05-22 00:00:00.0,6517
68850,2014-05-25 00:00:00.0,8451
68851,2014-05-26 00:00:00.0,11193
68854,2014-06-01 00:00:00.0,6528
68856,2014-06-03 00:00:00.0,8853
68858,2014-06-06 00:00:00.0,10744
68862,2014-06-15 00:00:00.0,7326
68864,2014-06-18 00:00:00.0,9634
68868,2014-06-24 00:00:00.0,10184
68869,2014-06-25 00:00:00.0,7456
68875,2014-07-04 00:00:00.0,10637
68877,2014-07-07 00:00:00.0,9692
68878,2014-07-08 00:00:00.0,6753
68882,2014-07-22 00:00:00.0,10000
68883,2014-07-23 00:00:00.0,5533

```

**Q5. Try to import any table at any location of hdfs but make sure the fields should be terminated by |(pipe operator)**

```
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -chmod 777 /
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/retail_db --username root --password cloudera --table orders --fields-terminated-by '|' --lines-terminated-by "\n" --target-dir /26aug -m 1
Warning: /usr/lib/sqoop/.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/08/26 00:33:31 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/08/26 00:33:31 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/08/26 00:33:31 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/08/26 00:33:31 INFO tool.CodeGenTool: Beginning code generation
22/08/26 00:33:32 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/26 00:33:32 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
22/08/26 00:33:32 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/c1848eefb49d51129ef8aa010fb273c7/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/08/26 00:33:35 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/c1848eefb49d51129ef8aa010fb273c7/orders.jar
22/08/26 00:33:35 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/08/26 00:33:35 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/08/26 00:33:35 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/08/26 00:33:35 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/08/26 00:33:35 INFO mapreduce.ImportJobBase: Beginning import of orders
22/08/26 00:33:35 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/08/26 00:33:36 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/08/26 00:33:36 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/08/26 00:33:37 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedIOException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 00:33:37 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedIOException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/08/26 00:33:38 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedIOException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:785)
```

```

22/08/26 00:33:39 INFO db.DBInputFormat: Using read committed transaction isolation
22/08/26 00:33:39 INFO mapreduce.JobSubmitter: number of splits:1
22/08/26 00:33:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1661493687528_0004
22/08/26 00:33:40 INFO impl.YarnClientImpl: Submitted application application_1661493687528_0004
22/08/26 00:33:40 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1661493687528_0004/
22/08/26 00:33:40 INFO mapreduce.Job: Running job: job_1661493687528_0004
22/08/26 00:33:53 INFO mapreduce.Job: Job job_1661493687528_0004 running in uber mode : false
22/08/26 00:33:53 INFO mapreduce.Job: map 0% reduce 0%
22/08/26 00:34:20 INFO mapreduce.Job: map 100% reduce 0%
22/08/26 00:34:21 INFO mapreduce.Job: Job job_1661493687528_0004 completed successfully
22/08/26 00:34:21 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172186
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=2999944
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=12196864
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=23822
    Total vcore-milliseconds taken by all map tasks=23822
    Total megabyte-milliseconds taken by all map tasks=12196864
  Map-Reduce Framework
    Map input records=68883
    Map output records=68883
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=162
    CPU time spent (ms)=3440
    Physical memory (bytes) snapshot=102715392
    Virtual memory (bytes) snapshot=700149760
    Total committed heap usage (bytes)=29949952
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=2999944
22/08/26 00:34:21 INFO mapreduce.ImportJobBase: Transferred 2.861 MB in 44.8901 seconds (65.2623 KB/sec)
22/08/26 00:34:21 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
[cloudera@quickstart ~]$ █
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfsadmin -safemode leave
Safe mode is OFF
[cloudera@quickstart ~]$ su - hdfs
Password:
-bash-4.1$ hdfs dfs -ld /
-ld: Unknown command
-bash-4.1$ hdfs dfs -ls /
Found 32 items
drwxr-xr-x  - hdfs  supergroup          0 2022-08-18 08:57 /12aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-16 07:36 /18aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-19 05:22 /19aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-12 23:37 /1aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-24 23:10 /25aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-24 23:17 /25aug.snapppy
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 00:49 /25aug_bzip
drwxr-xr-x  - hdfs  supergroup          0 2022-08-24 23:41 /25aug_cond
drwxr-xr-x  - hdfs  supergroup          0 2022-08-24 23:37 /25aug_cust
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 00:18 /25aug_deflate
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 00:23 /25aug_deflate1
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 01:11 /25aug_gzip
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 00:40 /25aug_lz4
drwxr-xr-x  - hdfs  supergroup          0 2022-08-25 00:58 /25aug_lzip
drwxr-xr-x  - hdfs  supergroup          0 2022-08-24 23:21 /25aug_snapppy
drwxr-xr-x  - cloudera supergroup        0 2022-08-26 00:34 /26aug
drwxr-xr-x  - hdfs  supergroup          0 2022-08-16 05:15 /5t
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
-rw-r--r--  1 hdfs  supergroup          0 2022-08-13 01:29 /datascience
drwxr-xr-x  - hdfs  supergroup          0 2022-08-18 09:53 /dk
-rw-r--r--  1 hdfs  supergroup          0 2022-08-16 05:14 /fi.txt
drwxr-xr-x  - hbase supergroup          0 2022-08-10 02:53 /hbase
drwxr-xr-x  - hdfs  supergroup          0 2022-08-14 09:34 /hil
drwxr-xr-x  - hdfs  supergroup          0 2022-08-18 09:56 /home
drwxr-xr-x  - hdfs  supergroup          0 2022-08-14 10:20 /l2
drwxr-xr-x  - hdfs  supergroup          0 2022-08-16 05:06 /le3
-rw-r--r--  3 hdfs  supergroup          0 2022-08-14 10:11 /learn2
drwxr-xr-x  - hdfs  supergroup          0 2022-08-14 10:53 /lern3
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-08-16 09:10 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-08-12 23:33 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
-bash-4.1$ hdfs dfs -ls /26aug
Found 2 items
-rw-r--r--  1 cloudera supergroup      0 2022-08-26 00:34 /26aug/_SUCCESS
-rw-r--r--  1 cloudera supergroup 2999944 2022-08-26 00:34 /26aug/part-m-000
00
-bash-4.1$ hdfs dfs -ls /26aug

```

