

Ryan Britton & Ashan Deen

ECE 4150

Professor Vijay Madisetti

Monday, April 15th

Project Report: Apache Spark

1. Introduction to Apache Spark

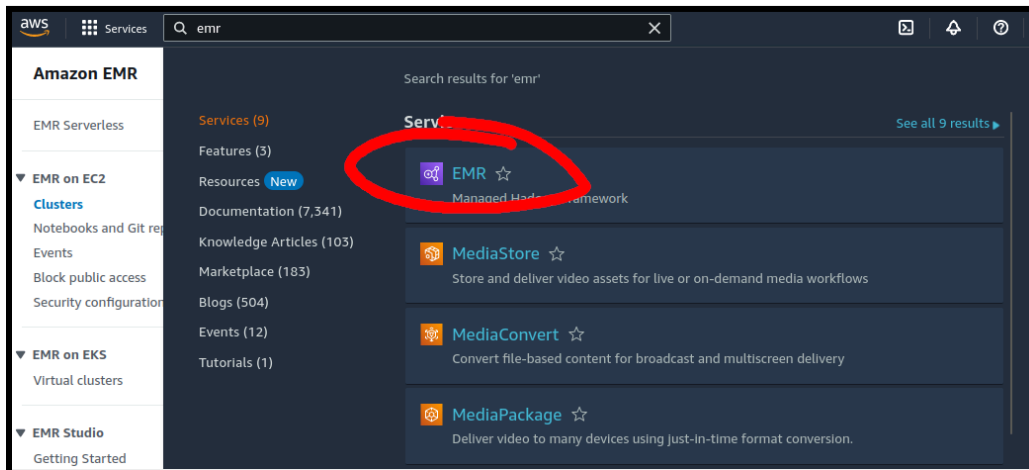


Apache Spark is a framework developed by contributors to the Apache Foundation designed for processing data at great scales. Some example use cases for Apache Spark could be as simple as dataset processing to complex as machine learning pipelines. Other notable use cases are for analytics and forecasting, such as for financial models, as well as image processing. Spark also supports the streaming of real time data, with integrations for Apache Kafka, TCP sockets, and more.

In this project, we will be performing data processing with Apache Spark by reading from a local file in our EMR cluster. Coincidentally, we will be counting the number of words in the book "The Adventures of Sherlock Holmes" by Arthur Conan Doyle, as well as counting the occurrences of letters of the alphabet.

2. Setting Up Apache Spark

Beginning in the AWS console, navigate to Amazon EMR.



Next you should create a new cluster. Be sure that "Spark Interactive" is selected for the Application Bundle.

▼ Name and applications - required [Info](#)

Name your cluster and choose the applications that you want to install to your cluster.

Name

Amazon EMR release [Info](#)

A release contains a set of applications which can be installed on your cluster.

emr-7.0.0

Application bundle

Spark
Interactive

Core
Hadoop

Flink

HBase

Presto

Trino

Custom

☐ AmazonCloudWatchAgent
1.300031.1

☐ HCatalog 3.1.3

☐ Hue 4.11.0

☒ Livy 0.7.1

☐ Phoenix 5.1.3

☒ Spark 3.5.0

☐ Tez 0.10.2

☐ ZooKeeper 3.5.10

☐ Flink 1.18.0

☒ Hadoop 3.3.6

☒ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☐ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 426

☐ HBase 2.4.17

☒ Hive 3.1.3

☐ JupyterHub 1.5.0

☐ Oozie 5.2.1

☐ Presto 0.283

☐ TensorFlow 2.11.0

☐ Zeppelin 0.10.1

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

☐ Use for Hive table metadata

☐ Use for Spark table metadata

Operating system options [Info](#)

☒ Amazon Linux release

☐ Custom Amazon Machine Image (AMI)

☒ Automatically apply latest Amazon Linux updates

For the remaining options, you may delete "Task 1 of 1", and assign or create a new VPC. Be sure the subnet is public.

Task 1 of 1

Name
Task - 1

Choose EC2 instance type
m5.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: \$0.192 per instance/hour
Lowest Spot price: \$0.078 (us-east-1f)

Actions ▾

► Node configuration - optional

Add task instance group

You can add up to 47 more task instance groups.

EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#)

Size (GiB)	IOPS	Throughput (MiB/s)
15	3000	125

15 - 100 GiB per volume
General Purpose SSD (gp3)

3000 - 16000 IOPS per volume.
Choose a maximum ratio of 500:1 between IOPS and volume size.

