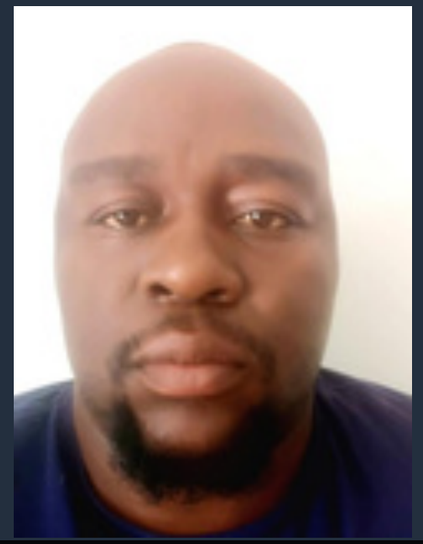
House Rules

- Welcome! Thank you for joining us for **Taming Big Data with Elastic MapReduce (EMR) on AWS.**
- Please feel free to pop your questions in the Q&A and our Engineers will answer them as we go along.
- Our Presenter will also answer some of the questions at different points.
- Please see Github link where you will find some of the information pertaining to our event tonight: https://github.com/patrick-muller/labs
- And please see the document for Launching your EMR Cluster, using the Hash Link which Nikki sent on email to you
- If you did not receive a Hash Link, please check your Junk-Mail or Spam

aws

About AWS Premium Support

AWS Premium Support is one-on-one, fast-response support from experienced technical support engineers. The service helps customers use AWS products and features. The **Big Data Cloud Support Engineering** team is one of our teams within Premium Support. Our Engineers support our customers around the globe, along with becoming Subject Matter Experts in their respective services, provide and receive training from global partners and have the opportunity to contribute to the development of our services.

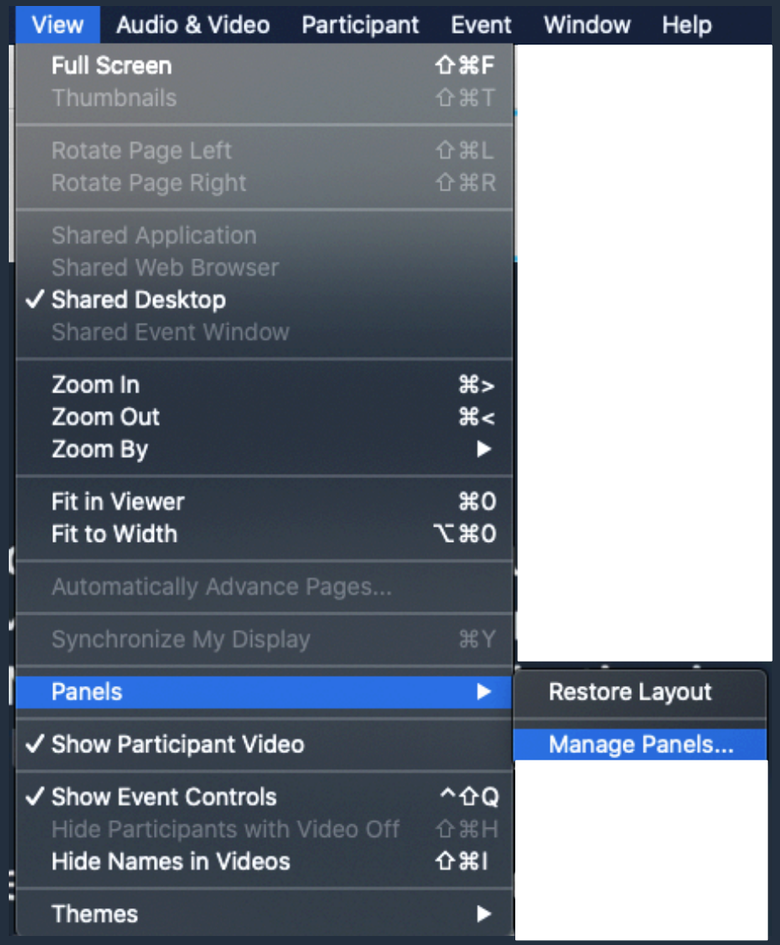


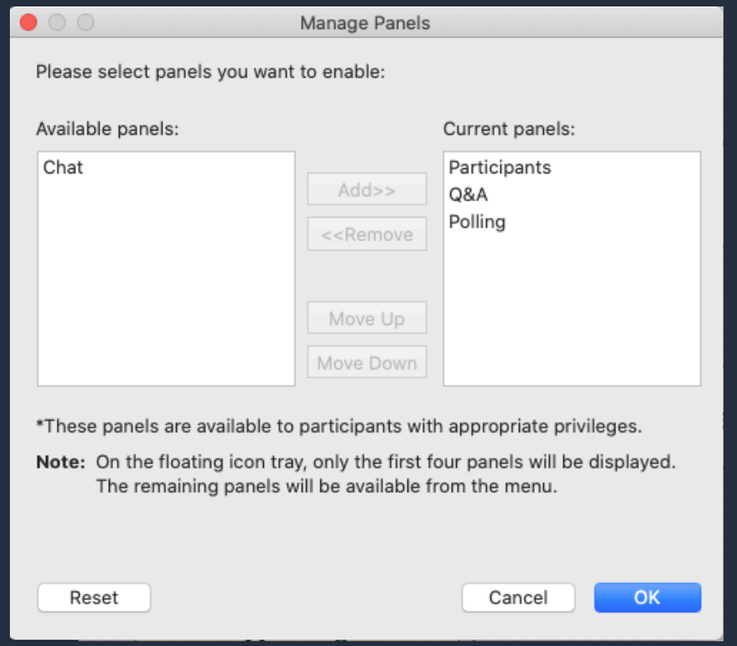
Ops Manager - Brian

The purpose of this event is to give you a chance to learn from our Big Data Cloud Support Engineers, learn more about one of our big data services and stand a chance to participate in our recruitment process, should you be successful in the completion of technical challenge on Day 3.

