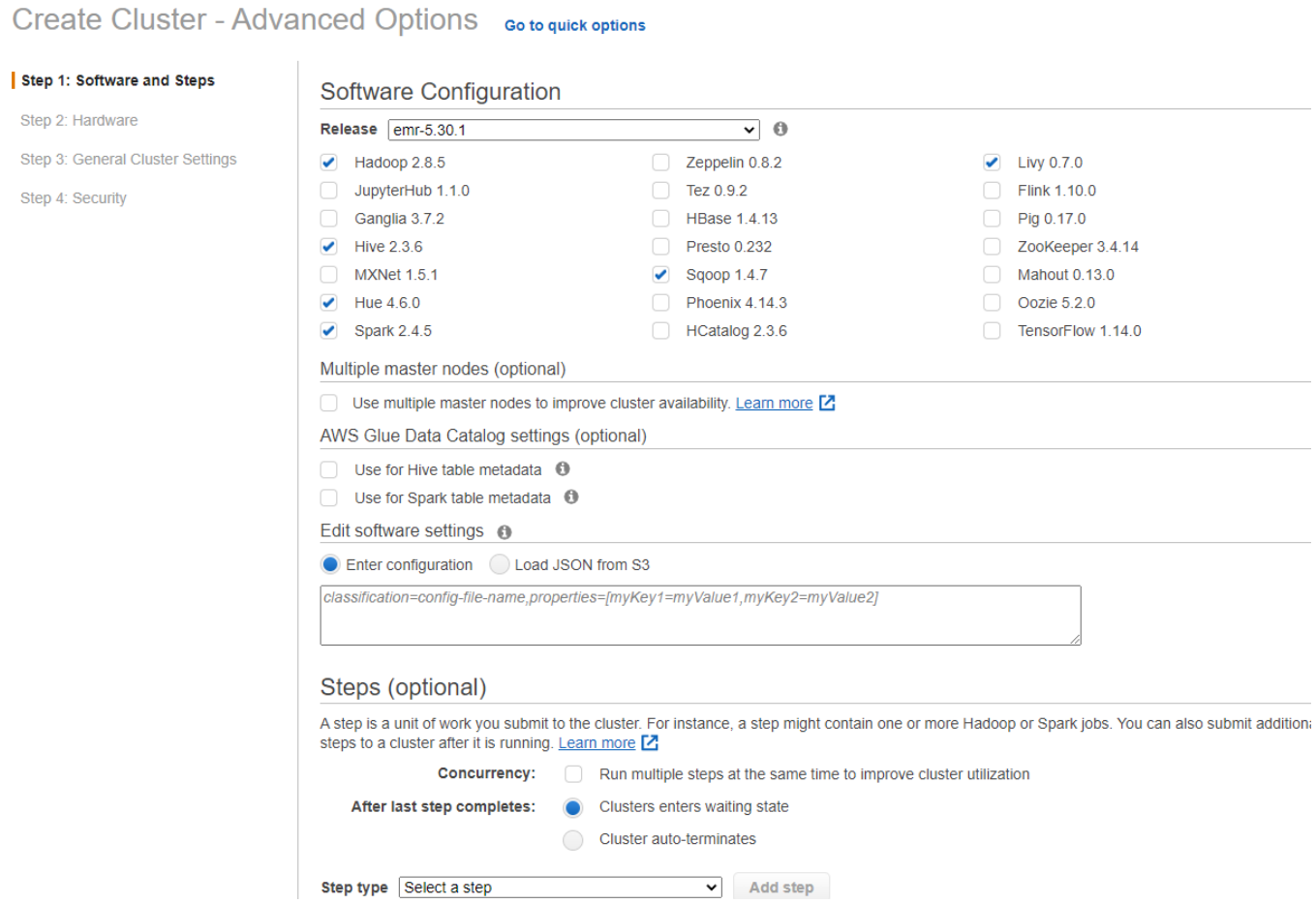
Data Ingestion from the RDS to HDFS using Sqoop

For Ingesting the data from RDS to HDFS, we are going to use the Apache Sqoop service in our EMR Cluster.
Below are the services used in provisioning EMR Cluster:



We are using Hadoop to setup the HDFS file system. We are using Hive, Hue and Spark for processing the big data and livy is used to access the data in cluster notebooks.

Sqoop:

Sqoop Import command used for importing table from RDS to HDFS:

- Sqoop connects to the database, it uses JDBC(Java Database Connectivity) to examine the table to be imported by retrieving a list of all the columns and their SQL data types.

- The SQL data types (integer, varchar, etc.) can be mapped to Java data types (integer, string, etc.).
- Sqoop has a code generator, which creates a table-specific Java class to hold the extracted records from the table by using the information provided by the JDBC about the data types.
- Then, Sqoop connects to the cluster to submit a MapReduce job using the Java class that is generated. The data set being transferred is split into multiple partitions and a map-only job is launched, which outputs a set of files containing the imported data.

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east1.rds.amazonaws.com/testdatabase\
--table SRC_ATM_TRANS \
--username student \
--password STUDENT123 \
--target-dir /home/hadoop/ETL_Project_SRC_data \
-m 1
```

After running this command, the data from the "SRC_ATM_TRANS" table in MySQL database – "test" will be imported to the HDFS cluster.

'import' keyword is used to define that this is an import command. The '**--target-dir**' argument is used to define the target directory where the MySQL table has to be imported, in this case: /home/hadoop/ETL_Project_SRC_data. The '**-m**' argument is used to set the number of mappers used for this job.

Command used to see the list of imported data in HDFS:

- Command used : hadoop fs -ls /home/hadoop/ETL_Project_SRC_data

Screenshot of the imported data:

```
[hadoop@ip-10-0-0-21 ~]$ hadoop fs -ls /home/hadoop/ETL_Project_SRC_data
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-04-30 11:37 /home/hadoop/ETL_Project_SRC_data/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2022-04-30 11:37 /home/hadoop/ETL_Project_SRC_data/part-m-00000
[hadoop@ip-10-0-0-21 ~]$
```