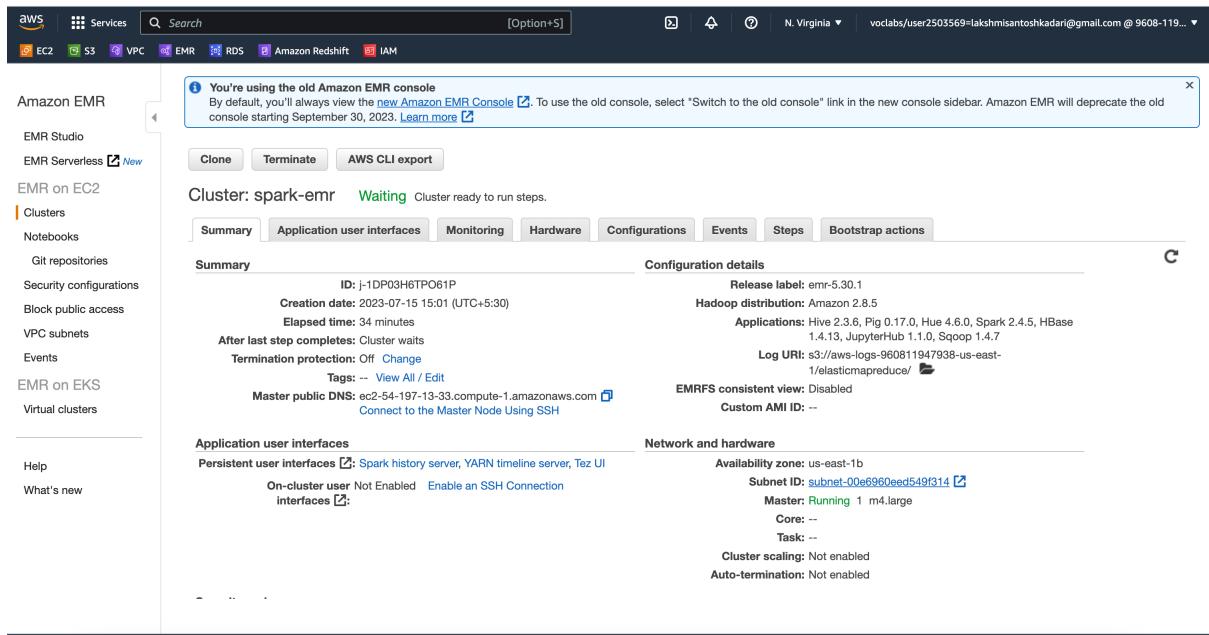


# Data Ingestion

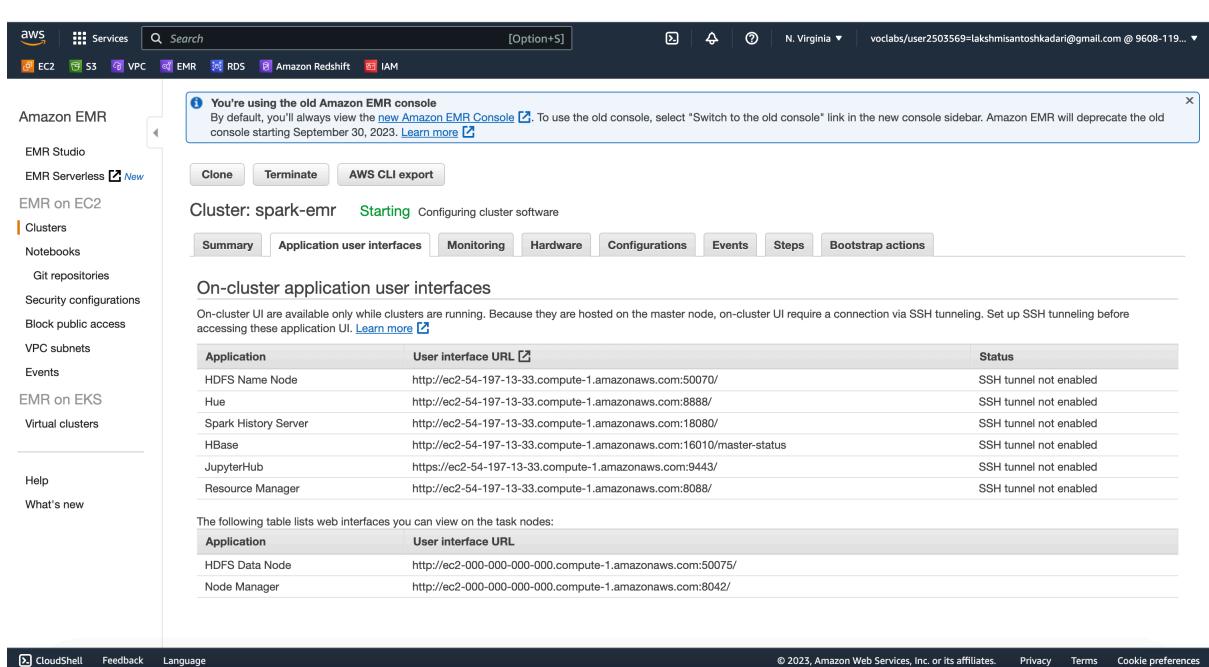
Step 1: Created an EMR cluster as shown in the screenshots.