aws

Cisco Webex

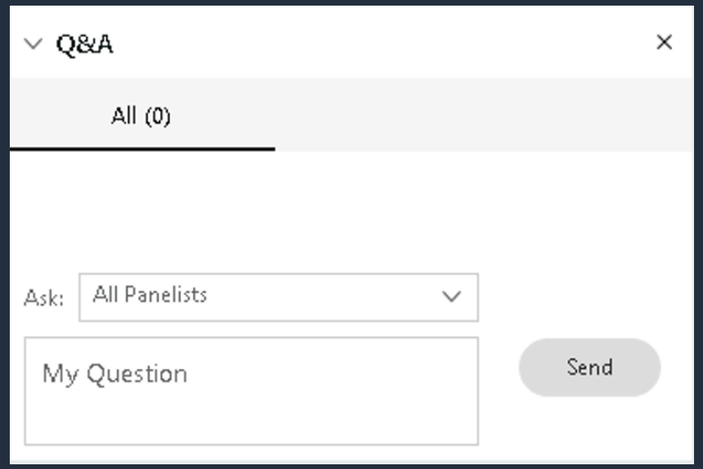




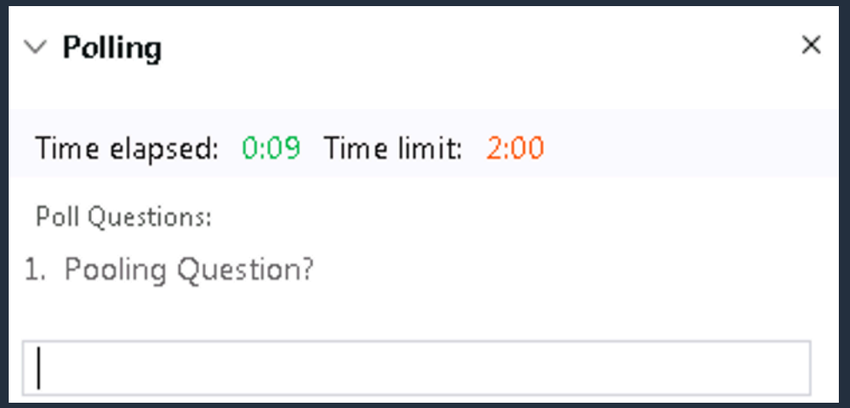
aws

Cisco Webex

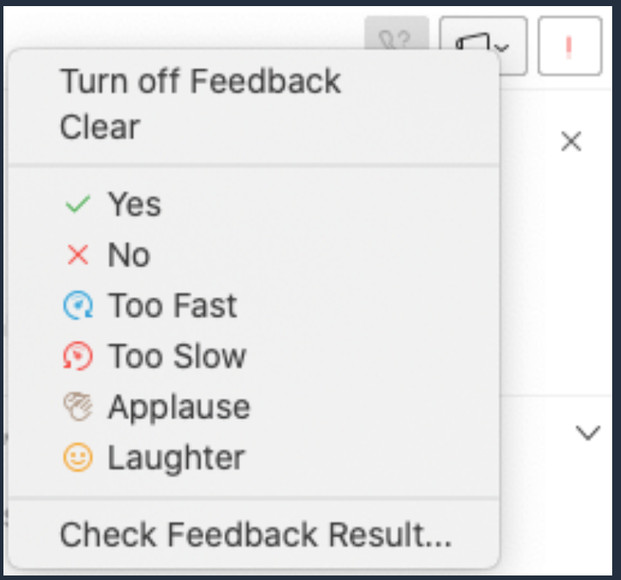
Q&A



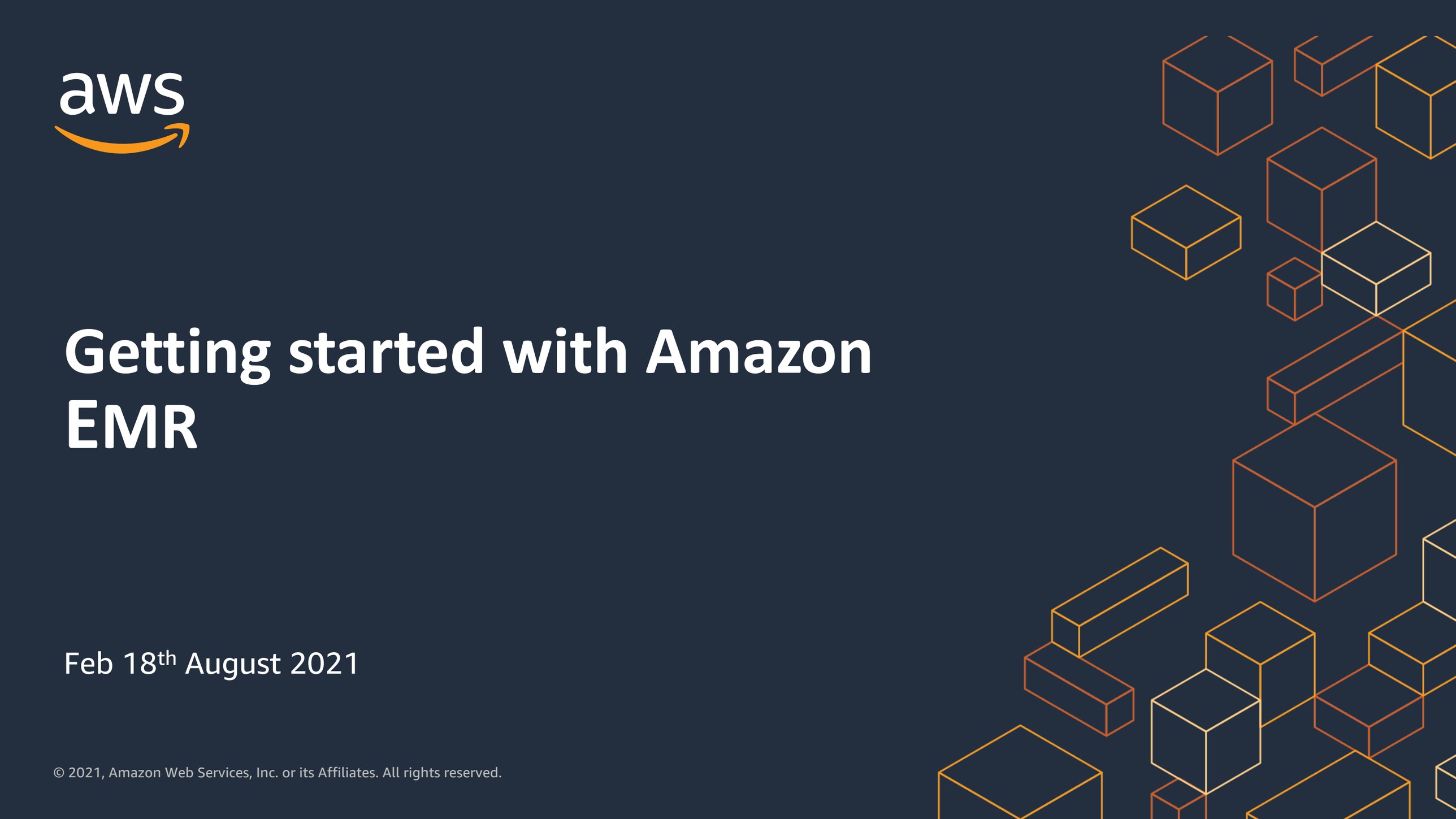
Pooling Questions



Interaction



aws



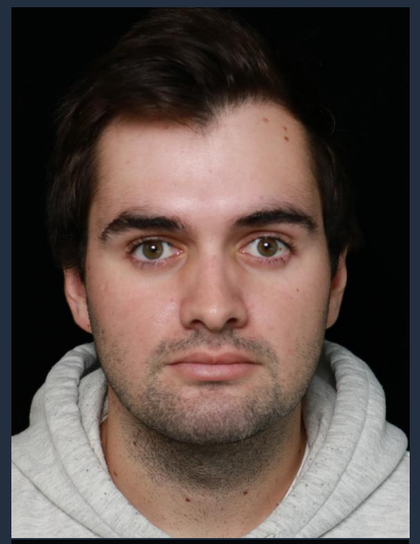
Getting started with Amazon EMR

Feb 18th August 2021

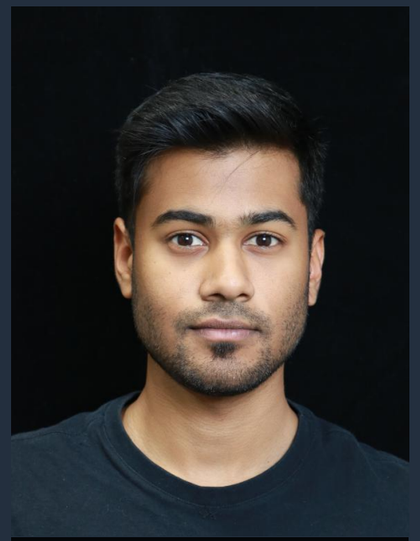
Introduction

- Instructors

Patrick



Peter



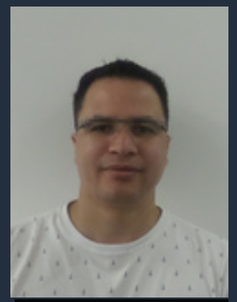
Vidushen

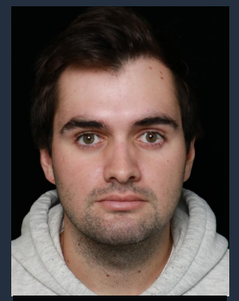
aws

Introduction

Team


Vidushen

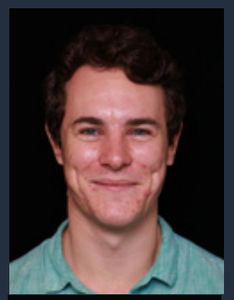

Patrick


Peter


Emmanuel

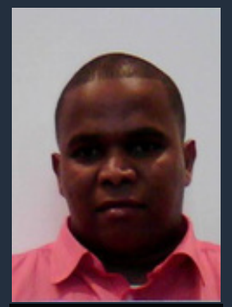

Umesh


Amit


David


Ridick


Andrew


Neil


Riyaad

aws

Agenda

- Cluster Setup
- Introduction to Big Data
- Introduction to Hadoop
- Hadoop Distributed File systems (HDFS)
- YARN Framework
- MapReduce
- Hadoop Ecosystem
- Apache Hadoop on Amazon EMR
- Advantages of Hadoop on Amazon EMR
- Q&A

