

Create an EMR instance

The screenshot displays the Amazon EMR console interface. The left sidebar contains navigation links for Amazon EMR, EMR Studio, EMR on EC2, Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events, EMR on EKS, and Virtual clusters. The main content area shows the details for a cluster named 'upgrad_etlproject', which is in the 'Waiting' state. The cluster is ready after the last step completed.

Cluster: upgrad_etlproject Waiting Cluster ready after last step completed

Summary

- ID: j-CJK18T2GH0E0
- Creation date: 2022-01-29 11:36 (UTC+5:30)
- Elapsed time: 34 minutes
- After last step completes: Cluster waits
- Termination protection: Off [Change](#)
- Tags: -- [View All / Edit](#)
- Master public DNS: ec2-3-239-101-255.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

- Release label: emr-5.30.1
- Hadoop distribution: Amazon 2.8.5
- Applications: Hive 2.3.6, Hue 4.6.0, Spark 2.4.5, HBase 1.4.13, Sqoop 1.4.7, HCatalog 2.3.6, Livy 0.7.0
- Log URI: s3://aws-logs-693162548630-us-east-1/elasticmapreduce/
- EMRFS consistent view: Disabled
- Custom AMI ID: --

Application user interfaces

- Persistent user interfaces: [Spark history server](#), [YARN timeline server](#), [Tez UI](#)
- On-cluster user interfaces: Not Enabled [Enable an SSH Connection](#)

Network and hardware

- Availability zone: us-east-1f
- Subnet ID: [subnet-3a52b63b](#)
- Master: Running 1 m4.xlarge
- Core: --
- Task: --
- Cluster scaling: Not enabled
- Auto-termination: Not enabled

Security and access

- Key name: d112721
- EC2 instance profile: EMR_EC2_DefaultRole
- EMR role: EMR_DefaultRole
- Auto Scaling role: EMR_AutoScaling_DefaultRole
- Visible to all users: All [Change](#)
- Security groups for Master: [sg-00582519a48381a26](#) (ElasticMapReduce-master)
- Security groups for Core & Task: [sg-0e432a0c546a8a87c](#) (ElasticMapReduce-slave)

© 2022, Amazon Internet Services Private Ltd. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

Setting up mysql connector and mysql

```
hadoop@ip-172-31-69-119:~/mysql-connector-java-8.0.25
before moving into a production environment.

Remove test database and access to it? [Y/n] Y
- Dropping test database...
... Success!
- Removing privileges on test database...
... Success!

Reloading the privilege tables will ensure that all changes made so far
will take effect immediately.

Reload privilege tables now? [Y/n] Y
... Success!

Cleaning up...

All done! If you've completed all of the above steps, your MariaDB
installation should now be secure.

Thanks for using MariaDB!
[hadoop@ip-172-31-69-119 mysql-connector-java-8.0.25]$ mysql -u root -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 72
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 [(none)]>
```

Import from AWS into HDFS using Sqoop

SQOOP Import Command

sqoop import \

--connect jdbc:mysql://upgradetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \

--table SRC_ATM_TRANS \

--username student --password STUDENT123 \

-m 1

```
hadoop@ip-172-31-49-119:~$ sqoop import \
--connect jdbc:mysql://upgradetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
-m 1
23/01/29 07:08:28 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM SRC_ATM_TRANS AS t LIMIT 1
23/01/29 07:08:28 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM SRC_ATM_TRANS AS t LIMIT 1
23/01/29 07:08:28 INFO manager.SqoopManager: HADOOP_PATH_HOME is /usr/lib/hadoop-mapreduce
Note: /usr/lib/hadoop-mapreduce/share/hadoop-mapreduce-hdfs-client.jar does not override a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/01/29 07:08:32 INFO manager.SqoopManager: Writing jar file: /tmp/sqoop-hadoop/compile/9a74391e39326802aef3a68b8b0b/SRC_ATM_TRANS.jar
23/01/29 07:08:32 INFO manager.SqoopManager: It looks like you are importing from mysql.
23/01/29 07:08:32 INFO manager.SqoopManager: This transfer can be faster! Use the --direct option.
23/01/29 07:08:32 INFO manager.SqoopManager: Option to execute a Hadoop-specific fast path.
23/01/29 07:08:32 INFO manager.SqoopManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/01/29 07:08:32 INFO manager.ImportDatabase: Retrieving report of SRC_ATM_TRANS
23/01/29 07:08:33 INFO Configuration.deprecation: mapred.job.jar is deprecated. Instead, use mapreduce.job.jar
23/01/29 07:08:33 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/01/29 07:08:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-49-119.ec2.internal/172.31.49.119:8032
23/01/29 07:08:34 INFO rm.RMProxyClient: Using read committed transaction isolation
23/01/29 07:08:37 INFO mapreduce.JobSubmitter: number of splits=1
23/01/29 07:08:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_164343702481_0001
23/01/29 07:08:38 INFO mapreduce.Job: Submitting application application_164343702481_0001
23/01/29 07:08:38 INFO mapreduce.Job: The url to track the job: http://ip-172-31-49-119.ec2.internal:10080/progress/application_164343702481_0001/
23/01/29 07:08:38 INFO mapreduce.Job: Running job: job_164343702481_0001
23/01/29 07:08:48 INFO mapreduce.Job: Job job_164343702481_0001 running in uber mode : false
23/01/29 07:08:48 INFO mapreduce.Job: map 0 reduce 0
23/01/29 07:09:18 INFO mapreduce.Job: map 100% reduce 0%
23/01/29 07:09:18 INFO mapreduce.Job: Job job_164343702481_0001 completed successfully
23/01/29 07:09:18 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=18934
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=7
  HDFS: Number of bytes written=114115
  HDFS: Number of read operations=4
  HDFS: Number of write operations=1
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=1
Job Counters
  Launched map tasks=1
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=120400
  Total time spent by all reducers in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=120400
  Total write=114115 bytes taken by all map tasks=140058
  Total mapreduce-write=114115 bytes taken by all map tasks=140058
MapReduce Framework
  Map input records=140058
  Map output records=140058
  Input split bytes=0
  Spilled Records=0
  Failed Shuffles=0
  MapReduce program=0
  GC time elapsed (ms)=954
  CPU time spent (ms)=2770
  Physical memory (bytes) mapped=413414960
  Virtual memory (bytes) mapped=32772464
  Total committed heap usage (bytes)=33402312
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=114115
23/01/29 07:09:18 INFO mapreduce.ImportDatabase: Transferred 954.6059 MB in 44.2824 seconds (11.4403 MB/sec)
23/01/29 07:09:18 INFO mapreduce.ImportDatabase: Retrieved 246052 records.
hadoop@ip-172-31-49-119:~$
```

