Building Batch Data Analytics Solutions on AWS:

Lab 2 - Batch Data Processing using Amazon EMR with Hive

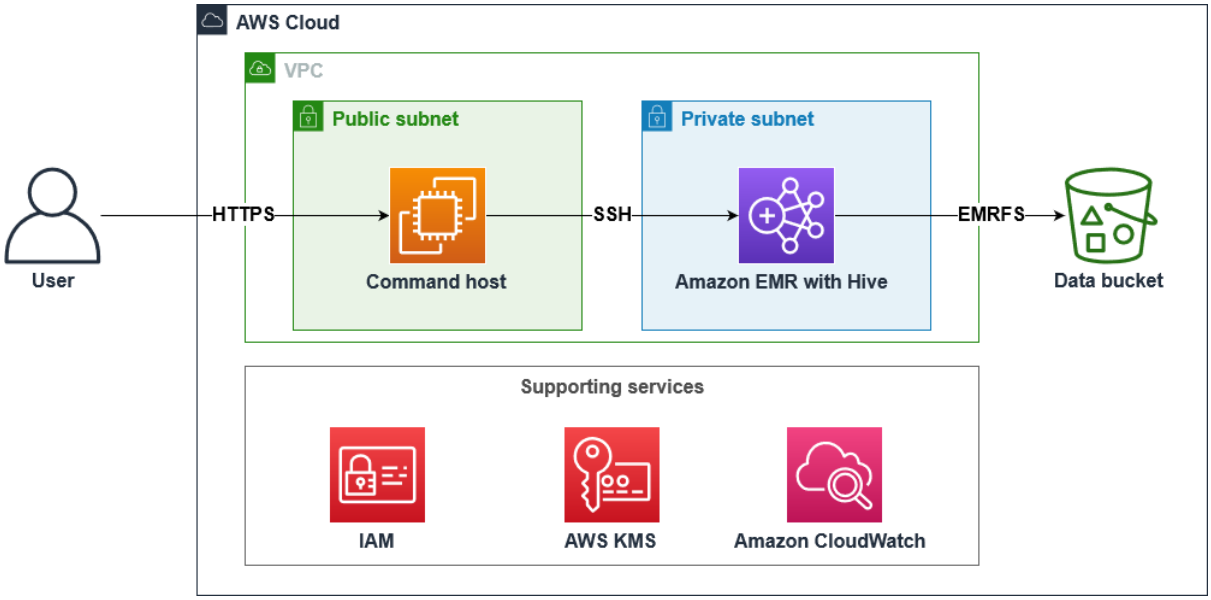
Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at *AWS Training and Certification*.

Lab overview

You have already reviewed the capabilities of the Amazon EMR architecture. Now, you are given some sample raw data to find the optimum solution that can scale up to petabytes during Any Company Financials' production rollout. You decide to use Amazon EMR and Apache Hive to meet this need.

Your task is to load the sample data in Amazon Simple Storage Service (Amazon S3). You will then connect to the EMR cluster, create an Apache Hive table, load data from Amazon S3, and run queries using HiveQL.



OBJECTIVES

By the end of this lab, you will be able to:

Review how Amazon EMR and Apache Hive can be used together to ingest and query data

Identify key components of an EMR cluster

Connect to an EMR cluster with SSH

Create a table using Apache Hive and load batch data from Amazon S3

Run queries using HiveQL

AWS SERVICES NOT USED IN THIS LAB

AWS service capabilities used in this lab are limited to what the lab requires. Expect errors when accessing other services or performing actions beyond those provided in this lab guide.

ICON KEY

Various icons are used throughout this lab to call attention to certain aspects of the guide. The following list explains the purpose for each one:

Specifies the command you must run.

Verify the output of a command or edited file.

Specifies important hints, tips, guidance, or advice.

Calls attention to information of special interest or importance. Failure to read the note does not result in physical harm to the equipment or data, but it could result in the need to repeat certain steps.

Specifies where to find more information.

Start lab

To launch the lab, at the top of the page, choose Start lab.

Caution: You must wait for the provisioned AWS services to be ready before you can continue.

To open the lab, choose Open Console.

You are automatically signed in to the AWS Management Console in a new web browser tab.

WARNING: Do not change the Region unless instructed.

COMMON SIGN-IN ERRORS

Error: You must first sign out

If you see the message, **You must first log out before logging into a different AWS account:**

Choose the **click here** link.

Close your **Amazon Web Services Sign In** web browser tab and return to your initial lab page.

Choose Open Console again.

