Linux Academy

Live! Lab

Analyzing Log Files with EMR

Contents

Lab Connection Information

- Labs may take up to five minutes to build
- Access to an AWS Console is provided on the Live! Lab page, along with your login credentials
- Ensure you are using the N. Virginia region
- Labs will automatically end once the alloted amount of time finishes

Related Courses

AWS CSA - Pro Course

Related Videos

Amazon EMR

Need Help?

Linux Academy Community

... and you can always send in a support ticket on our website to talk to an instructor!

Introduction

Amazon Elastic MapReduce (EMR) provides users a cost-effective way to process large amounts of data. In this lab, we use EMR to review NASA web server log files and group the data by 404 entries.

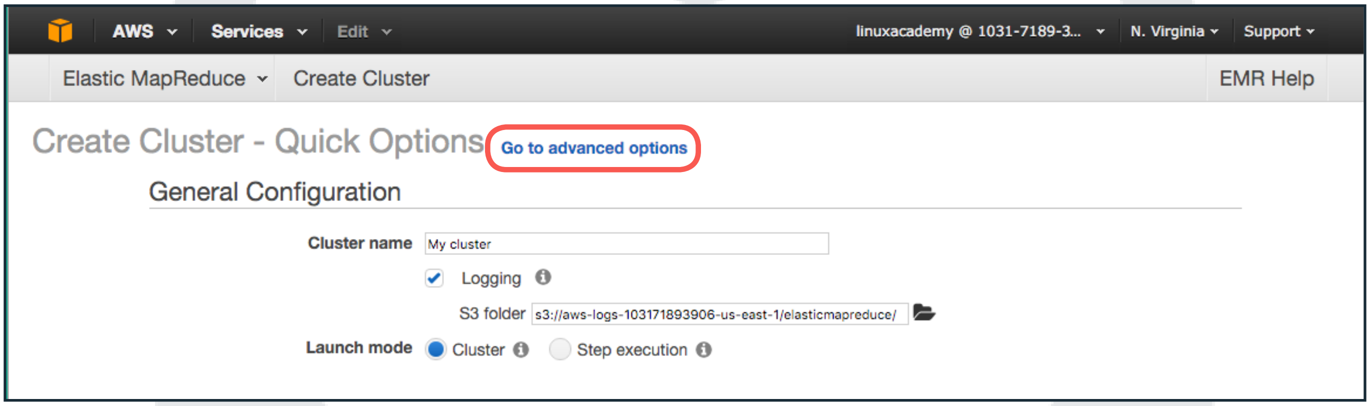
Log into the AWS console using the provided credentials before continuing.

Creating the EMR Job

For the following steps, we need to switch between our **Amazon S3** and **EMR dashboards**. Open both in separate tabs in your browser to streamline the process.

Upon viewing the S3 dashboard, you should see an already-created bucket containing your Linux Academy username. Enter the bucket to see a folder, *data*, which contains the NASA data, and a *hive.q* script. Download the script to review later in this lab.

Switch to the EMR tab. You may be able to view previously terminated instances, but these do not concern us. Press **Create cluster**, then **Go to advanced options**.



Create Cluster - Quick Options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-103171893906-us-east-1/elasticmapreduce/