Using validation document to verify the count of records matching

Verifying the import using -ls command

```
hadoop@ip-172-31-49-119:~$ sqoop import \
--connect jdbc:mysql://upgradetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
-m 1
23/01/29 07:08:28 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM SRC_ATM_TRANS AS t LIMIT 1
23/01/29 07:08:28 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM SRC_ATM_TRANS AS t LIMIT 1
23/01/29 07:08:28 INFO manager.SqoopManager: HADOOP_PATH_HOME is /usr/lib/hadoop-mapreduce
Note: /usr/lib/hadoop-mapreduce/share/hadoop-mapreduce-hdfs-client.jar does not override a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/01/29 07:08:32 INFO manager.SqoopManager: Writing jar file: /tmp/sqoop-hadoop/compile/9a74391e39326802aef3a68b8b0b/SRC_ATM_TRANS.jar
23/01/29 07:08:32 INFO manager.SqoopManager: It looks like you are importing from mysql.
23/01/29 07:08:32 INFO manager.SqoopManager: This transfer can be faster! Use the --direct option.
23/01/29 07:08:32 INFO manager.SqoopManager: Option to execute a Hadoop-specific fast path.
23/01/29 07:08:32 INFO manager.SqoopManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/01/29 07:08:32 INFO manager.ImportDatabase: Retrieving report of SRC_ATM_TRANS
23/01/29 07:08:33 INFO Configuration.deprecation: mapred.job.jar is deprecated. Instead, use mapreduce.job.jar
23/01/29 07:08:33 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/01/29 07:08:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-49-119.ec2.internal/172.31.49.119:8032
23/01/29 07:08:34 INFO rm.RMProxyClient: Using read committed transaction isolation
23/01/29 07:08:37 INFO mapreduce.JobSubmitter: number of splits=1
23/01/29 07:08:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_164343702481_0001
23/01/29 07:08:38 INFO mapreduce.Job: Submitting application application_164343702481_0001
23/01/29 07:08:38 INFO mapreduce.Job: The url to track the job: http://ip-172-31-49-119.ec2.internal:10080/progress/application_164343702481_0001/
23/01/29 07:08:38 INFO mapreduce.Job: Running job: job_164343702481_0001
23/01/29 07:08:48 INFO mapreduce.Job: Job job_164343702481_0001 running in uber mode : false
23/01/29 07:08:48 INFO mapreduce.Job: map 0 reduce 0
23/01/29 07:09:18 INFO mapreduce.Job: map 100% reduce 0%
23/01/29 07:09:18 INFO mapreduce.Job: Job job_164343702481_0001 completed successfully
23/01/29 07:09:18 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=18934
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=7
  HDFS: Number of bytes written=114115
  HDFS: Number of read operations=4
  HDFS: Number of write operations=1
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=1
Job Counters
  Launched map tasks=1
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=120400
  Total time spent by all reducers in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=120400
  Total write=114115 bytes taken by all map tasks=140058
  Total mapreduce-write=114115 bytes taken by all map tasks=140058
MapReduce Framework
  Map input records=140058
  Map output records=140058
  Input split bytes=0
  Spilled Records=0
  Failed Shuffles=0
  MapReduce program=0
  GC time elapsed (ms)=954
  CPU time spent (ms)=2770
  Physical memory (bytes) mapped=413414960
  Virtual memory (bytes) mapped=32772464
  Total committed heap usage (bytes)=33402312
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=114115
23/01/29 07:09:18 INFO mapreduce.ImportDatabase: Transferred 954.6059 MB in 44.2824 seconds (11.4403 MB/sec)
23/01/29 07:09:18 INFO mapreduce.ImportDatabase: Retrieved 246052 records.
hadoop@ip-172-31-49-119:~$
```