The screenshot shows the AWS EMR console interface. On the left, there's a sidebar with navigation links for Amazon EMR, EMR Studio, EMR Serverless (with a 'New' button), and sections for EMR on EC2 (Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events) and EMR on EKS (Virtual clusters). The main content area has tabs for Summary, Application user interfaces, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The 'Summary' tab is selected, displaying information about a cluster named 'spark-emr'. The cluster status is 'Waiting' with the message 'Cluster ready to run steps.' Key details include:

- ID: j-1DP03H6TP061P
- Creation date: 2023-07-15 15:01 (UTC+5:30)
- Elapsed time: 34 minutes
- After last step completes: Cluster waits
- Termination protection: Off Change
- Tags: -- View All / Edit
- Master public DNS: ec2-54-197-13-33.compute-1.amazonaws.com
- Connect to the Master Node Using SSH

The 'Configuration details' section shows the release label as 'mr-5.3.0', Hadoop distribution as 'Amazon 2.8.5', and applications like Hive 2.3.6, Pig 0.17.0, Hue 4.6.0, Spark 2.4.5, HBase 1.4.13, JupyterHub 1.1.0, and Sqoop 1.4.7. The 'Network and hardware' section provides details about the subnet ID, master instance type (m4.large), and task configuration.

This screenshot shows the same AWS EMR console interface, but the cluster status has changed to 'Starting'. The main content area displays the 'On-cluster application user interfaces' table, which lists various services and their corresponding URLs:

Application	User interface URL	Status
HDFS Name Node	http://ec2-54-197-13-33.compute-1.amazonaws.com:50070/	SSH tunnel not enabled
Hue	http://ec2-54-197-13-33.compute-1.amazonaws.com:8888/	SSH tunnel not enabled
Spark History Server	http://ec2-54-197-13-33.compute-1.amazonaws.com:18080/	SSH tunnel not enabled
HBase	http://ec2-54-197-13-33.compute-1.amazonaws.com:16010/master-status	SSH tunnel not enabled
JupyterHub	https://ec2-54-197-13-33.compute-1.amazonaws.com:9443/	SSH tunnel not enabled
Resource Manager	http://ec2-54-197-13-33.compute-1.amazonaws.com:8088/	SSH tunnel not enabled

Below this table, there's a note about viewing web interfaces on task nodes, followed by another table for HDFS Data Node and Node Manager URLs.

Step 2: Logging in to the cluster using below command.

```
ssh -i EC2-VPC.pem hadoop@ec2-54-147-45-157.compute-1.amazonaws.com
```

Step3: Installed JDBC connector using below commands.

```
sudo su -  
wget https://dev.mysql.com/get/Downloads/Connector-Java/mysql-connector-java-8.0.25.tar.gz  
tar -xvf mysql-connector-java-8.0.25.tar.gz  
cd mysql-connector-java-8.0.25/  
cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
```

```
[root@ip-172-31-98-110 ~]# tar -xvf mysql-connector-java-8.0.25.tar.gz
[root@ip-172-31-98-110 ~]# cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
[root@ip-172-31-98-110 mysql-connector-java-8.0.25]#
[root@ip-172-31-98-110 mysql-connector-java-8.0.25]# cd
[root@ip-172-31-98-110 ~]#
[root@ip-172-31-98-110 ~]#
[root@ip-172-31-98-110 ~]# ||
```

Step 3: Imported the data from the RDS using below sqoop command:

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--target-dir /user/root/SRC_ATM_TRANS \
-m 1
```

