Loading Data into HDFS using Sqoop

Sqoop Command Used for the import:

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student  --password STUDENT123 \
--null-string ‘\\N’ –null-non-string ‘\\N’  \
--target-dir /user/root/atm_data \
-m 1
```

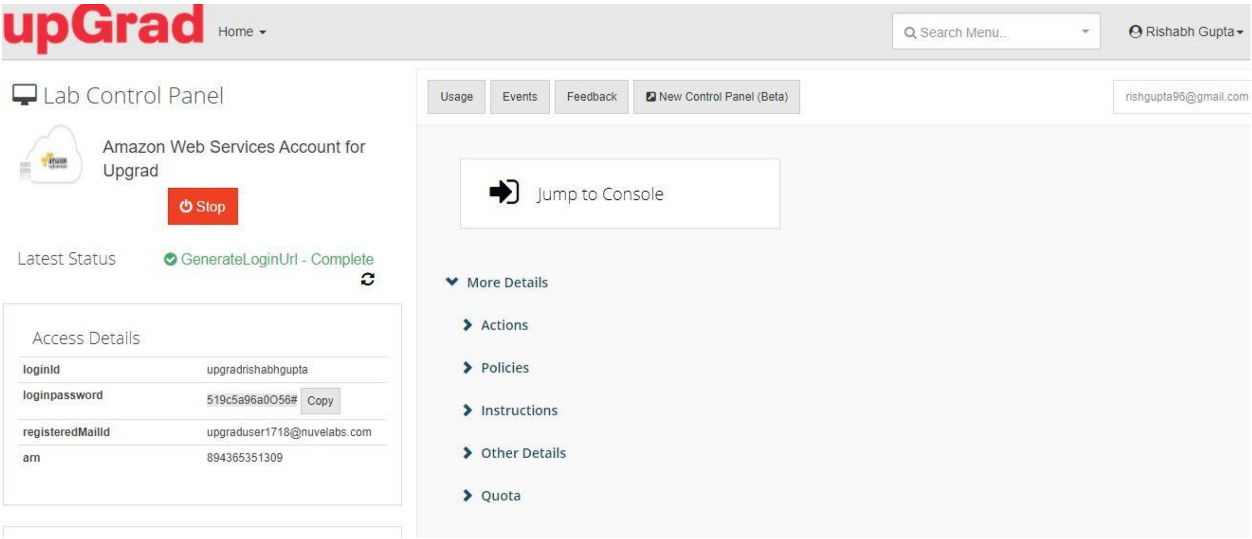
Command used to check the records inserted in HDFS:

```
hadoop fs  –ls  /user/root/atm_data
```

When we used this command we saw 2 files stored in this folder – one is SUCCESS_ 0 byte file and other is the part* file which actually contains our records. Another check for this was to after the MapReduce job ran it gave the number of records it imported. The number in our case exactly matched the validation given for the assignment

Steps Involved with Screenshots:

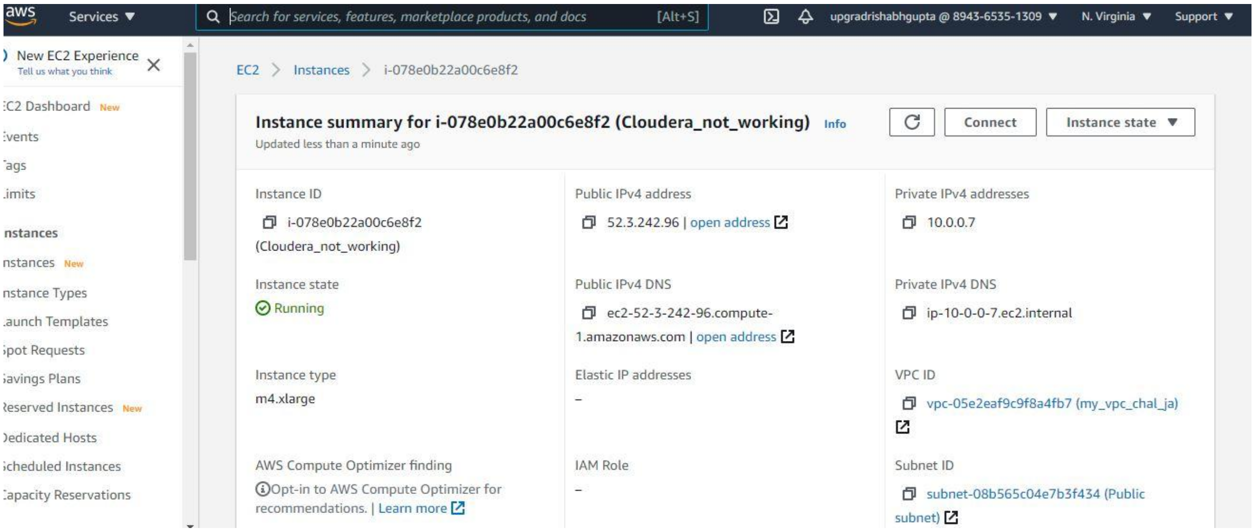
Step 1: Opening AWS dashboard using Nuvepro credentials



Latest Status GenerateLoginUrl - Complete

Access Details