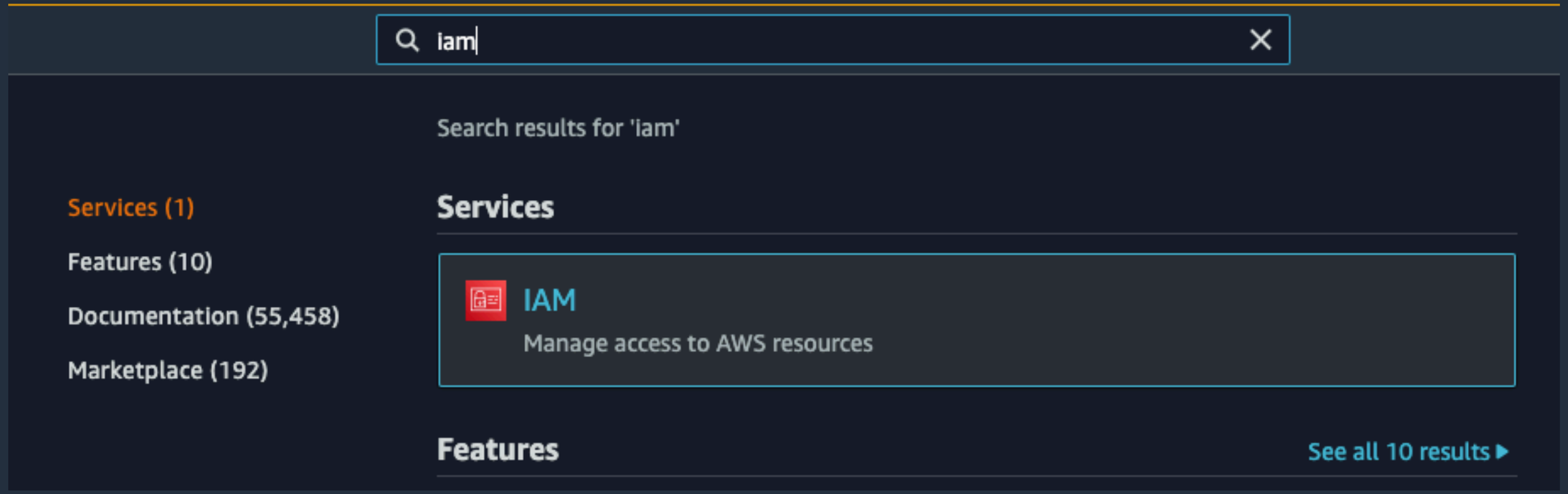
aws

Cluster Setup

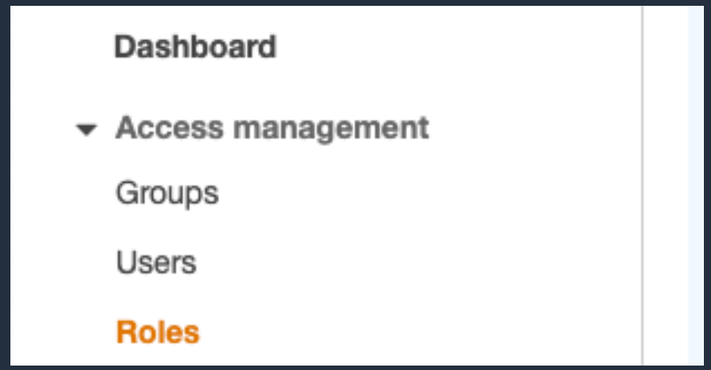
aws

How to launch an EMR cluster

First we need to create a Role, open the IAM console



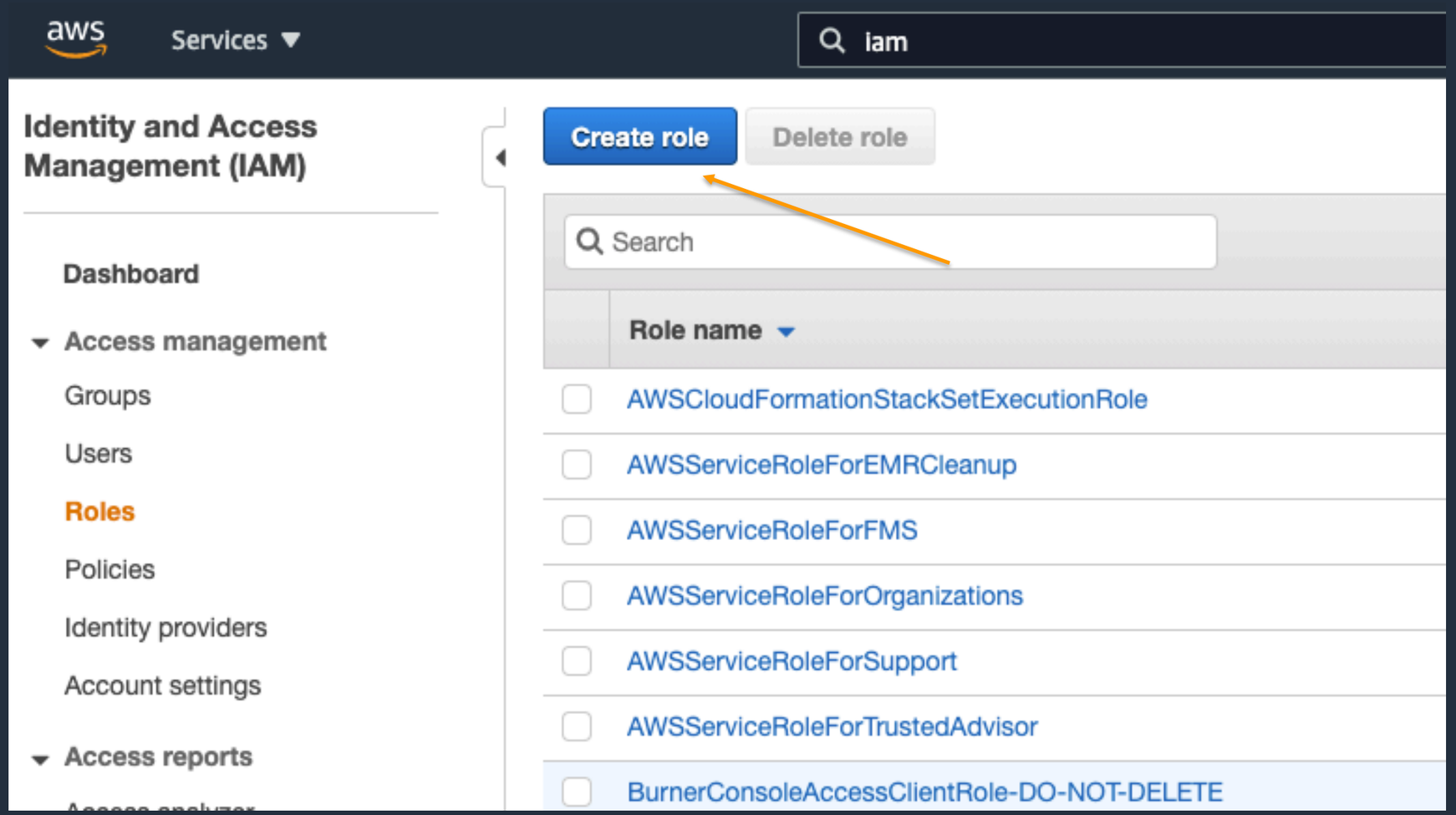
And select the Roles



aws

How to launch an EMR cluster

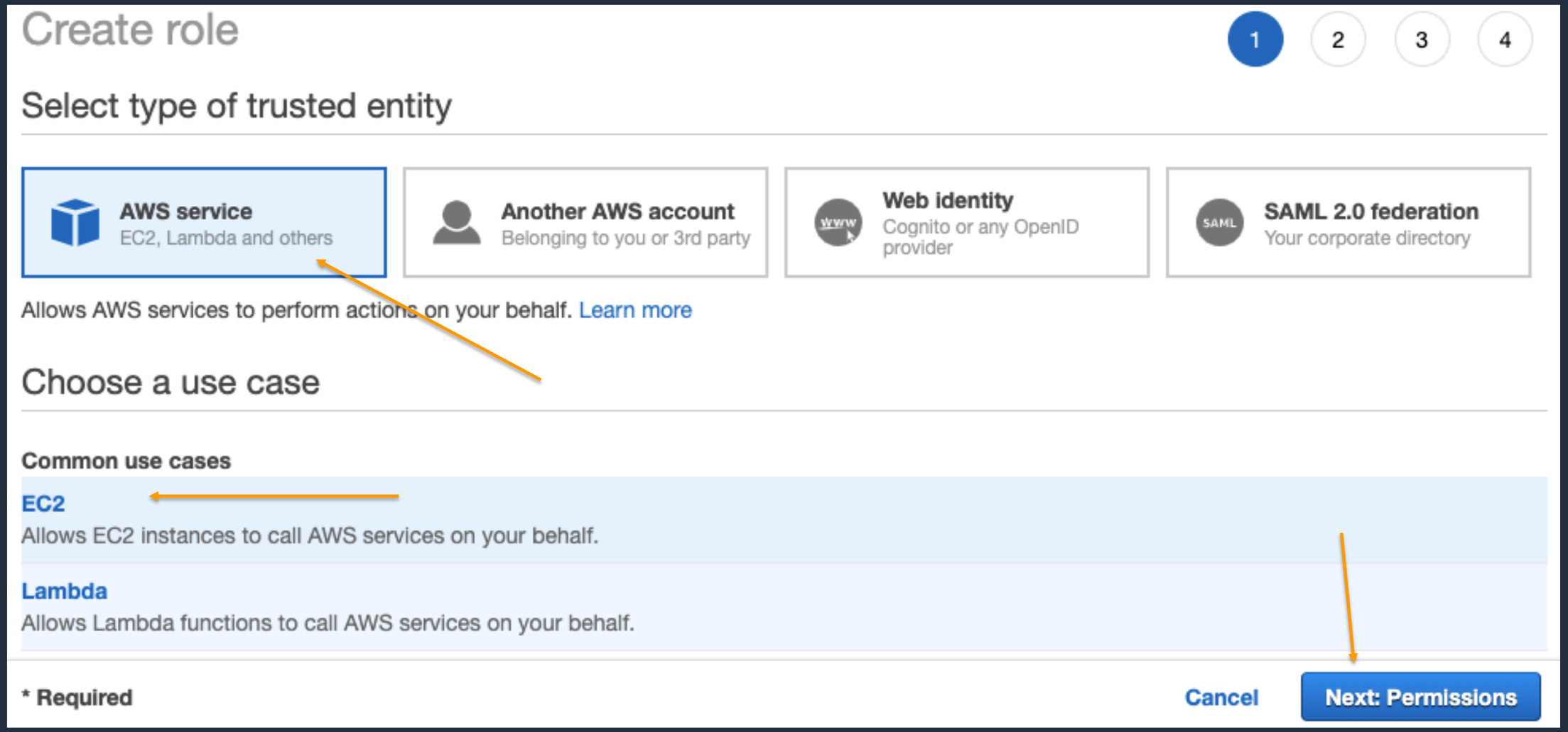
Click in Create role



aws

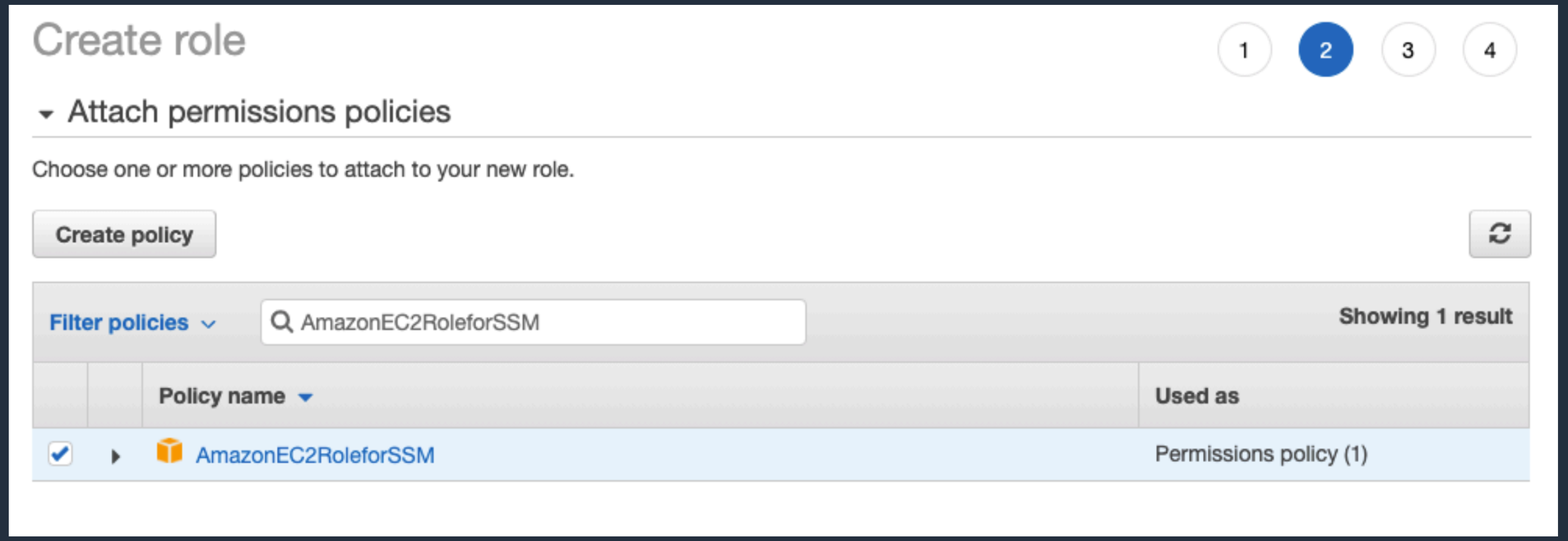
How to launch an EMR cluster

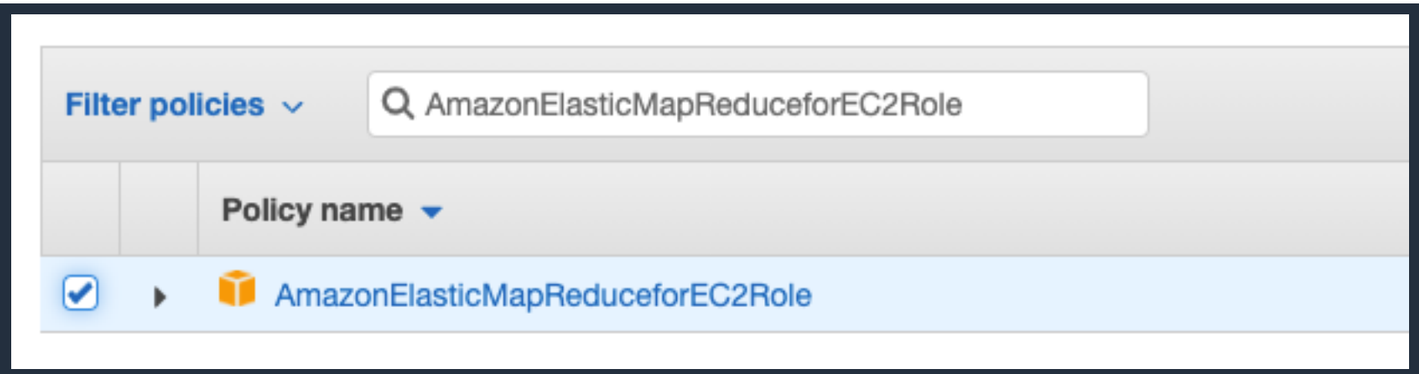
Select AW Service as trusted entity, EC2 as common use cases and click in Next



How to launch an EMR cluster

Search and select AmazonEC2RoleforSSM, after that search and select AmazonElasticMapReduceforEC2Role. Click in Next





