

Step 4: Use Sqoop to load data into S3 Bucket

Run the following Sqoop command to list tables in your RDS database (use your endpoint, username and password):

```
sqoop list-tables --connect
```

```
jdbc:mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com/happiness --username  
admin --password REDACTED
```

```
[ec2-user@ip-172-31-43-216 /]$ sqoop list-tables --connect jdbc:mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com/happiness --username admin --password REDACTED  
[ec2-user@ip-172-31-43-216 ~]$ sqoop list-tables --connect jdbc:mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com/happiness --username admin --password REDACTED  
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]  
2023-11-05 19:47:04,676 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2023-11-05 19:47:04,795 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2023-11-05 19:47:05,181 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.  
2015  
2016  
2017  
2018  
2019  
[ec2-user@ip-172-31-43-216 ~]$
```

```
2015  
2016  
2017  
2018  
2019  
[ec2-user@ip-172-31-43-216 /]$
```

RDS endpoint: [mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com](https://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com)

EMR Master Public DNS: ec2-54-86-214-57.compute-1.amazonaws.com

Hue Application Interface Link: <http://ec2-54-86-214-57.compute-1.amazonaws.com:8888/>

Note: Via troubleshooting, I accidentally terminated my EMR instances a few times and thus this is the public DNS of the most recent EMR instance I spun up. The same goes for the Hue link

Run the following Sqoop command to move data from your RDS database to AWS S3 bucket you created:

```
sqoop import --connect
```

```
jdbc:mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com/happiness --username  
admin --password REDACTED --table 2018 --target-dir s3://fm-671happy/Lab5/
```

```
2023-11-05 20:01:27,251 INFO mapreduce.Job: Job job_1699211210135_0001 completed successfully
2023-11-05 20:01:27,395 INFO mapreduce.Job: Counters: 39
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=1195474
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=467
    HDFS: Number of bytes written=0
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=0
    HDFS: Number of bytes read erasure-coded=0
    S3: Number of bytes read=0
    S3: Number of bytes written=8369
    S3: Number of read operations=0
    S3: Number of large read operations=0
    S3: Number of write operations=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)=3294096
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=68627
    Total vcore-milliseconds taken by all map tasks=68627
    Total megabyte-milliseconds taken by all map tasks=105411072
  Map-Reduce Framework
    Map input records=156
    Map output records=156
    Input split bytes=467
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=2189
    CPU time spent (ms)=14730
    Physical memory (bytes) snapshot=1805864960
    Virtual memory (bytes) snapshot=12369584128
    Total committed heap usage (bytes)=1508900864
    Peak Map Physical memory (bytes)=473292800
    Peak Map Virtual memory (bytes)=3100315648
  File Input Format Counters
```

```
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=8369
2023-11-05 20:01:27,407 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 67.4688 seconds (0 bytes/sec)
2023-11-05 20:01:27,411 INFO mapreduce.ImportJobBase: Retrieved 156 records.
```

Lab5/ Copy S3 URI

Objects Properties

Objects (5)
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	._SUCCESS	-	November 5, 2023, 15:01:26 (UTC-05:00)	0 B	Standard
<input type="checkbox"/>	part-m-00000	-	November 5, 2023, 15:01:07 (UTC-05:00)	2.0 KB	Standard
<input type="checkbox"/>	part-m-00001	-	November 5, 2023, 15:01:07 (UTC-05:00)	2.0 KB	Standard
<input type="checkbox"/>	part-m-00002	-	November 5, 2023, 15:01:24 (UTC-05:00)	2.1 KB	Standard
<input type="checkbox"/>	part-m-00003	-	November 5, 2023, 15:01:25 (UTC-05:00)	2.0 KB	Standard

S3 Bucket: s3://fm-671happy/Lab5/

Run the following Sqoop command to move data from your RDS database to HDFS:

sqoop import --connect

jdbc:mysql://happiness.cqjkb4mswjw.us-east-1.rds.amazonaws.com/happiness --username








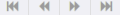
admin --password REDACTED --table 2018 --target-dir /user/hadoop/lab5s3-dist-cp --src