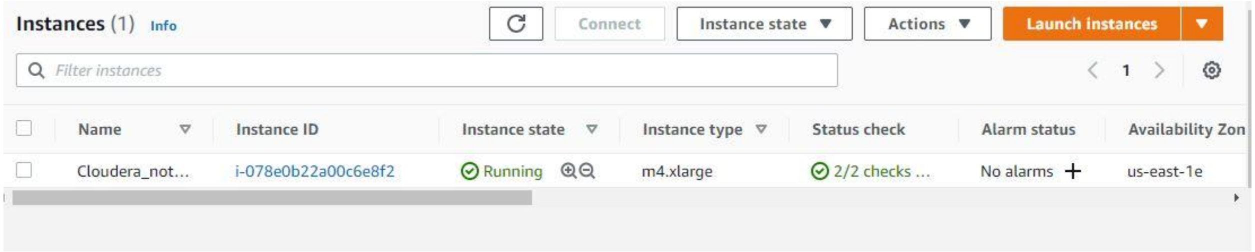
loginId	upgradrishabhgupta
loginpassword	519c5a96a0O56# Copy
registeredMailId	upgraduser1718@nuvelabs.com
arn	894365351309



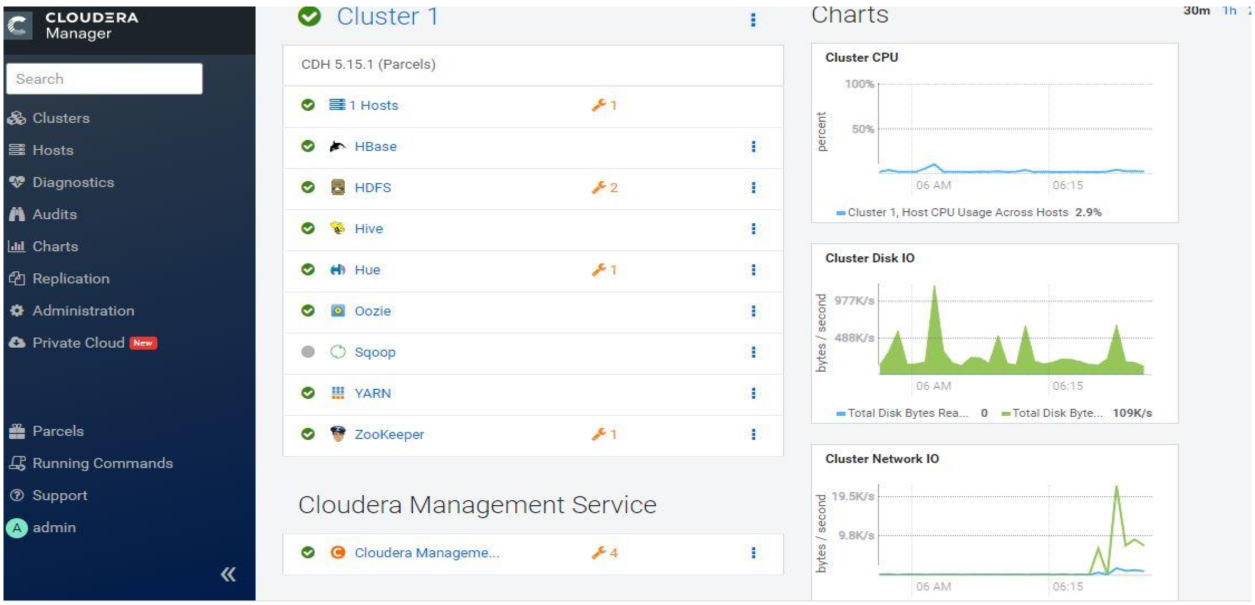
Instance summary for i-078e0b22a00c6e8f2 (Cloudera_not_working) Info
Updated less than a minute ago

Instance ID i-078e0b22a00c6e8f2 (Cloudera_not_working)	Public IPv4 address 52.3.242.96 | open address	Private IPv4 addresses 10.0.0.7
Instance state Running	Public IPv4 DNS ec2-52-3-242-96.compute-1.amazonaws.com | open address	Private IPv4 DNS ip-10-0-0-7.ec2.internal
Instance type m4.xlarge	Elastic IP addresses –	VPC ID vpc-05e2eaf9c9f8a4fb7 (my_vpc_chal_ja)
AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendations. | Learn more	IAM Role –	Subnet ID subnet-08b565c04e7b3f434 (Public subnet)

Step 2: Running EC2 instance and using PuTTy for interacting with Hadoop ecosystem