```

root@ip-172-31-98-110 ~# 
root@ip-172-31-98-110 ~# ./sqoop import \
> --connect jdbc:mysql://upgratedate.cyieci9bmmf.us-east-1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --target-dir /user/root/SRC_ATM_TRANS \
> -m 1
Warning: /usr/lib/sqoop../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUACCUMO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.212-b04-0ubuntu1~16.04.1-b17-jdk:/lib/slf4j-j-reload4j-1.7.36.jar!/org/slf4j/jimpl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jarfile:/usr/lib/nive/Slf4j-j14j-impl-2.1.7.1.jar!/org/slf4j/jimpl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jarfile:/usr/lib/nive/Slf4j-j14j-impl-1.7.33.jar!/org/slf4j/jimpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.jimpl.ReloadableLoggerFactory]
2023-07-20 15:28:49.939 INFO sqoop.BasedSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2023-07-20 15:28:50.697 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2023-07-20 15:28:50.698 INFO tool.CodeGenTool: Beginning code generation
>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.
2023-07-20 15:28:50.266 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
2023-07-20 15:28:50.238 INFO manager.SqoopManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
2023-07-20 15:28:50.408 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2023-07-20 15:28:50.968 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/tmp/87fb0eae563ac4a495933572d9e93a8/SRC_ATM_TRANS.jar
2023-07-20 15:28:50.969 INFO manager.MySQLManager: This connection was not explicitly created by the user application. It will be closed after the last activity.
2023-07-20 15:28:50.995 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2023-07-20 15:28:50.995 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNuN (mysql)
2023-07-20 15:28:50.104 INFO mapred.ImportJobBase: Beginning import of SRC_ATM_TRANS
2023-07-20 15:28:50.781 INFO Configuration.deprecation.mapred.jar is deprecated. Instead, use mapreduce.job.jar
2023-07-20 15:28:50.781 INFO Configuration.deprecation.mapred.local.dir is deprecated. Instead, use mapreduce.local.dir
2023-07-20 15:29:00.318 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds(s).
2023-07-20 15:29:00.319 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-07-20 15:29:00.597 INFO db.DBInputFormat: Using read committed transaction isolation
2023-07-20 15:29:00.626 INFO mapreduce.JobSubmitter: number of splits: 1
2023-07-20 15:29:00.626 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1924772745_0001
2023-07-20 15:29:01.476 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-07-20 15:29:02.288 INFO mapred.local.DistributedCacheManager: Creating symlink: /tmp/hadoop-root/reserved/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars <- /root/libjars/*
2023-07-20 15:29:02.213 WARN fs.FileUtil: Command 'ln -s /tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars /root/libjars/*' failed 1 with: ln: failed to create symbolic link '/root/libjars/*': No such file or directory

2023-07-20 15:29:02.214 WARN mapred.local.DistributedCacheManager: Failed to create symlink: /tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars <- /root/libjars/*
2023-07-20 15:29:02.214 INFO mapred.local.DistributedCacheManager: Localized file:/tmp/hadoop/mapred/staging/root1924772745.staging/job_local1924772745_0001/libjars as file:/tmp/hadoop-root/mapred/local/job_local1924772745_0001_ccb764c3-c0e4-47b5-9f7d-fc6f46d32a48/libjars
2023-07-20 15:29:03.370 INFO mapred.LocalDistributedCacheManager: The url to track the job: http://localhost:8888/
2023-07-20 15:29:03.339 INFO mapred.LocalDistributedCacheManager: Running job: job_local1924772745_0001
2023-07-20 15:29:03.339 INFO mapred.LocalDistributedCacheManager: Status after 14 seconds: null
2023-07-20 15:29:07.375 INFO output.FileOutputFormat$OptimizedOutputDirRunOnce: EM Optimized Committer is not supported by org.apache.hadoop.hive ql.io.ProxyLocalFileSystem
2023-07-20 15:29:07.377 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-07-20 15:29:07.378 INFO output.FileOutputCommitter: File Output Committer skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-07-20 15:29:07.379 INFO output.FileOutputCommitter: Direct Write: DISABLED
2023-07-20 15:29:07.380 INFO output.FileOutputCommitter: org.apache.hadoop.mapreduce.lib.output.DirectFileOutputCommitter
2023-07-20 15:29:07.493 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-07-20 15:29:07.496 INFO mapred.LocalJobRunner: Starting task: attempt_local1924772745_0001_0_000000_0
2023-07-20 15:29:07.579 INFO output.FileOutputFormat$OptimizedOutputDirRunOnce: EM Optimized Committer is not supported by org.apache.hadoop.hive ql.io.ProxyLocalFileSystem
2023-07-20 15:29:07.588 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-07-20 15:29:07.588 INFO output.FileOutputCommitter: File Output Committer skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-07-20 15:29:07.589 INFO output.FileOutputCommitter: org.apache.hadoop.mapreduce.lib.output.DirectFileOutputCommitter
2023-07-20 15:29:07.621 INFO mapred.Task: Using ResourceCalculatorPartitionTree: []
2023-07-20 15:29:07.659 INFO db.DBInputFormat: Using read committed transaction isolation

```