aws

How to launch an EMR cluster

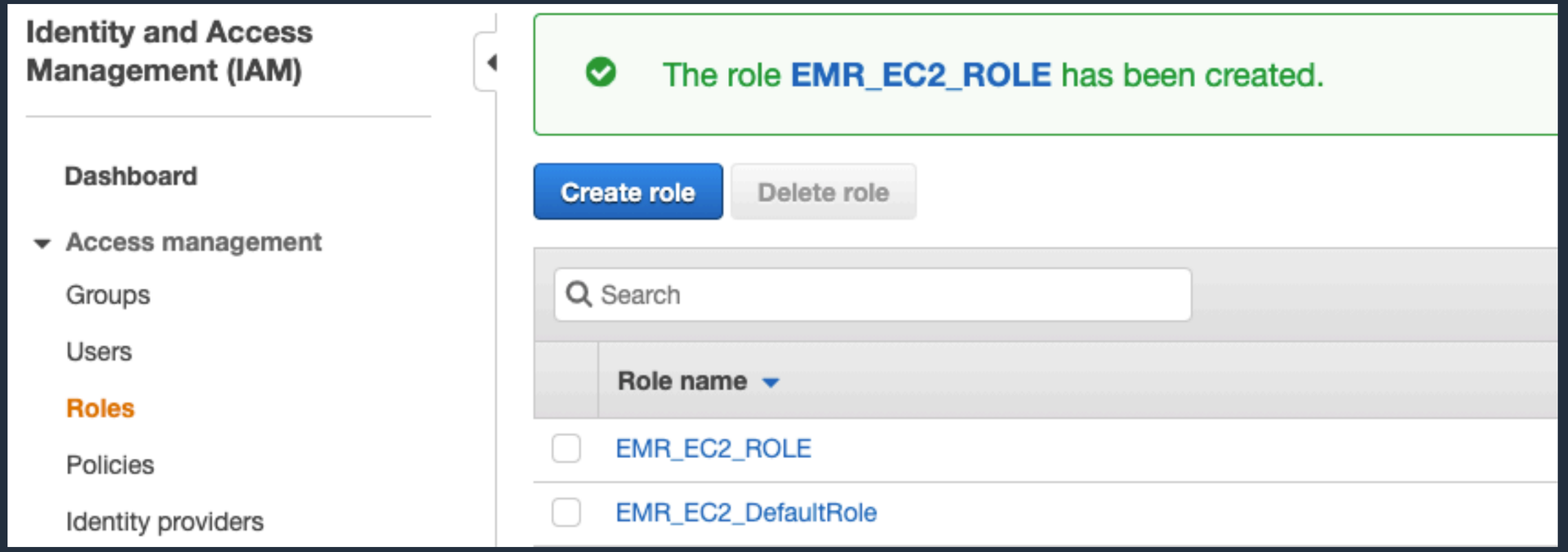
In the Tags session click in Next, In the Review add EMR_EC2_ROLE as Role name.

Check if both policies(AmazonEC2RoleforSSM, AmazonElasticMapReduceforEC2Role) are selected and if Trusted entities is correct, after that click in Create Role

aws

How to launch an EMR cluster

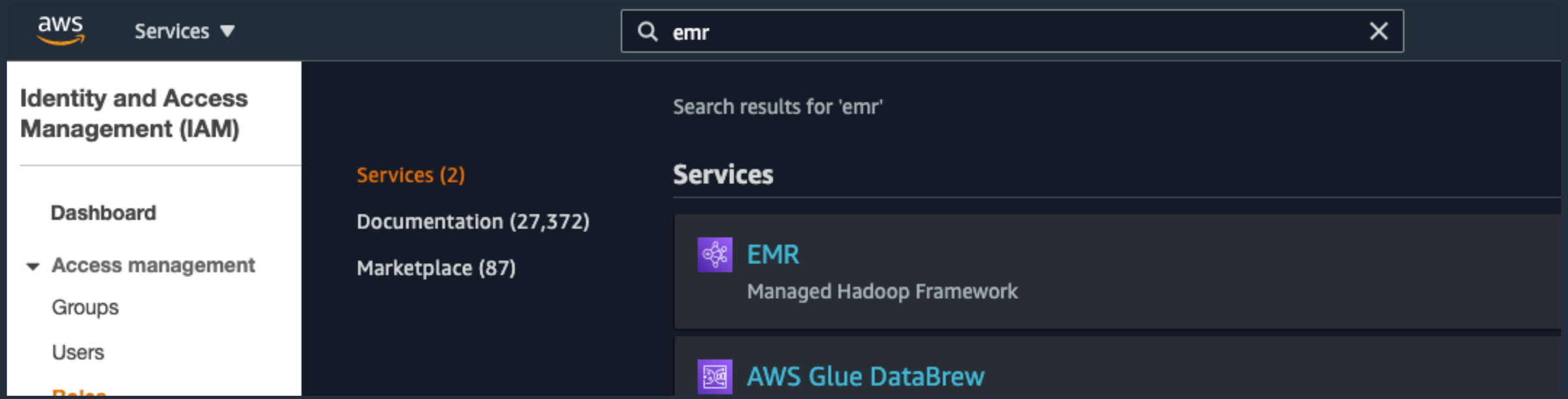
Check if you can see the role EMR_EC2_ROLE in the Role name List



aws

How to launch an EMR cluster

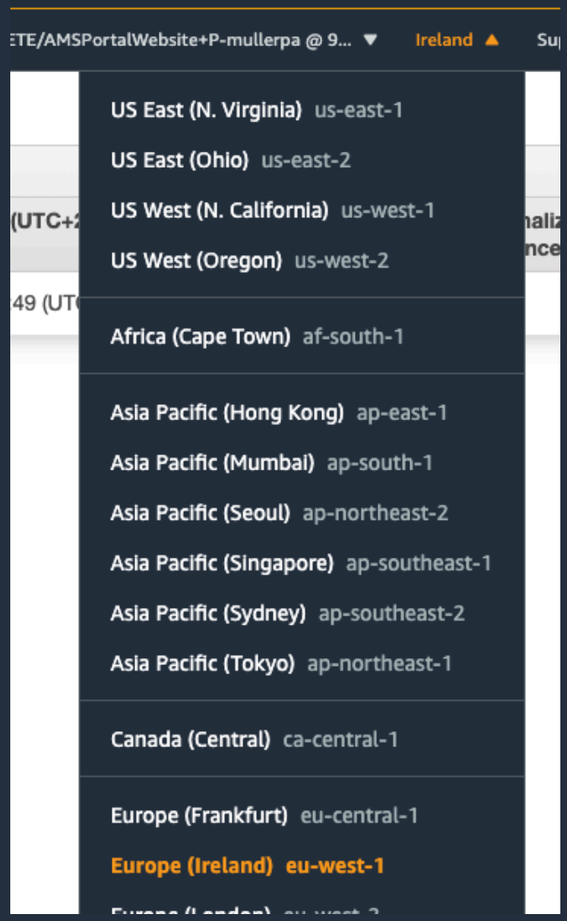
On the AWS console, select search and select EMR from the list of services