□	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zon
□	Cloudera_not…	i-078e0b22a00c6e8f2	Running	m4.xlarge	2/2 checks …	No alarms +	us-east-1e



Cluster 1

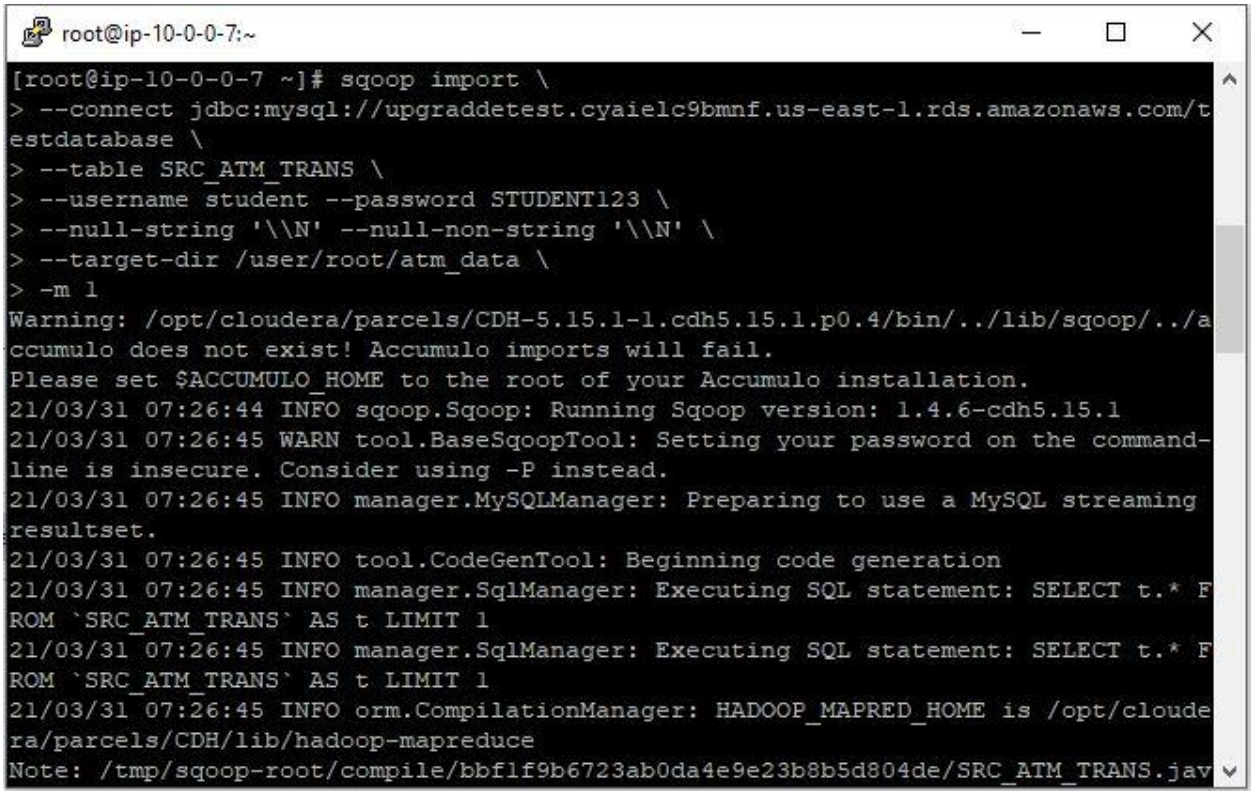
CDH 5.15.1 (Parcels)

Service	Alerts
1 Hosts	1
HBase	
HDFS	2
Hive	
Hue	1
Oozie	
Sqoop	
YARN	
ZooKeeper	1

Cloudera Management Service

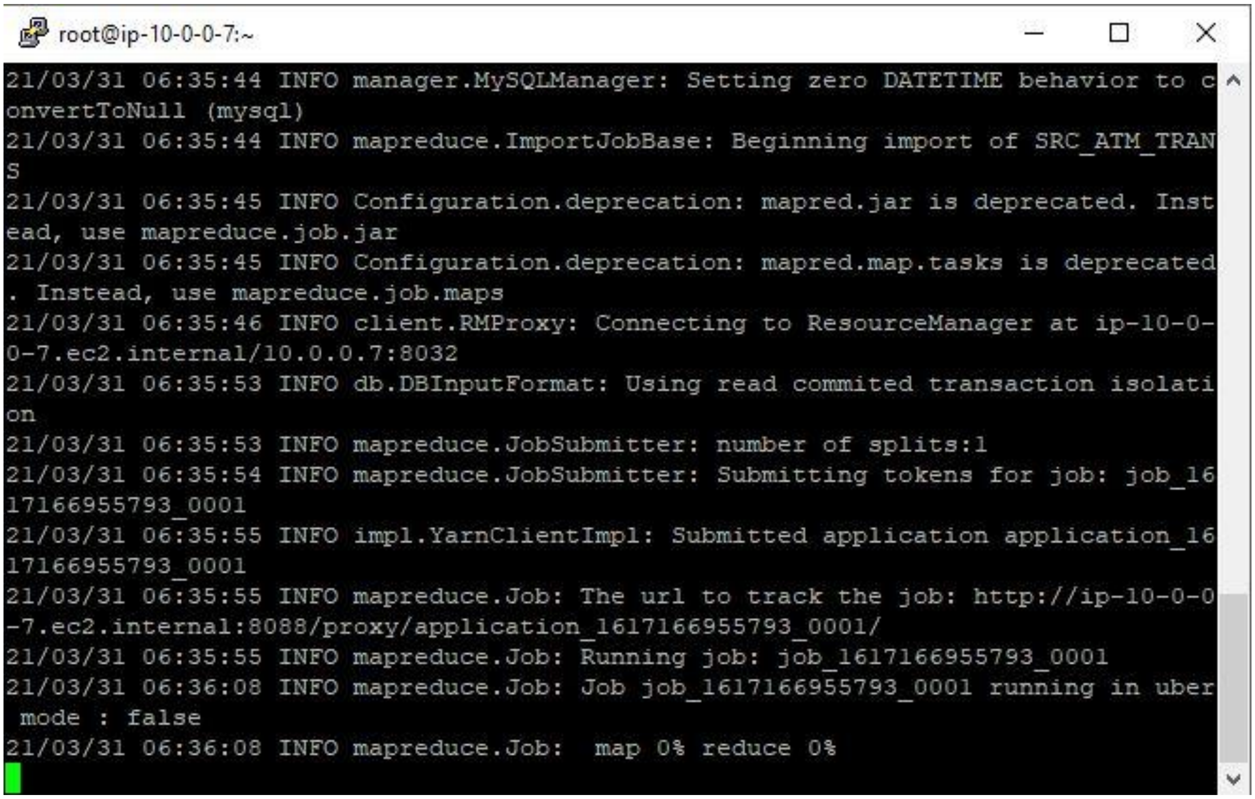
Service	Alerts
Cloudera Manageme...	4

Step3 : Creating directory in hadoop where we want to import data from RDB and then writing sqoop import command to actually import and store data in HDFS



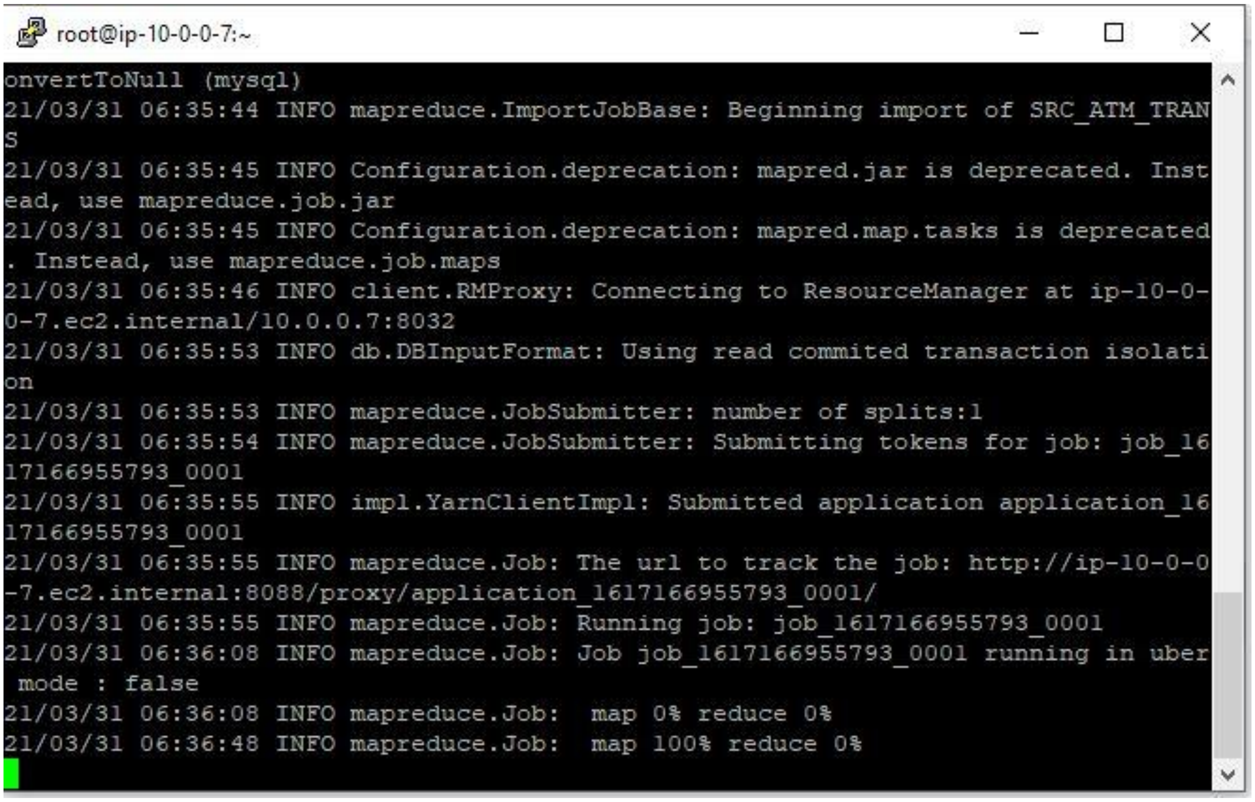
```
[root@ip-10-0-0-7 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/t
estdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/atm_data \
> -m 1
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/../a
ccumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
21/03/31 07:26:44 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