Error: Choosing Start Lab has no effect

In some cases, certain pop-up or script blocker web browser extensions might prevent the **Start Lab** button from working as intended. If you experience an issue starting the lab:

Add the lab domain name to your pop-up or script blocker's allow list or turn it off.

Refresh the page and try again.

Task 1: Explore the lab environment

In this task, you review the account resources created when the lab was started.

SAMPLE DATA

In the Amazon Simple Storage Service (Amazon S3) bucket with **databucket** in its name, there is a **data/** folder that has a **stock_prices.csv** file. This file contains the information of stock prices of some of the big tech companies (AAPL, SQ, AMZN, GE, M, TSLA, and MSFT) for the year 2020. Data columns you can find include **Trade_Date**, **Ticker**, **High**, **Low**, **Open**, **Close**, **Volume**, and **Adj_Close**.

Sample Data

Trade_Date	Ticker	High	Low	Open	Close	Volume	Adj_Close
2020-01-02	aapl	75.1500015258789	73.79750061035156	74.05999755859375	75.0875015258789	135480400.0	74.20746612548828
2020-01-02	sq	64.05000305175781	62.95000076293945	62.9900016784668	63.8300018310546 9	5264700	63.8300018310546 9
2020-01-02	amzn	1898.010009765625	1864.1500244140625	1875.0	1898.010009765625	4029000	1898.010009765625
2020-01-02	ge	11.960000038146973	11.229999542236328	11.229999542236328	11.930000305175781	87421800.0	11.861019134521484
2020-01-02	m	17.270000457763672	16.389999389648438	17.180000305175781	16.520000457763672	26388100.0	15.861986160278 32
2020-01-02	tsla	86.13999938964844	84.34200286865234	84.9000015258789	86.0520019531250 0	47660500.0	86.0520019531250 0
2020-01-02	msft	160.72999572753906	158.33000183105469	158.77999877929688	160.61999511718750	22622100.0	158.20576477050781

Using this file in the Amazon S3 bucket as your data source, you will import the data to Amazon EMR and perform analysis using HiveQL.

REVIEW YOUR EMR CLUSTER CONFIGURATION

At the top of the page, in the unified search bar, search for and choose

EMR

In the left navigation pane, in the **EMR on EC2** section, choose **Clusters**.

Select **lab cluster** to view more details.

You will be presented with the **Summary** page of the EMR cluster. Use this tab to view the basics of your cluster configuration.

CHALLENGE A

Choose each tab to review the EMR cluster details.

Can you answer the following questions based on your review?

What is the master node public **DNS address**?

What are the number and type of **Core** instances?

What is the **release** of Amazon EMR used to create the cluster?

What are the open-source **applications** Amazon EMR installed when the cluster was created?

In this instance, you will notice that the cluster is preloaded with the Hive application.

Task 2: Connect to the EMR leader node using Session Manager

In this task, you use Session Manager, a capability of AWS Systems Manager, to connect to your EMR leader node.

On the left side of this lab instruction page, copy the **CommandHostSessionManagementUrl** value.

Open a new tab in your browser, paste the value in, and press ENTER.

This will open a Session Manager terminal.

To connect to your EMR leader node, paste the following commands into the Session Manager terminal:

```
# Get EMR cluster ID and export to the Environment.
export ID=$(aws emr list-clusters | jq '.Clusters[0].Id' | tr -d "")

# Use the ID to get the PublicDNS name of the EMR cluster
# and export to the Environment.
export HOST=$(aws emr describe-cluster --cluster-id $ID | jq '.Cluster.MasterPublicDnsName' | tr -d "")
```

```
# SSH to the EMR cluster
ssh -i ~/EMRKey.pem hadoop@$HOST
```

When prompted to allow a first connection to this remote server, type

`yes` and press ENTER.

This command accesses a predefined key pair for authentication, so you are not prompted for a password.

You are connected to your EMR leader EC2 instance. A message similar to the following indicates a successful connection to the leader node:

```
__|  __|_  )
      _|  (     /   Amazon Linux AMI
     ___|\___|___|

EEEEEEEEEEEEEEEEEEEE    MMMMMMMMM                                MMMMMMMMM
RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

Task 3: Access your Amazon S3 data using Amazon S3 Select with Hive

In this task, you start an interactive Hive session with the leader node. Then, you create the Hive table using Amazon S3 Select.

For Amazon EMR, the computational work of filtering large datasets for processing is pushed down from the cluster to Amazon S3, which can improve performance in some applications. With Amazon S3 Select, you can use simple Structured Query Language (SQL) statements to filter the contents of an Amazon S3 object and retrieve just the subset of data that you need.

To create a logging directory that will be used by Hive, paste the following commands into the SSH window:

```
sudo chown hadoop -R /var/log/hive
```

```
mkdir /var/log/hive/user/hadoop
```

The **hive.log** file is stored in this directory, which contains logs related to Hive.

To connect to the Hive command line interface (CLI), paste the following command into the SSH window:

```
hive
```

You should be presented with a **hive>** prompt. It might take about 10 seconds to appear.

To create a table you can use with Amazon S3 Select, paste the following Hive statement in a text editor:

Replace <*dataBucket*> with the **dataBucket** value shown to the left of these instructions.