```
-bash-4.1$ hdfs dfs -cat /26aug/part-m-00000
```

68836	2014-05-03	00:00:00.0	8009	PENDING_PAYMENT
68837	2014-05-07	00:00:00.0	1223	COMPLETE
68838	2014-05-08	00:00:00.0	371	PENDING
68839	2014-05-10	00:00:00.0	10090	COMPLETE
68840	2014-05-11	00:00:00.0	14399	CLOSED
68841	2014-05-14	00:00:00.0	12258	PROCESSING
68842	2014-05-15	00:00:00.0	11268	PENDING
68843	2014-05-16	00:00:00.0	10347	PROCESSING
68844	2014-05-17	00:00:00.0	443	COMPLETE
68845	2014-05-18	00:00:00.0	16584	COMPLETE
68846	2014-05-20	00:00:00.0	13454	PROCESSING
68847	2014-05-21	00:00:00.0	2543	COMPLETE
68848	2014-05-22	00:00:00.0	16517	CLOSED
68849	2014-05-23	00:00:00.0	2356	COMPLETE
68850	2014-05-25	00:00:00.0	18451	COMPLETE
68851	2014-05-26	00:00:00.0	11193	PENDING_PAYMENT
68852	2014-05-29	00:00:00.0	4596	CLOSED
68853	2014-05-31	00:00:00.0	1202	PENDING_PAYMENT
68854	2014-06-01	00:00:00.0	16528	ON_HOLD
68855	2014-06-02	00:00:00.0	403	PENDING_PAYMENT
68856	2014-06-03	00:00:00.0	18853	PROCESSING
68857	2014-06-04	00:00:00.0	312	COMPLETE
68858	2014-06-06	00:00:00.0	10744	COMPLETE
68859	2014-06-11	00:00:00.0	1428	COMPLETE
68860	2014-06-12	00:00:00.0	4229	PENDING
68861	2014-06-13	00:00:00.0	3031	PENDING_PAYMENT
68862	2014-06-15	00:00:00.0	17326	PROCESSING
68863	2014-06-16	00:00:00.0	3361	CLOSED
68864	2014-06-18	00:00:00.0	19634	ON_HOLD
68865	2014-06-19	00:00:00.0	14567	SUSPECTED_FRAUD
68866	2014-06-20	00:00:00.0	3890	PENDING_PAYMENT
68867	2014-06-23	00:00:00.0	1869	CANCELED
68868	2014-06-24	00:00:00.0	10184	PENDING
68869	2014-06-25	00:00:00.0	7456	PROCESSING
68870	2014-06-26	00:00:00.0	3343	COMPLETE
68871	2014-06-28	00:00:00.0	4960	PENDING
68872	2014-06-29	00:00:00.0	3354	COMPLETE
68873	2014-06-30	00:00:00.0	14545	PENDING
68874	2014-07-03	00:00:00.0	1601	COMPLETE
68875	2014-07-04	00:00:00.0	10637	ON_HOLD
68876	2014-07-06	00:00:00.0	4124	COMPLETE
68877	2014-07-07	00:00:00.0	9692	ON_HOLD
68878	2014-07-08	00:00:00.0	16753	COMPLETE
68879	2014-07-09	00:00:00.0	1778	COMPLETE
68880	2014-07-13	00:00:00.0	1117	COMPLETE
68881	2014-07-19	00:00:00.0	2518	PENDING_PAYMENT
68882	2014-07-22	00:00:00.0	10000	ON_HOLD
68883	2014-07-23	00:00:00.0	5533	COMPLETE

-