125 - 1000 MiB/s per volume.
Choose a maximum ratio of 0.25:1 between throughput and IOPS.

► Cluster scaling and provisioning - required [Info](#)
Choose how Amazon EMR should size your cluster.

▼ Networking - required [Info](#)
Choose the network settings that determine how you and other entities communicate with your cluster.

Virtual private cloud (VPC) [Info](#)
vpc-0ef31b5329cd673e6 [Browse](#) [Create VPC](#)

Subnet [Info](#)
subnet-00f4568b72a73dd9a [Browse](#) [Create subnet](#)

Next you must create or reuse an SSH key to access the cluster. Download this key to your local machine.

▼ **Security configuration and EC2 key pair** [Info](#)

Choose a security configuration or create a new one that you can reuse with other clusters.

Security configuration

Select your cluster encryption, authentication, authorization, and instance metadata service settings.

Q Choose a security configuration

↺

Browse ↗

Create security configuration ↗

Amazon EC2 key pair for SSH to the cluster [Info](#)

Q ECE4150

×

Browse

Create key pair ↗

Finally, select "Create a service role" and "Create an instance profile"

▼ Identity and Access Management (IAM) roles - *required* [Info](#)

Choose or create a service role and instance profile for the EC2 instances in your cluster.

Amazon EMR service role [Info](#)

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

☐ Choose an existing service role

Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

☒ Create a service role

Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Networking resources

We've already added the resources that you configured in the [Networking](#) section. Choose the VPC, subnet, and security groups that the service role can access.

Virtual Private Cloud (VPC)

Choose one or more VPCs ▼

Subnet

Choose one or more subnets ▼

Security group

Choose one or more security groups ▼

EC2 instance profile for Amazon EMR

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

☐ Choose an existing instance profile

Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

☒ Create an instance profile

Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

S3 bucket access [Info](#)

☒ Specific S3 buckets or prefixes in your account [Info](#)

Choose the buckets or prefixes that you want this instance profile to access.

☐ All S3 buckets in this account with read and write access

Grant the instance profile access to all buckets that have read and write access enabled in your account.

S3 buckets

We've already added the resources that you configured in the [Cluster logs](#) section. Choose the S3 buckets and bucket prefixes where you store logs and data for your cluster, bootstrap actions, and steps.

S3 URI

View [↗](#)

Browse S3

Add

S3 bucket	Prefix	Permission	
aws-logs-59018410... Inherited from Cluster logs	elasticmapreduce	Read and write	<div>Edit</div>

Custom automatic scaling role - *optional*

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. [Learn more](#) [↗](#)

Custom automatic scaling role

Choose IAM role ▼

↻

Create IAM role [↗](#)

6

Now you can create the cluster. Ensure the options here match.

Summary [Info](#)

Name and applications - required

Name
FinalProject

Amazon EMR release
emr-7.0.0

Application bundle
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3,
JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Spark 3.5.
...)

Cluster configuration - required

Uniform instance groups
Primary (m5.xlarge), Core (m5.xlarge)

Cluster scaling and provisioning - required

Provisioning configuration
Core size: 1 Instance

Cancel **Create cluster**

Once the EMR cluster is created, you must allow access to the from your IP address. Locate the "Core and task nodes - EMR managed security group" on the EMR cluster summary page.

Network and security [Info](#)

Network Virtual Private Cloud (VPC) vpc-0ef31b5329cd673e6 🔗 Subnet(s) and Availability Zone(s) (AZ) subnet-00f4568b72a73dd9a 🔗 us-east-1b ▼ EC2 security groups (firewall) Primary node EMR managed security group sg-0a641360ae7a17eb1 🔗 Additional security groups - Core and task nodes EMR managed security group sg-0ccfff653ded2677c 🔗 Additional security groups - Service access (private subnet) sg-04e8e7d82fcae3425 🔗	Security configuration Security configuration None EC2 key pair ECE4150	Permissions Service role for Amazon EMR AmazonEMR-ServiceRole-20240412T195005 🔗 EC2 Instance profile AmazonEMR-InstanceProfile-20240412T194949 Custom automatic scaling role Not configured
--	--	--

For this security group, edit the inbound rules to create a new rule that allows SSH access from any IP.

Edit inbound rules [Info](#)

Inbound rules control the incoming traffic that's allowed to reach the instance.