aws

How to launch an EMR cluster

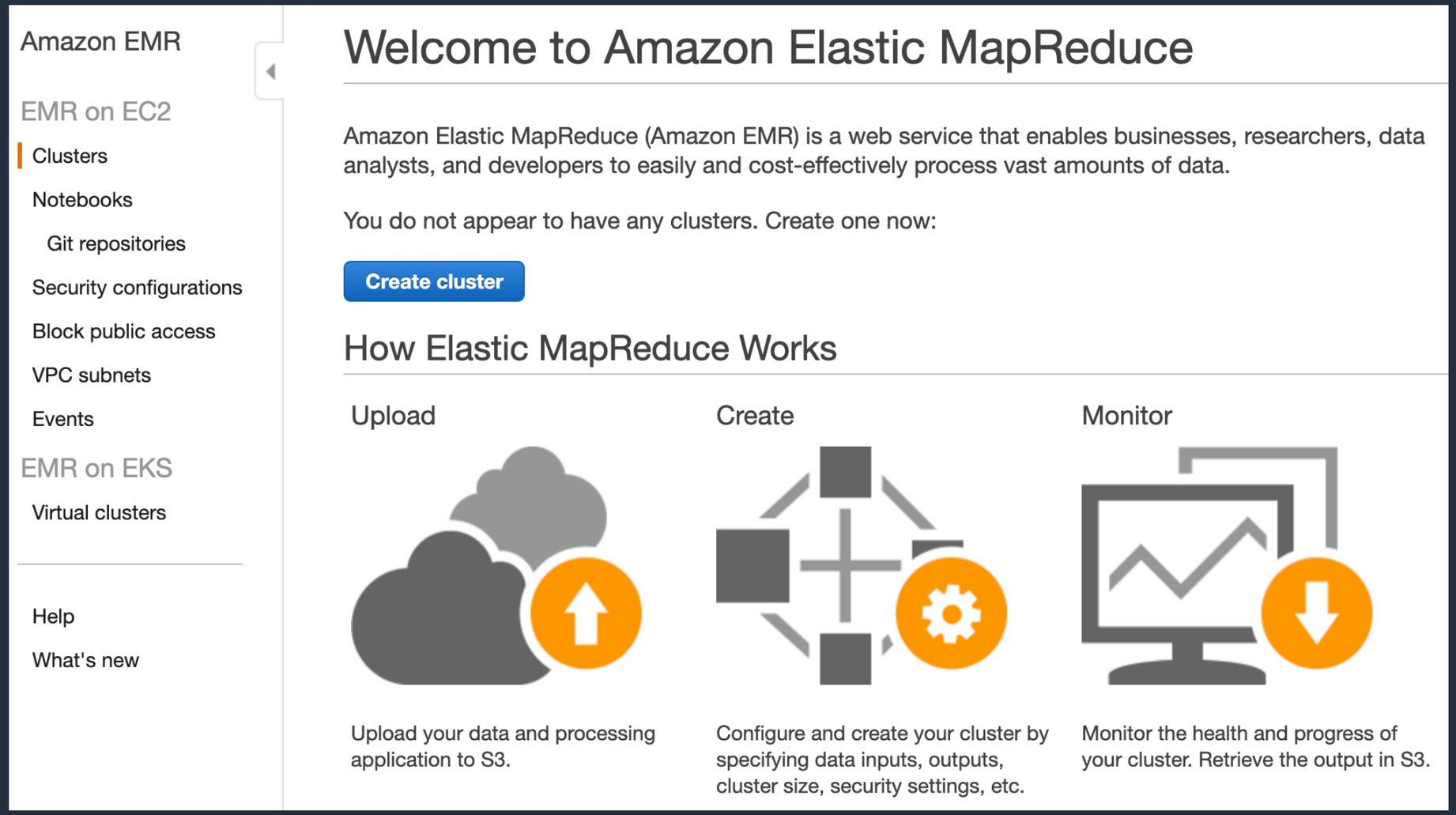
Select the Ireland Region



How to launch an EMR cluster

Click on

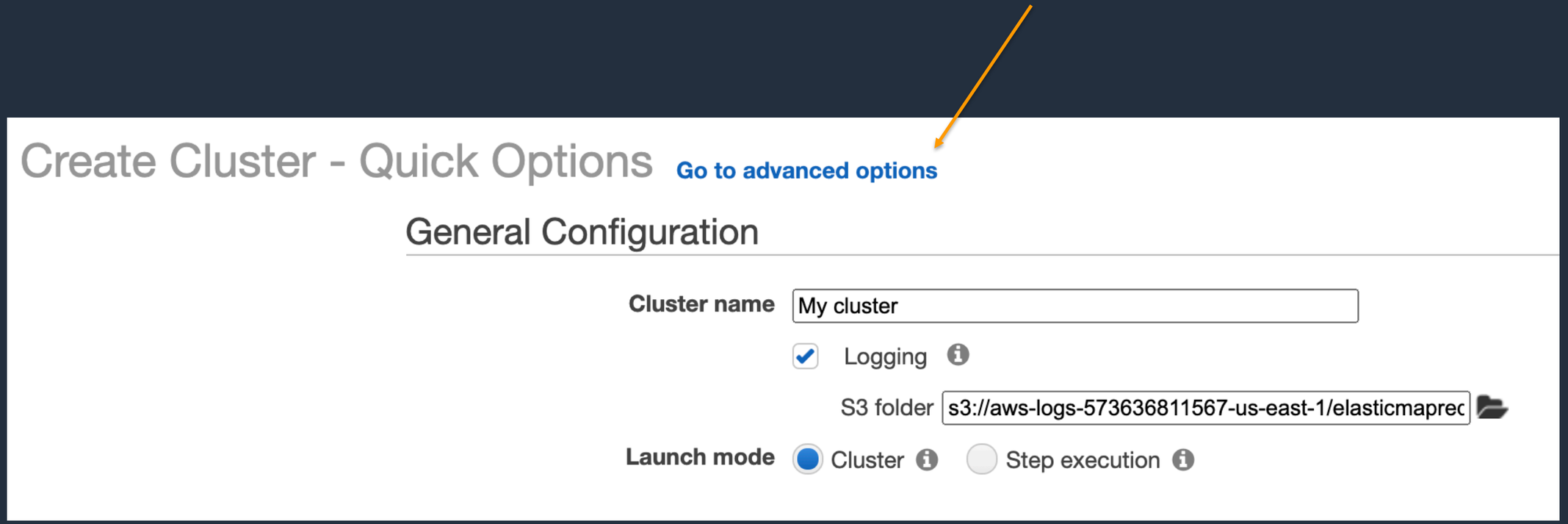
"Create cluster"



aws

How to launch an EMR cluster

On the create cluster page, you'll see an option to "Go to advanced options". Click on that



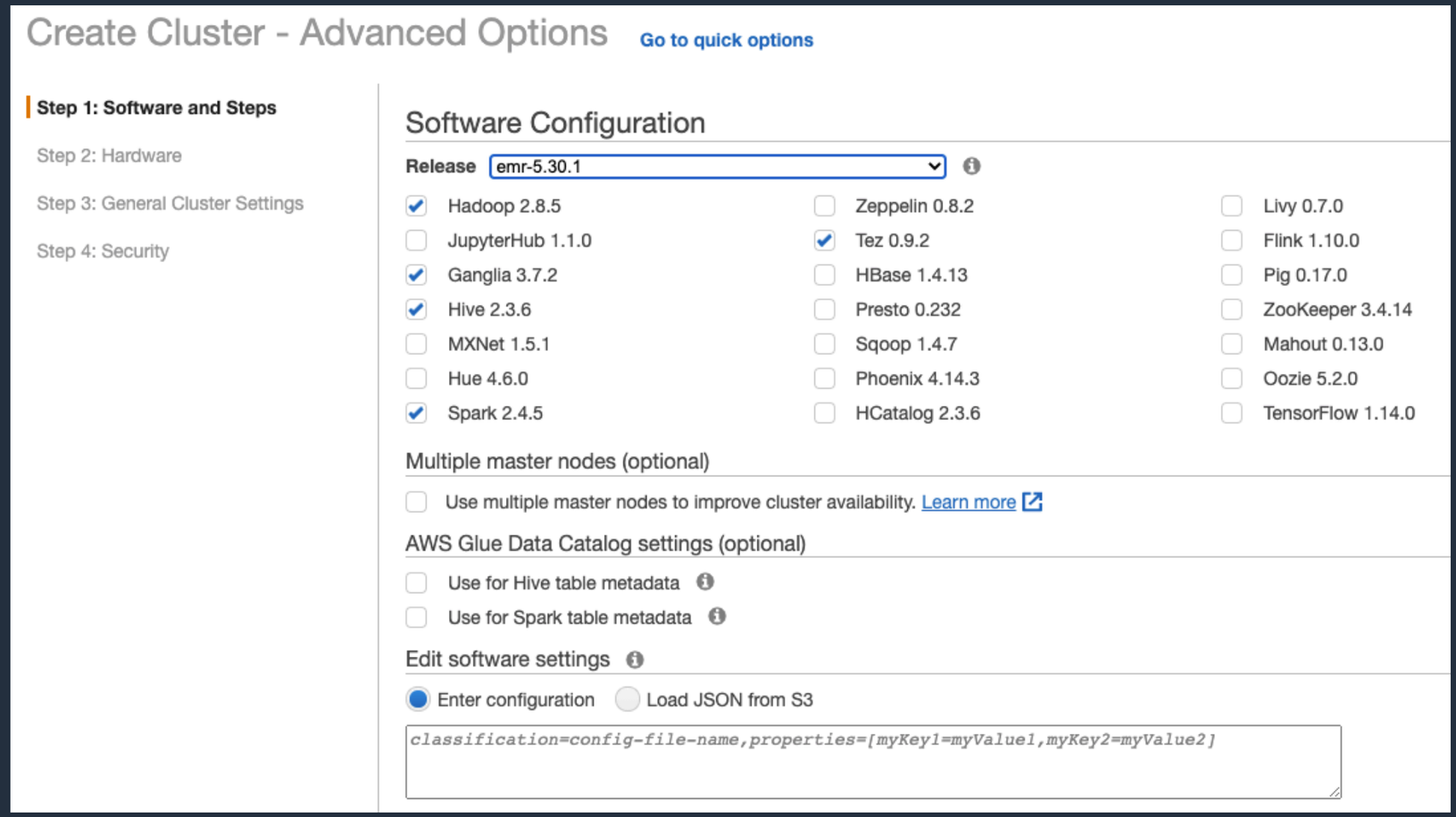
aws

How to launch an EMR cluster

During the Advanced Options, you will be able to select the applications you will be using during this course. Please select EMR release 5.32.0 and applications:

- Hadoop
- Ganglia
- Hive
- Spark
- Tez

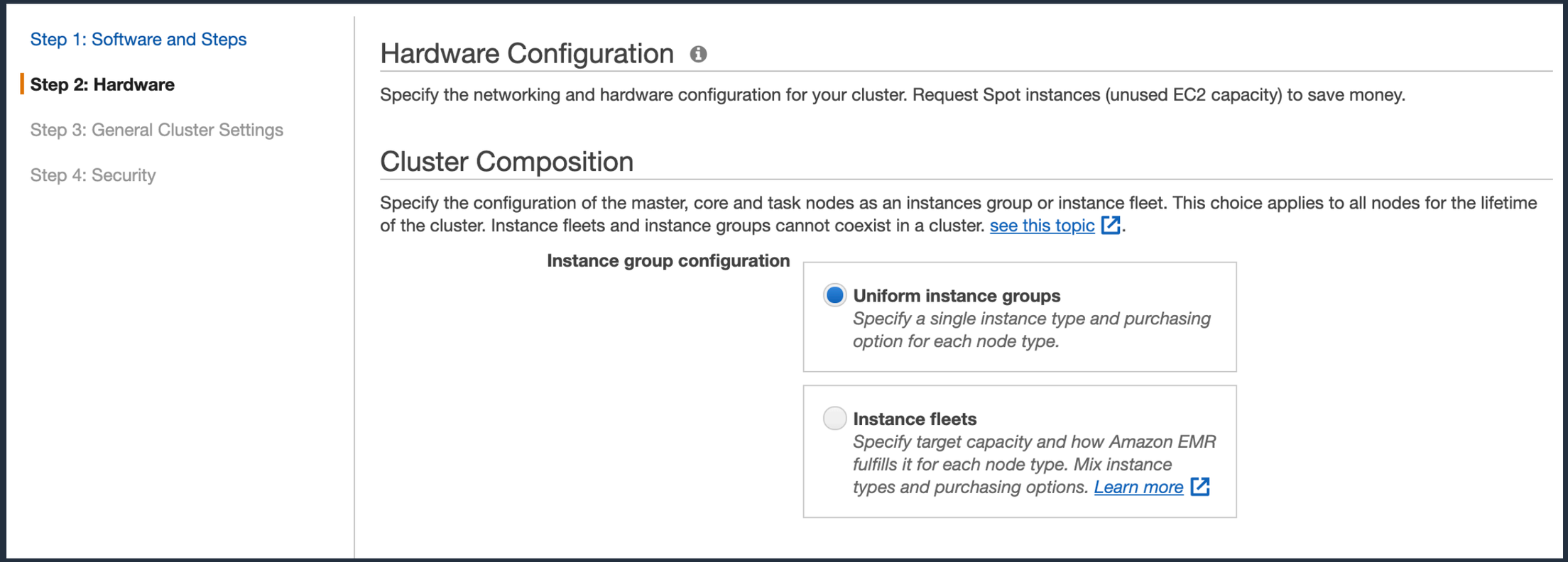
Then click “Next”



aws

How to launch an EMR cluster

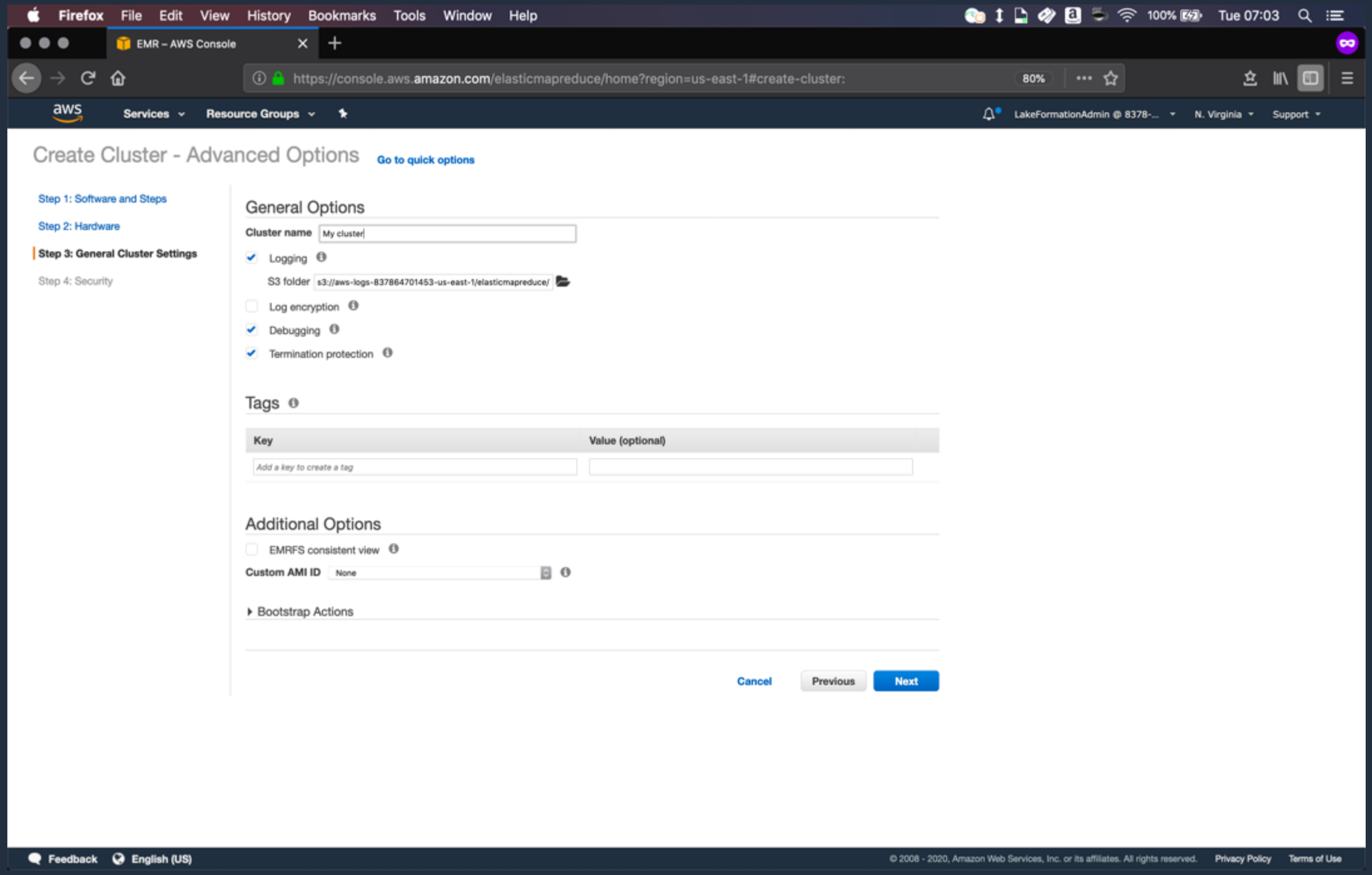
In the Hardware Configuration just click in Next



aws

How to launch an EMR cluster

On the “General Cluster
Settings” page, you
can name your cluster
and leave the rest of
The config as it is.
Click “Next”.