```
CREATE TABLE stockprice  (
`Trade_Date` string,
`Ticker` string,
`High` double,
`Low` double,
`Open` double,
`Close` double,
`Volume` double,
`Adj_Close` double
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS
INPUTFORMAT
  'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
```

```
LOCATION 's3://<dataBucket>/data/'
TBLPROPERTIES (
  "s3select.format" = "csv",
  "s3select.headerInfo" = "ignore",
  "skip.header.line.count"="1"
);
```

Take a moment to read through the query. A new table is created with specified headers and an input and output format are defined to help with processing steps later. In the Hive statement, pay close attention to TBLPROPERTIES. In this query, you are using a CSV file and ignoring the first line containing the column headers.

From your text editor, to create the `stockprice` table from the `stock_price.csv` file stored in Amazon S3, paste the updated Hive statement into the SSH window.

The output should display the following message to confirm that the command ran successfully:

```
OK
Time taken: 11.03 seconds
```

To validate data in the `stockprice` table, paste the following Hive statement.

By default, Amazon S3 Select is disabled when you run queries. You can enable Amazon S3 Select by setting `s3select.filter` to `true` in your Hive session, as we've done here.

```
SET s3select.filter=true;
SELECT * FROM stockprice WHERE `Trade_Date` LIKE '2020-01-03' ORDER BY `Ticker`;
```

The output should display seven rows of data. A sample row is shown below:

```
2020-01-03 amzn 1886.199951171875 1864.5 1864.5 1874.969970703125 3764400.0
```

CHALLENGE B

Can you list the top 10 records by volume with dates using the `stockprice` table?

Navigate **here** for a solution.

Task 4: Challenge – ingest and query movie data

We have uploaded movie data to the **challengeBucket**. Your task is to create a

movies table and find the number of movies that actor

Tom Hanl is associated with as an actor.

Navigate **here** for a solution.

Hint: The column names are:

year

title

directors_0

rating

genres_0

genres_1

rank

running_time_secs

actors_0

actors_1

actors_2

directors_1

directors_2

Conclusion

Congratulations! You now have successfully:

Reviewed how Amazon EMR and Apache Hive can be used together to ingest and query data

Identified key components of an EMR cluster

Connected to an EMR cluster with SSH

Created a table using Apache Hive and loaded batch data from Amazon S3

Run queries using HiveQL

End lab

Follow these steps to close the console and end your lab.

Return to the **AWS Management Console**.

At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.

Choose End lab and then confirm that you want to end your lab.

For more information about AWS Training and Certification, see *https://aws.amazon.com/training/*.

Your feedback is welcome and appreciated.
If you would like to share any feedback, suggestions, or corrections, please provide the details in our *AWS Training and Certification Contact Form*.

Appendix

CHALLENGE B SOLUTION

```
SELECT `Trade_Date`, `Ticker`, `Volume` FROM stockprice ORDER BY `Volume` DESC LIMIT 10;
```

To continue this lab, move on to Task 4.

TASK 4 CHALLENGE SOLUTION

Replace <*challengeBucket*> with the **challengeBucket** value shown to the left of these instructions

```
CREATE TABLE movies  (
`year` int,
`title` string,
`directors_0` string,
`rating` string,
`genres_0` string,
`genres_1` string,
`rank` string,
`running_time_secs` string,
`actors_0` string,
`actors_1` string,
`actors_2` string,
`directors_1` string,
`directors_2` string
```

```
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS
INPUTFORMAT
  'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://<challengeBucket>/data/'
TBLPROPERTIES (
  "s3select.format" = "csv",
  "s3select.headerInfo" = "ignore",
  "skip.header.line.count"="1"
);
```

The number of movies that actor

Tom Hanl is associated with:

```
SELECT COUNT(title) FROM movies WHERE actors_0='Tom Hanks' OR actors_1='Tom Hanks' OR actors_2='Tom Hanks';
```