Inbound rules [Info](#)

Security group rule ID	Type Info	Protocol Info	Port range Info	Source Info	Description - optional Info	
sgr-05b87cef9f0aa78b7	All ICMP - IPv4	ICMP	All	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0ccff653ded2677c"/> <input type="button" value="X"/>		
sgr-082682f9f55a18d21	All TCP	TCP	0 - 65535	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0a641360ae7a17eb1"/> <input type="button" value="X"/>		
sgr-008ce88e105f7ab34	All UDP	UDP	0 - 65535	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0ccff653ded2677c"/> <input type="button" value="X"/>		
sgr-0a956652529870e2b	All TCP	TCP	0 - 65535	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0ccff653ded2677c"/> <input type="button" value="X"/>		
sgr-0dff507af71ab942c	All ICMP - IPv4	ICMP	All	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0a641360ae7a17eb1"/> <input type="button" value="X"/>		
sgr-0e9f37c0b011c5f1e	All UDP	UDP	0 - 65535	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-0a641360ae7a17eb1"/> <input type="button" value="X"/>		
sgr-02cd38c583fb8be3	Custom TCP	TCP	8443	Cus... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="sg-04e8e7d82fcae3425"/> <input type="button" value="X"/>		
-	SSH	TCP	22	An... <input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>
				<input type="text" value="0.0.0.0/0"/> <input type="button" value="X"/>		

Rules with source of 0.0.0.0/0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

Now that SSH access is permitted, connect to the primary node using the steps listed on the EMR cluster summary page.

FinalProject Updated less than a minute ago Terminate Clone in AWS CLI Clone

▼ **Summary**

Cluster info	Applications	Cluster management	Status and time
Cluster ID J-3B1EOACANP57A Cluster configuration Instance groups Capacity 1 Primary 1 Core 0 Task	Amazon EMR version emr-7.0.0 Installed applications Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Spark 3.5.0	Log destination in Amazon S3 aws-logs-590184107193-us-east-1/ elasticmapreduce Primary node public DNS ec2-3-85-237-244.compute-1.amazo naws.com Connect to the Primary node using SSH Connect to the Primary node using SSM	Status Starting Creation time April 12, 2024, 20:08 (UTC-04:00) Elapsed time 2 minutes, 7 seconds

3. Task 1: Counting Unique Words

For counting the words in the book, we will use a modified version of a sample python script provided by Apache in the Spark source code GitHub page:

(<https://github.com/apache/spark>)

Once you have confirmed you can SSH into the cluster, use scp to copy the "book.txt" and "wordcount.py" from the task1 folder to the cluster.

```

• ((.venv) ) ~/git/ece4150-labs/finalproject $ scp -i ECE4150.pem -r resources/task1 hadoop@ec2-3-85-237-244.compute-1.amazonaws.com:~
  book.txt                                     100% 581KB   2.5MB/s   00:00
  wordcount.py                                100% 1418    47.0KB/s   00:00
○ ((.venv) ) ~/git/ece4150-labs/finalproject $ █

```

Next ssh into the cluster and copy the dataset to hadoop, as that is how Spark in the EMR cluster has been configured to stream data from.

```
[hadoop@ip-10-0-8-250 task1]$ _hadoop fs -copyFromLocal book.txt /user/hadoop/book.txt
```

Finally, submit the job to Apache Spark for processing

```
[hadoop@ip-10-0-8-250 task1]$ spark-submit wordcount.py book.txt
```

If that is successful, you may copy the results.txt file back to your computer for submission for task 1.

```
$ scp -i ECE4150.pem -r hadoop@ec2-3-85-237-244.compute-1.amazonaws.com:~/task1/results.txt resources/task1/
```

For completion, we've included results.txt, which should match the output results.txt.

4. Task 2: Counting Character Occurrences

For the second task, if actually a task, we would then ask the students to modify the wordcount.py script to count the occurrences of characters in the text. For the sake of simplicity and completion, we've provided characteroccurrences.py files in task2/, as well as the expected result. The file may be modified in EMR or modified locally and SCP'd later. We will use the same book.txt from the previous task. The output should be sorted by character, with the format of "Character: '{}', Count: {}" for each line.

```
scp -i ECE4150.pem -r resources/task2/ hadoop@ec2-3-85-237-244.compute-1.amazonaws.com:~
```

For completion, we've included results.txt, which should match the output results.txt.