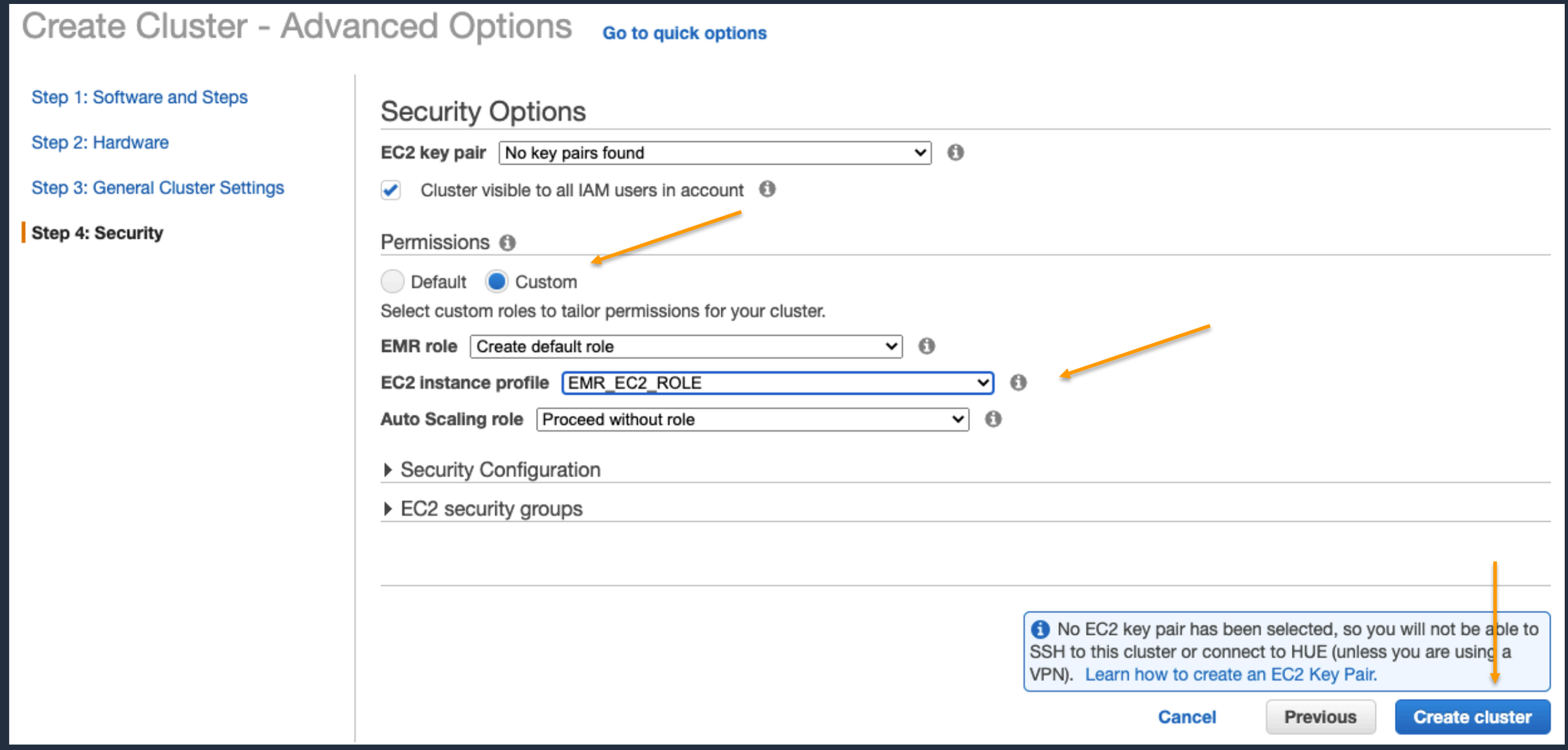


aws

How to launch an EMR cluster

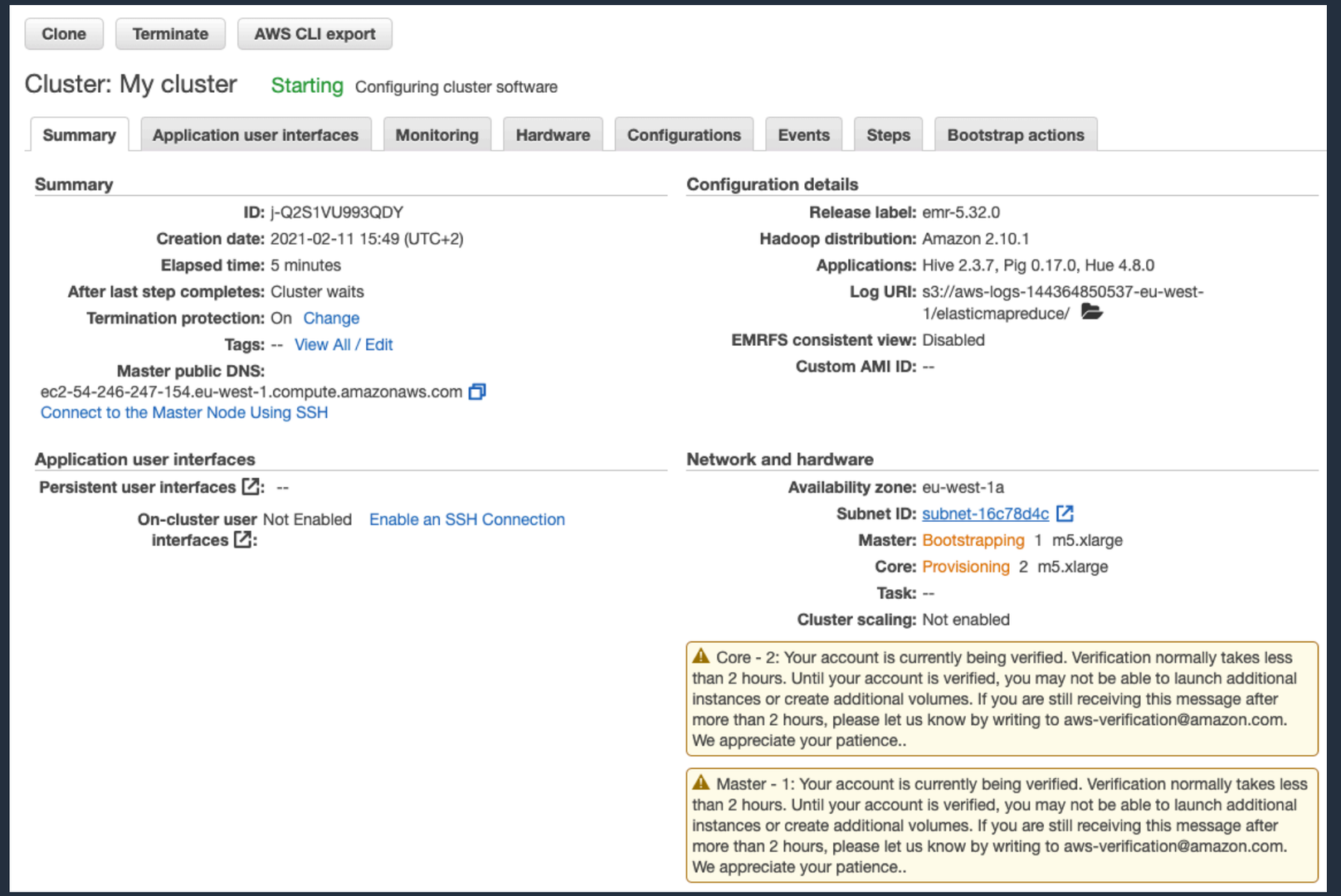
In the Security Tab, in the Permissions session, select Custom

And in the EC2 Instance profile select EMR_EC2_ROLE and click in Create Cluster



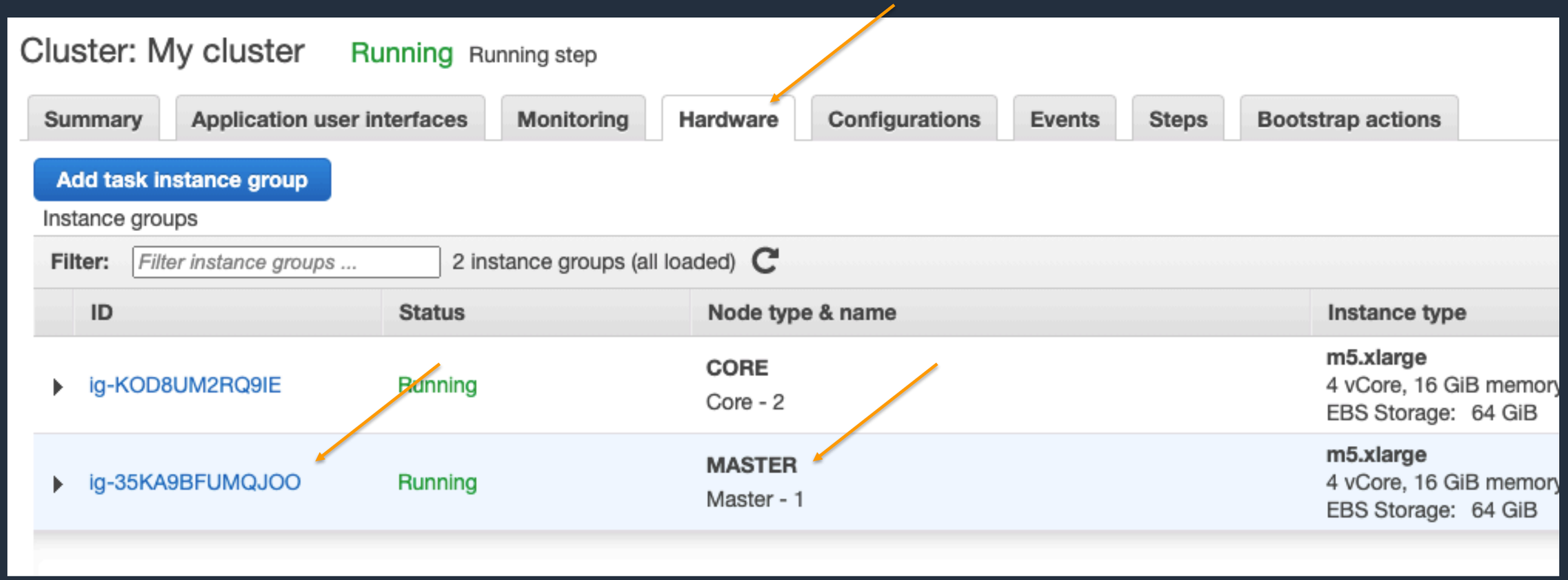
How to launch an EMR cluster

Your cluster will now be
starting, and resources
will be provisioning.
Please allow 5-10min for your cluster to provision successfully. The status will move from starting,
to waiting.