```

atm_zipcode", "atm_lat", "atm_lon", "Currency", "Card_Type", "transaction_amount", "service", "message_code", "message_text", "weather_lat", "weather_lon", "weather_city_id", "weather_city_name", "temp", "pressure
2023-07-20 15:29:03.344 INFO mapreduce.Job: Job job_local192477245_0001 running in uber mode : false
2023-07-20 15:29:03.365 INFO mapreduce.Job: map %0 reduce %
2023-07-20 15:29:14.452 INFO mapred.LocalJobRunner: map > map
2023-07-20 15:29:26.676 INFO mapred.LocalJobRunner: mapreduce.JobProgressMapper: mapreduce.JobProgress thread is finished. keepGoing=false
2023-07-20 15:29:26.689 INFO mapred.Task: Task attempt_local192477245_0001_m_000000_0 is done. And is in the process of committing
2023-07-20 15:29:26.163 INFO mapred.Task: map > map
2023-07-20 15:29:26.163 INFO mapred.Task: Task attempt_local192477245_0001_m_000000_0 is allowed to commit now
2023-07-20 15:29:26.173 INFO mapred.Task: OutputCommitter: Saved output of task attempt_local192477245_0001_m_000000_0' to file:/user/root/SRC_ATM_TRANS
2023-07-20 15:29:26.173 INFO mapred.Task: Task attempt_local192477245_0001_m_000000_0' done.
2023-07-20 15:29:26.208 INFO mapred.Task: Final Counters for attempt_local192477245_0001_m_000000_0: Counters: 16
  File System Counters
    FILE: Number of bytes read=31564
    FILE: Number of bytes written=530468302
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=2468572
    Map output records=2468572
    Input split bytes=86
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=72
    Total committed heap usage (bytes)=1587544064
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=535364999
2023-07-20 15:29:26.288 INFO mapred.LocalJobRunner: Finishing task: attempt_local192477245_0001_m_000000_0
2023-07-20 15:29:26.281 INFO mapred.LocalJobRunner: map task executor complete.
2023-07-20 15:29:26.381 INFO mapreduce.Job: map 100% reduce 0%
2023-07-20 15:29:26.381 INFO mapreduce.Job: Job job_local192477245_0001 completed successfully
2023-07-20 15:29:26.381 INFO mapreduce.Job: Counters: 10
  File System Counters
    FILE: Number of bytes read=31564
    FILE: Number of bytes written=530468302
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=2468572
    Map output records=2468572
    Input split bytes=86
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=71
    Total committed heap usage (bytes)=1587544064
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=535364999
2023-07-20 15:29:26.392 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 26.3556 seconds (0 bytes/sec)
2023-07-20 15:29:26.392 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-98-110 ~]# 

```

Step 4: Checked the data from the location using below command to ensure it is imported.

```
hadoop fs -ls /user/root/SRC_ATM_TRANS
```

```
[hadoop@ip-172-31-90-110 ~]$  
[hadoop@ip-172-31-90-110 ~]$ sudo su -  
Last login: Thu Jul 20 15:22:11 UTC 2023 on pts/0  
  
EEEEEEEEEEEEEE MMMMMMM          MMMMMMM RRRRRRRRRRRRRR  
E:::::::::::E M:::::M          M:::::M R:::::::::::R  
EE:::::EEEEE:::E M:::::M          M:::::M R:::::RRRRR:::::R  
   E:::E     EEEE M:::::M          M:::::M RR:::::R      R:::::R  
E:::::E     M:::::M::::M          M:::::M R:::::R      R:::::R  
EE:::::EEEEE E M::::M M::::M          M:::::M R:::::RRRR:::::R  
E:::::::::::E M:::::M M:::::M          M:::::M R:::::::::::RR  
E:::::EEEEE E M:::::M M:::::M          M:::::M R:::::RRRRR:::::R  
E:::::E     M:::::M M:::::M          M:::::M R:::::R      R:::::R  
E:::::E     EEEE M:::::M          M:::::M R:::::R      R:::::R  
EE:::::EEEEE E M:::::M          M:::::M R:::::R      R:::::R  
E:::::::::::E M:::::M          M:::::M RR:::::R      R:::::R  
EEEEEEEEEEEEEE MMMMM          MMMMM RRRRRRR  
  
[root@ip-172-31-90-110 ~]# hadoop fs -copyFromLocal /user/root/SRC_ATM_TRANS/ /user/root/  
[root@ip-172-31-90-110 ~]#  
[root@ip-172-31-90-110 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS  
Found 2 items  
-rw-r--r-- 1 root hdfsadmingroup 0 2023-07-20 15:38 /user/root/SRC_ATM_TRANS/_SUCCESS  
-rw-r--r-- 1 root hdfsadmingroup 531214815 2023-07-20 15:38 /user/root/SRC_ATM_TRANS/part-m-00000  
[root@ip-172-31-90-110 ~]#  
[root@ip-172-31-90-110 ~]#  
[root@ip-172-31-90-110 ~]#
```