Analyze Big Data with Hadoop

Step 1

Click **Start Lab** to launch lab

Step 2

Click **Open Console**, which will automatically log you in to the AWS Console.

Task 1: Create an Amazon S3 bucket

Step 3

In the **AWS Management Console**, on the **Service** menu, click **S3**

Step 4

Click **Create bucket**

Step 5

For **Bucket name**, enter **hadoop-** folllowed by a random number.

Step 6

Click **Create**

Task 2: Launch an Amazon EMR cluster

Step 7

On the **Services** menu, click **EMR**

Step 8

Click **Create cluster**

Step 9

In the **General Configuration** section, configure the following:

```
- Cluster name: My cluster
- S3 folder:
    Click the folder icon
    Click the hadoop- bucket that you created eariler
    Click Select
```

Step 10

In the **Hardware configuration** section, configure:

```
- Instance type: m4.large
- Number of instances: 2
```

Step 11

In the **Security and access** section, configure:

```
- EC2 key pair: Proceed without an EC2 key pair
- Permissions: Custom
- EMR role: EMR_DefaultRole
- EC2 instance profile: EMR_EC2_DefaultRole
```

Step 12

Click **Create cluster** to launch your EMR cluster

```
Hadoop Applications
- Apache Hadoop
- Ganglia
- Apache Tez
- Hive
- Hue
- Pig
```

Task 3: Process Your Sample Data by Running a Hive Script

Step 13

Wait until your cluster is showing a status of Waiting.

Step 14

Click the Steps tab

Step 15

Click Add step

Step 16

In the Add step dialog, configure the following settings:

```
- Step type: Hive program
- Nmae: Process logs
- Script S3 locationn: Copy and paste this location
    s3://us-west-2.elasticmapreduce.samples/cloudfront/code/Hive_Clo
- Input S3 location: Copy and paste:
    s3://us-west-2.elasticmapreduce.samples
```

```
- Output S3 location: Click the folder icon and select the hadoop- b
- Arguments: Copy and paste
    hiveconf hive.support.sql11.reserved.keywords=false
- Click Add
```

The Hive script does the following:

```
Creates a Hive table named cloudfront_logs
Reads the CloudFront log files from Amazon S3 and parses the files u
Writes the parsed results to the cloudfront_logs Hive table.
Subbmits a HiveQL query against the data to retrieve the total reque
Writes the query results to your Amazon S3 ouput bucket.
```

Step 17

Wait for the status of the step to change to Completed

Task 4: View the Results

Step 18

On the Services menu, click S3

Step 19

Click on the name of the hadoop- bucket

Step 20

Click the os_requests folder

The Hive query results are stored in a text file

Step 21

Select the 000000_0 file

Step 22

Click Download in the pop-up box and save the file to your computer

Step 23

Open the file using a text editor such as WordPad, TextEdit, or gEdit

In the output file, you should see the number of access requests by operating system

Task 5: Terminate your Amazon EMR Cluster

Step 24

On the Services menu, click EMR

Step 25

Select My cluster

Step 26

Click Terminate

Step 27

In the Terminnate cluster dialog, click Terminate.