s3://fm-671happy/Lab5/ --dest=hdfs:///user/hadoop/lab5

```
2023-11-05 23:03:23,902 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2023-11-05 23:03:24,161 INFO client.DefaultHadoopFileProxyProvider: Connecting to ResourceManager at ip-172-31-38-220.ec2.internal/172.31.38.220:8032
2023-11-05 23:03:24,593 INFO client.AMSProxy: Connecting to Application History server at ip-172-31-38-220.ec2.internal/172.31.38.220:10200
2023-11-05 23:03:25,049 INFO mapreduce.JobResourceMonitor: Enabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1699224220342_0001
2023-11-05 23:03:28,034 INFO db.DBInputFormat: Using read committed transaction isolation
2023-11-05 23:03:28,035 INFO db.DataDrivenDBInputFormat: BoundingValueQuery: SELECT MIN('OverallRank'), MAX('OverallRank') FROM '2018'
2023-11-05 23:03:28,040 INFO db.IntegerSplitter: Split size: 38; Num splitters: 4 from: 1 to: 156
2023-11-05 23:03:28,096 INFO mapreduce.JobSubmitter: number of splits:4
2023-11-05 23:03:28,525 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1699224220342_0001
2023-11-05 23:03:28,525 INFO mapreduce.JobSubmitter: Executing with tokens: {}
2023-11-05 23:03:29,111 INFO conf.Configuration: resource-types.xml not found
2023-11-05 23:03:29,112 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-11-05 23:03:29,555 INFO emr1.YarnClientImpl: Submitted application application_1699224220342_0001
2023-11-05 23:03:29,661 INFO mapreduce.Job: The url to track the job: http://ip-172-31-38-220.ec2.internal:20888/proxy/application_1699224220342_0001/
2023-11-05 23:03:29,663 INFO mapreduce.Job: Running job: job_1699224220342_0001
2023-11-05 23:03:41,856 INFO mapreduce.Job: Job job_1699224220342_0001 running in uber mode : false
2023-11-05 23:03:41,857 INFO mapreduce.Job: map 0% reduce 0%
2023-11-05 23:03:56,054 INFO mapreduce.Job: map 25% reduce 0%
2023-11-05 23:03:57,063 INFO mapreduce.Job: map 100% reduce 0%
2023-11-05 23:03:57,070 INFO mapreduce.Job: Job job_1699224220342_0001 completed successfully
2023-11-05 23:03:57,170 INFO mapreduce.Job: Counters: 34
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=118432
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=467
    HDFS: Number of bytes written=3369
    HDFS: Number of read operations=24
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=0
    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)=238640
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=48670
    Total voice-milliseconds taken by all map tasks=48670
    Total mapreduce-milliseconds taken by all map tasks=4757120
  Map-Reduce Framework
    Map input records=156
    Map output records=156
    Input split bytes=467
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=7734
    CPU time spent (ms)=7910
    Physical memory (bytes) snapshot=118641492
    Virtual memory (bytes) snapshot=121643920
    Total committed heap usage (bytes)=1031798784
    Peak Map Physical memory (bytes)=303697520
    Peak Map Virtual memory (bytes)=308564196
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=3369
2023-11-05 23:03:57,172 INFO mapreduce.ImportJobBase: Transferred 8.1729 MB in 33.2421 seconds (251.7588 bytes/sec)
2023-11-05 23:03:57,177 INFO mapreduce.ImportJobBase: Retrieved 156 records.
[hadoop@ip-172-31-38-220 ~]$
```

```
2023-11-05 23:03:29,661 INFO mapreduce.Job: The url to track the job: http://ip-172-31-38-220.ec2.internal:20888/proxy/application_1699224220342_0001/
2023-11-05 23:03:29,663 INFO mapreduce.Job: Running job: job_1699224220342_0001
2023-11-05 23:03:41,856 INFO mapreduce.Job: Job job_1699224220342_0001 running in uber mode : false
2023-11-05 23:03:41,857 INFO mapreduce.Job: map 0% reduce 0%
2023-11-05 23:03:56,054 INFO mapreduce.Job: map 25% reduce 0%
2023-11-05 23:03:57,063 INFO mapreduce.Job: map 100% reduce 0%
2023-11-05 23:03:57,070 INFO mapreduce.Job: Job job_1699224220342_0001 completed successfully
2023-11-05 23:03:57,170 INFO mapreduce.Job: Counters: 34
```

```
2023-11-05 23:03:57,070 INFO mapreduce.Job: Job job_1699224220342_0001 completed successfully
```

Home /user/hadoop/lab5						
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 f		hadoop	hdfsadmingroup	drwxrwxrwx	November 05, 2023 03:24 PM
<input type="checkbox"/>	 .		hadoop	hdfsadmingroup	drwxr-xr-x	November 05, 2023 03:24 PM
<input type="checkbox"/>	 _SUCCESS	0 bytes	hadoop	hdfsadmingroup	-rw-r--r--	November 05, 2023 03:24 PM
<input type="checkbox"/>	 part-m-00000	2.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 05, 2023 03:24 PM
<input type="checkbox"/>	 part-m-00001	2.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 05, 2023 03:24 PM
<input type="checkbox"/>	 part-m-00002	2.1 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 05, 2023 03:24 PM
<input type="checkbox"/>	 part-m-00003	2.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 05, 2023 03:24 PM
Show 45 of 5 items					Page 1 of 1	

Use s3-dist-cp to copy files from your S3 bucket to hdfs:

s3-dist-cp --src s3://fm-671happy/Lab5/ --dest=hdfs:///user/hadoop/lab5

```
2023-11-05 23:07:50,384 INFO mapreduce.Job: Running job: job_1699224220342_0002
2023-11-05 23:07:59,646 INFO mapreduce.Job: Job job_1699224220342_0002 running in uber mode : false
2023-11-05 23:07:59,647 INFO mapreduce.Job: map 0% reduce 0%
2023-11-05 23:08:07,838 INFO mapreduce.Job: map 100% reduce 0%
2023-11-05 23:08:19,902 INFO mapreduce.Job: map 100% reduce 33%
2023-11-05 23:08:27,932 INFO mapreduce.Job: map 100% reduce 100%
2023-11-05 23:08:27,940 INFO mapreduce.Job: Job job_1699224220342_0002 completed successfully
2023-11-05 23:08:28,031 INFO mapreduce.Job: Counters: 59
```

Home /user/hadoop/lab5step4

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		hadoop	hdfsadmin	drwxrwxrwx	November 05, 2023 03:21 PM
<input type="checkbox"/>	.SUCCESS	0 bytes	hadoop	hdfsadmin	drwxr-xr-x	November 05, 2023 03:21 PM
<input type="checkbox"/>	part-m-00000	2.0 KB	hadoop	hdfsadmin	-rw-r--r--	November 05, 2023 03:21 PM
<input type="checkbox"/>	part-m-00001	2.0 KB	hadoop	hdfsadmin	-rw-r--r--	November 05, 2023 03:21 PM
<input type="checkbox"/>	part-m-00002	2.1 KB	hadoop	hdfsadmin	-rw-r--r--	November 05, 2023 03:21 PM
<input type="checkbox"/>	part-m-00003	2.0 KB	hadoop	hdfsadmin	-rw-r--r--	November 05, 2023 03:21 PM