

## **Lab – Data Processing using EMR – Step Execution – Hive Script**

**Step 1:** Create a S3 bucket name “retail-db” and upload the files/directories from the link [https://github.com/aws-training/retail\\_db.git](https://github.com/aws-training/retail_db.git) or download from the following link “[https://github.com/aws-training/retail\\_db/archive/refs/heads/main.zip](https://github.com/aws-training/retail_db/archive/refs/heads/main.zip)” and upload all the contents.

**Step 2:** In AWS Console, select EMR and Click on create a cluster and give a name or use default. Select step execution and step type as Hive program. Select latest release and select m4.large machines(1).

Create Cluster - Quick Options [Go to advanced options](#)

#### General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☐ Cluster ⓘ ☒ Step execution ⓘ

#### Add steps

A step is a unit of work submitted to an application running on your EMR cluster. EMR programmatically installs the applications needed to execute the added steps. [Learn more](#)

Name	Action on failure	JAR location	Arguments
Step type <input type="text" value="Hive program"/> <input type="button" value="Configure"/>			

#### Software configuration

Release  ⓘ

Applications  ⓘ

#### Hardware configuration

Instance type  ⓘ The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#)

**Step 3:** Click Configure beside the step type and select the script location from S3 (s3://retail-db)

Add step

Step type

Name

Script S3 location\*   
s3://<bucket-name>/<path-to-file>

Input S3 location   
s3://<bucket-name>/<folder>/

Output S3 location   
s3://<bucket-name>/<folder>/

Arguments   
Specify optional arguments for your script.

Action on failure  ⓘ What happens if the step fails

**Step 4:** Click create cluster and wait till the status changes to **Waiting**

**Step 5:** Keep checking the status of your cluster and when it starts terminating go to S3 bucket and check if a directory is created with name “daily\_product\_revenue” .

**Step 6:** Click the file under the “daily\_product\_revenue” folder and query it through S3 select to check the contents