21/03/31 07:26:45 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
21/03/31 07:26:45 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
21/03/31 07:26:45 INFO tool.CodeGenTool: Beginning code generation
21/03/31 07:26:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `SRC_ATM_TRANS` AS t LIMIT 1
21/03/31 07:26:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `SRC_ATM_TRANS` AS t LIMIT 1
21/03/31 07:26:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloude
ra/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/bbf1f9b6723ab0da4e9e23b8b5d804de/SRC_ATM_TRANS.jav
```

Step 4: Map Job begins



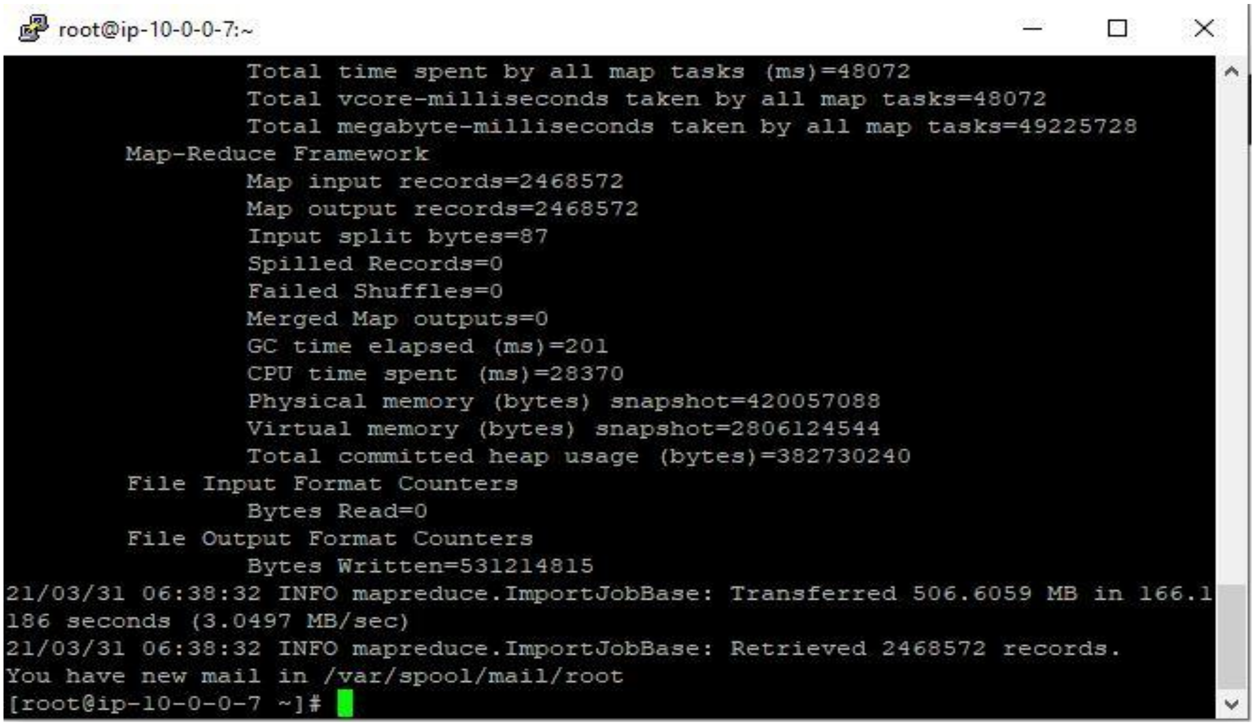
```
21/03/31 06:35:44 INFO manager.MySQLManager: Setting zero DATETIME behavior to c
onvertToNull (mysql)
21/03/31 06:35:44 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRAN
S