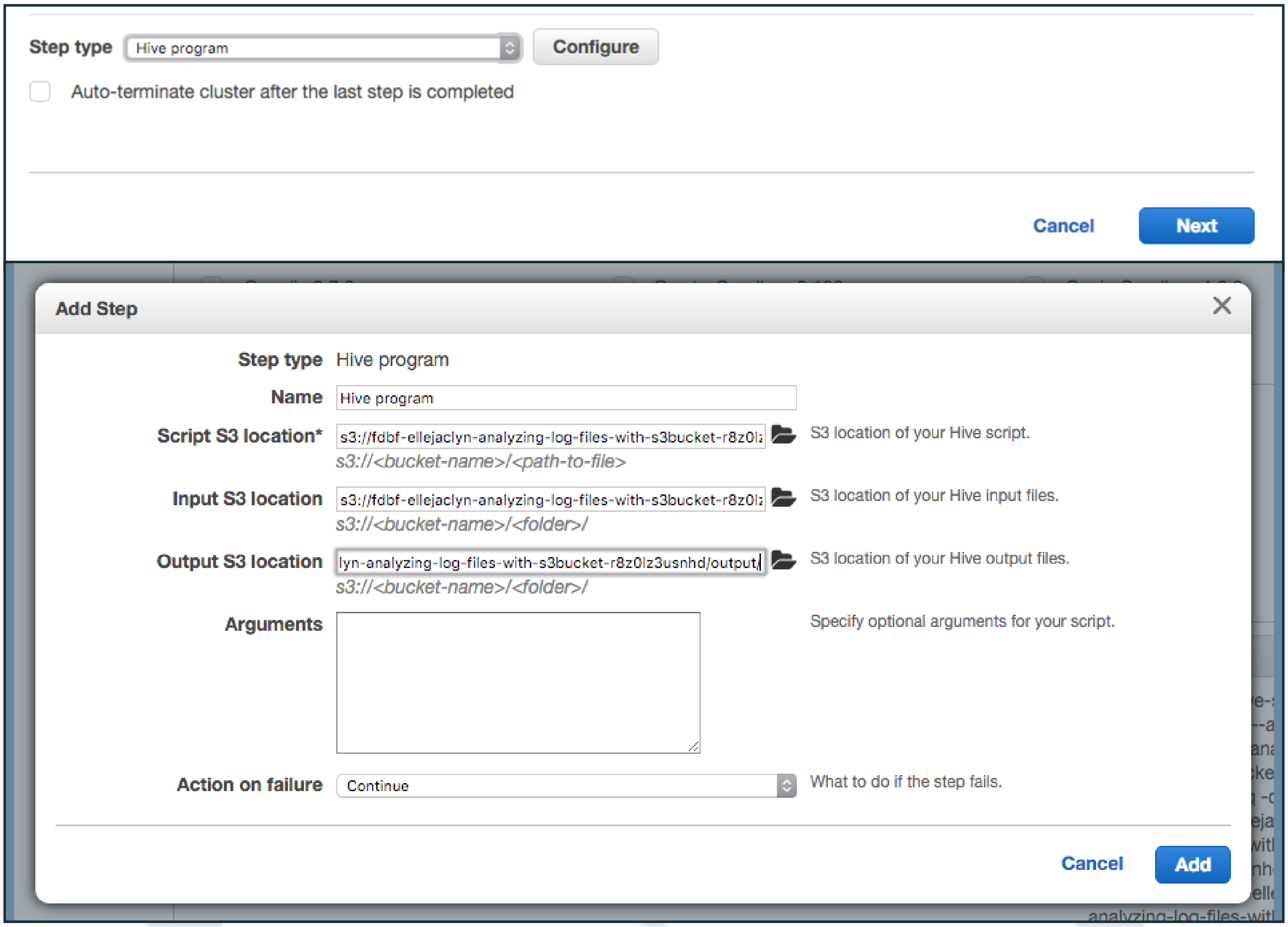
Launch mode: ● Cluster ○ Step execution

Under **Software Configuration**, you can leave the settings as-is, ensuring that *Hive*, a SQL-query language, is selected.

Move to **Add steps (optional)** and select *Hive program* as the **Step type**. Press **Configure**. It will ask you for a script.

For the **Script S3 location**, choose your bucket, then select the *hive.q* file. **Input S3 location** should then be set to the *data/* folder (**not** the file in the folder).

The **Output S3 location** has yet to be made. We do not want to select our root or data folders because the current data (including the script and NASA logs) will be removed. Instead, return to the S3 dashboard, enter the bucket containing your username, and create a folder named *output*. From the EMR dashboard, select the newly-created folder at the **Output S3 location**.



Step type: Hive program — Configure

□ Auto-terminate cluster after the last step is completed

Add Step

Step type: Hive program

Name: Hive program

Script S3 location*: s3://fdbf-ellejaclyn-analyzing-log-files-with-s3bucket-r8z0lz — S3 location of your Hive script.
s3://<bucket-name>/<path-to-file>

Input S3 location: s3://fdbf-ellejaclyn-analyzing-log-files-with-s3bucket-r8z0lz — S3 location of your Hive input files.
s3://<bucket-name>/<folder>/

Output S3 location: lyn-analyzing-log-files-with-s3bucket-r8z0lz3usnhd/output/ — S3 location of your Hive output files.
s3://<bucket-name>/<folder>/

Arguments: — Specify optional arguments for your script.

Action on failure: Continue — What to do if the step fails.

Since we do not need to pass in any arguments, select **Add** to return to the initial configuration page. Press **Next**.

On the following page, ensure that the **Network** selected is the one containing a subnet. Make sure the subnet is selected.

The task instance groups should be determined based on cost, which is addressed in greater detail in the related video. Due to the nature of this lab, we are leaving the initial groups, but changing the **EC2 instance type** to *c1.medium* for all three types. Select **Next**.

You can leave the **Cluster name** as-is, or select one more specific to the task. The logging, debugging, and termination checkboxes that follow can also be left as-is. Press **Next**.

For **Permissions**, press *Custom*, The field below will auto-populate. Because we do not need to directly log into the instances, no EC2 key pairs are needed. **Create cluster**.

If you view **Steps** on this page, you can see the steps EMR is taking to process the data.

The Hive Script

While the EMR cluster processes data, we can review the Hive script. Open the *hive.q* document to view it:

```
CREATE EXTERNAL TABLE IF NOT EXISTS nasa_logs (
Source STRING,
Date STRING,
Request STRING,
Status STRING
)


ROW FORMAT SERDE “org.apache.hadoop.hive.serde2.RegexSerDe”
WITH SERDEPROPERTIES(
“input.regex” = “^([^ ]*) - - \\[([^ ]*[\\d]{4})[^ ]* [^ ]* \”([^ ]* [^ ]*) [^ ]* ([\\d]
{3}) [^ ]*$”,
“output.format.string” = “%1$s %2$s %3$s %4$s”
)


LOCATION ‘${INPUT}’;


INSERT OVERWRITE DIRECTORY ‘${OUTPUT}’
select date, source, request, status, count(*) count
from nasa_logs
where status = “404”
group by date, source, request, status;
```

The code begins by creating a table for the imported data (lines 1-6). The table, called *nasa_logs*, contains rows for the source, date, request, and status.

The next segment (lines 9-12) takes the information and converts each line to seperate fields to populate the table.

Finally, the `INSERT OVERWRITE DIRECTORY` segment sorts the information by status and then counts the amount of logs that include a *404* status.

The Results

Refresh the EMR dashboard and view the **Steps** section again. It can take anywhere from ten to twenty minutes for the results to process, but once the **Status** for all steps reads as *complete*, you are ready to move on in the lab.

Return to the S3 bucket, and open the **output** folder. The results are located in a single file. Download the file to your workstation computer.

Open the file in your text editor of choice, ignoring any characters that seem to be inserted in error. Each line will resemble:

```
01/Aug/1995^A128.158.42.193^AGET /pub^A404^A3
```

This data includes the requested information from the hive script, above. The first segment (`01/Aug/1995`) includes the date, the source is the IP address (`128.158.42.193`), request is `GET`, and the status is `404`. The last number, `3` in the example above, is how many times this result was hit.

With the data processed and reviewed, you can now complete the lab.