21/03/31 06:35:45 INFO Configuration.deprecation: mapred.jar is deprecated. Inst
ead, use mapreduce.job.jar
21/03/31 06:35:45 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
21/03/31 06:35:46 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-
0-7.ec2.internal/10.0.0.7:8032
21/03/31 06:35:53 INFO db.DBInputFormat: Using read commited transaction isolati
on
21/03/31 06:35:53 INFO mapreduce.JobSubmitter: number of splits:1
21/03/31 06:35:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
17166955793_0001
21/03/31 06:35:55 INFO impl.YarnClientImpl: Submitted application application_16
17166955793_0001
21/03/31 06:35:55 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0
-7.ec2.internal:8088/proxy/application_1617166955793_0001/
21/03/31 06:35:55 INFO mapreduce.Job: Running job: job_1617166955793_0001
21/03/31 06:36:08 INFO mapreduce.Job: Job job_1617166955793_0001 running in uber
 mode : false
21/03/31 06:36:08 INFO mapreduce.Job:  map 0% reduce 0%
```

Step 5: Map job finished running 100 %



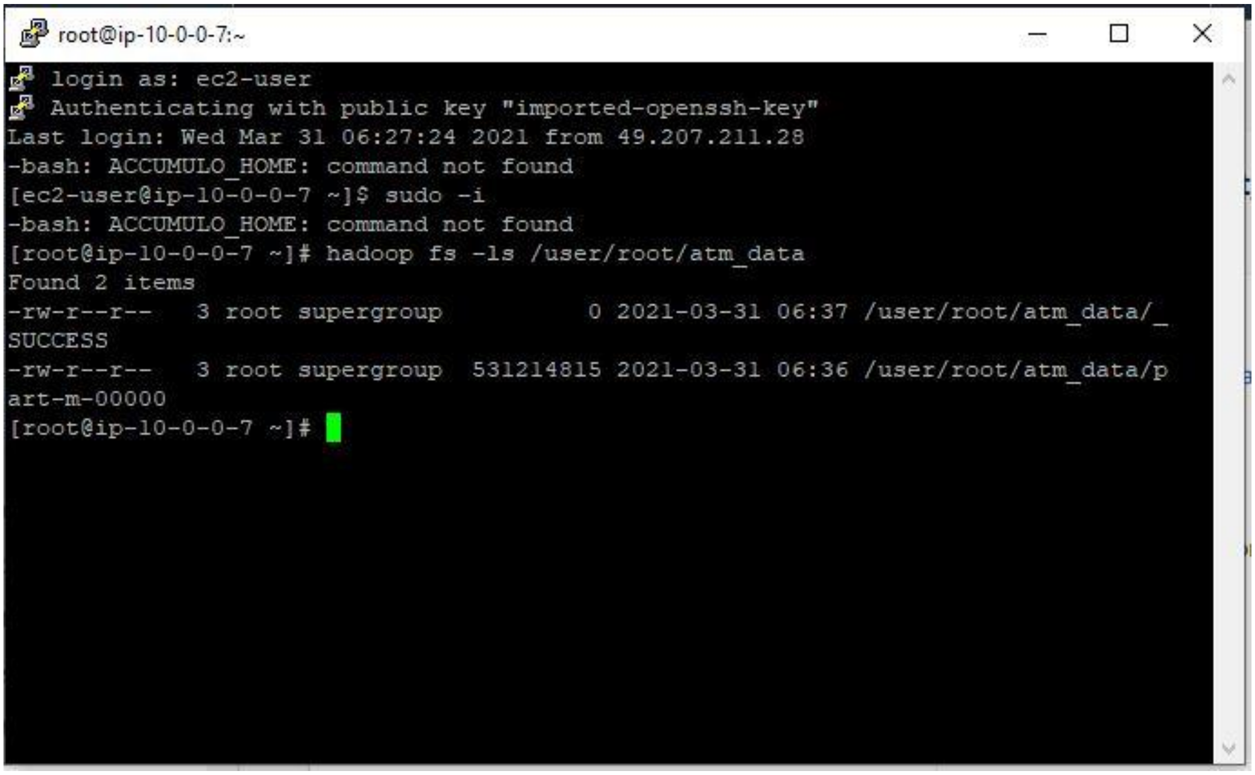
```
onvertToNull (mysql)
21/03/31 06:35:44 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRAN
S
21/03/31 06:35:45 INFO Configuration.deprecation: mapred.jar is deprecated. Inst
ead, use mapreduce.job.jar
21/03/31 06:35:45 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
21/03/31 06:35:46 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-
0-7.ec2.internal/10.0.0.7:8032
21/03/31 06:35:53 INFO db.DBInputFormat: Using read commited transaction isolati
on
21/03/31 06:35:53 INFO mapreduce.JobSubmitter: number of splits:1
21/03/31 06:35:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
17166955793_0001
21/03/31 06:35:55 INFO impl.YarnClientImpl: Submitted application application_16
17166955793_0001
21/03/31 06:35:55 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0
-7.ec2.internal:8088/proxy/application_1617166955793_0001/
21/03/31 06:35:55 INFO mapreduce.Job: Running job: job_1617166955793_0001
21/03/31 06:36:08 INFO mapreduce.Job: Job job_1617166955793_0001 running in uber
 mode : false
21/03/31 06:36:08 INFO mapreduce.Job:  map 0% reduce 0%
21/03/31 06:36:48 INFO mapreduce.Job:  map 100% reduce 0%
```

Step 6: After the sqoop job runs, it spits out the number of records it copied into the HDFS, in our case matches the validation



```
                Total time spent by all map tasks (ms)=48072
                Total vcore-milliseconds taken by all map tasks=48072
                Total megabyte-milliseconds taken by all map tasks=49225728
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=201
                CPU time spent (ms)=28370
                Physical memory (bytes) snapshot=420057088
                Virtual memory (bytes) snapshot=2806124544
                Total committed heap usage (bytes)=382730240
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
21/03/31 06:38:32 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 166.1
186 seconds (3.0497 MB/sec)
21/03/31 06:38:32 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
You have new mail in /var/spool/mail/root
[root@ip-10-0-0-7 ~]#
```

Step 7: Check the hadoop directory for creation of file



```
login as: ec2-user
Authenticating with public key "imported-openssh-key"
Last login: Wed Mar 31 06:27:24 2021 from 49.207.211.28
-bash: ACCUMULO_HOME: command not found
[ec2-user@ip-10-0-0-7 ~]$ sudo -i
-bash: ACCUMULO_HOME: command not found
[root@ip-10-0-0-7 ~]# hadoop fs -ls /user/root/atm_data
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-03-31 06:37 /user/root/atm_data/_
SUCCESS
-rw-r--r--   3 root supergroup  531214815 2021-03-31 06:36 /user/root/atm_data/p
art-m-00000
[root@ip-10-0-0-7 ~]# 
```