How to launch an EMR cluster

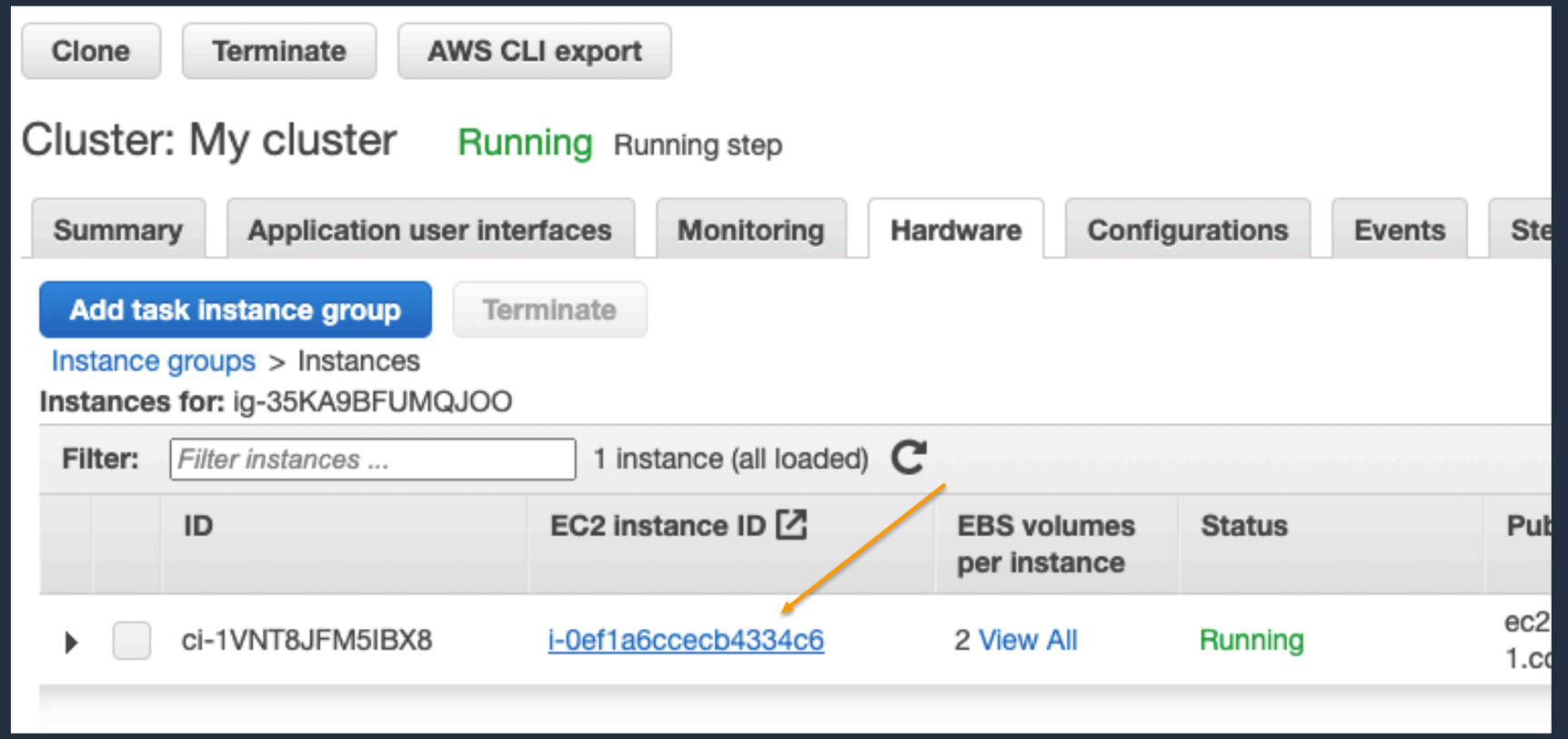
To connect to your cluster select the Hardware Tab, and click in the ID for the Master Node



aws

How to launch an EMR cluster

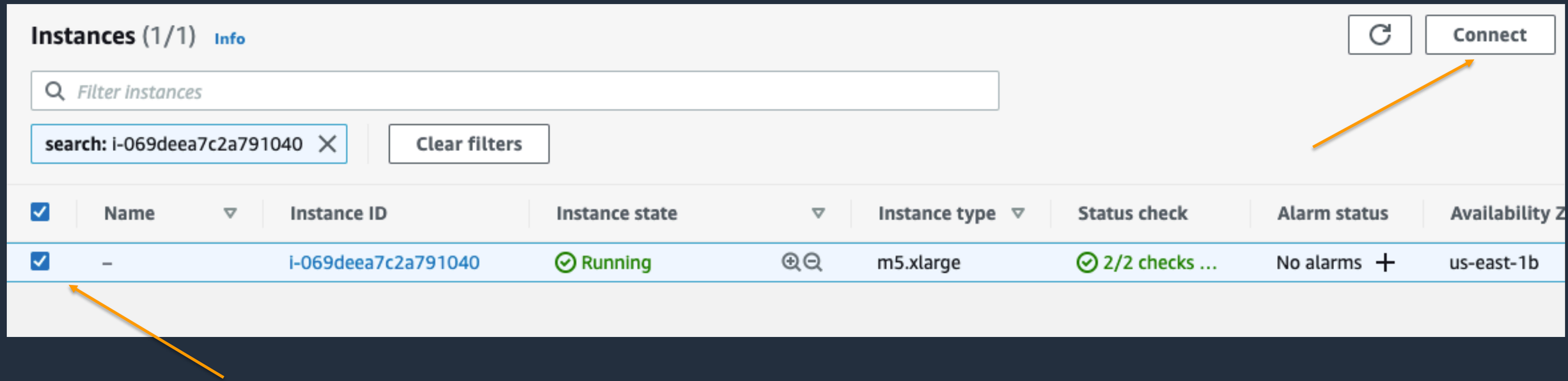
Click in the EC2 Instance ID Link



aws

How to launch an EMR cluster

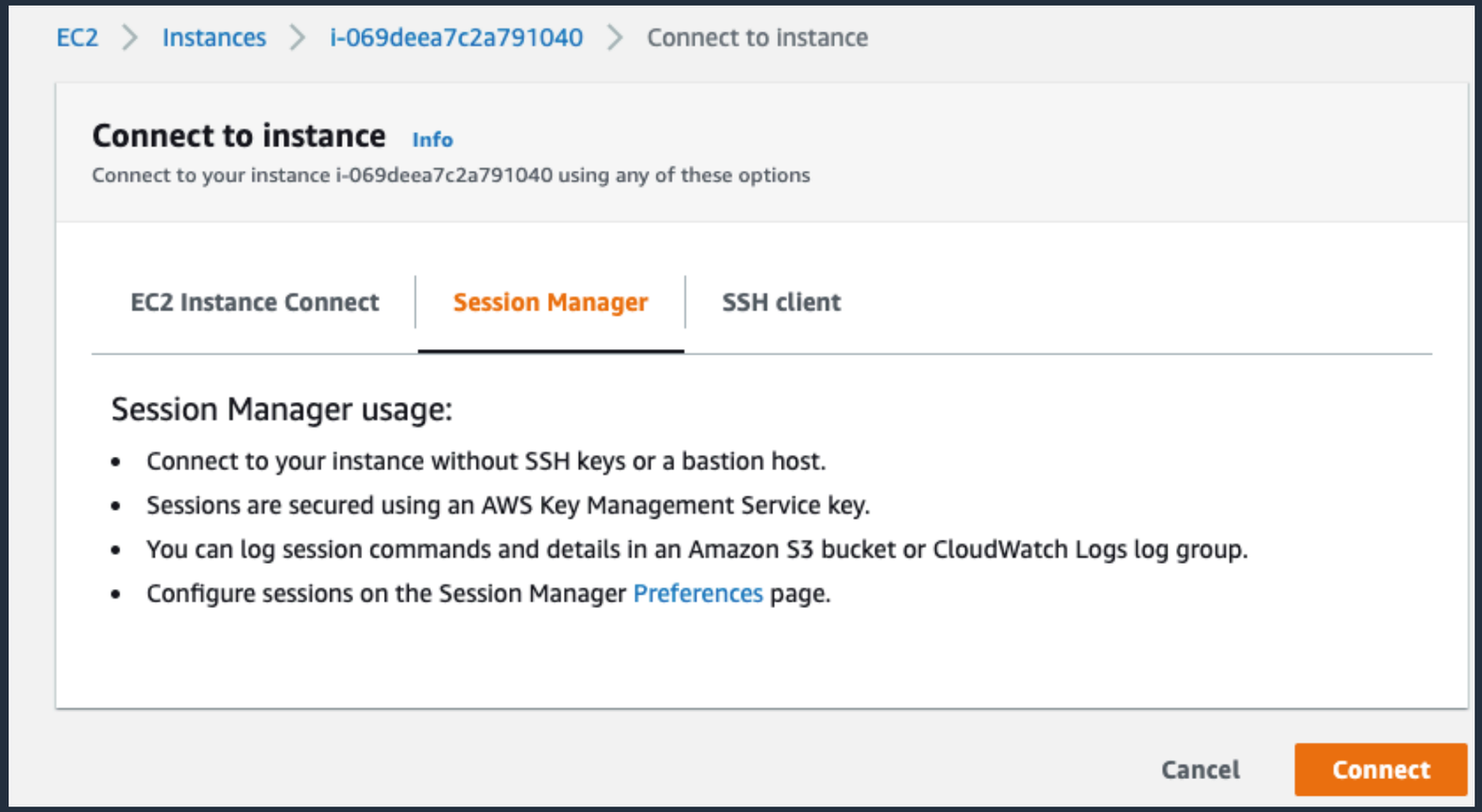
In the EC2 Instance Page, select the instance ID and Click in Connect



aws

How to launch an EMR cluster

Select the Session Manager option and click in Connect



How to launch an EMR cluster

In the terminal change run the below commands

sudo su - hadoop

You need to work with the hadoop user

aws

Introduction to Big Data

aws

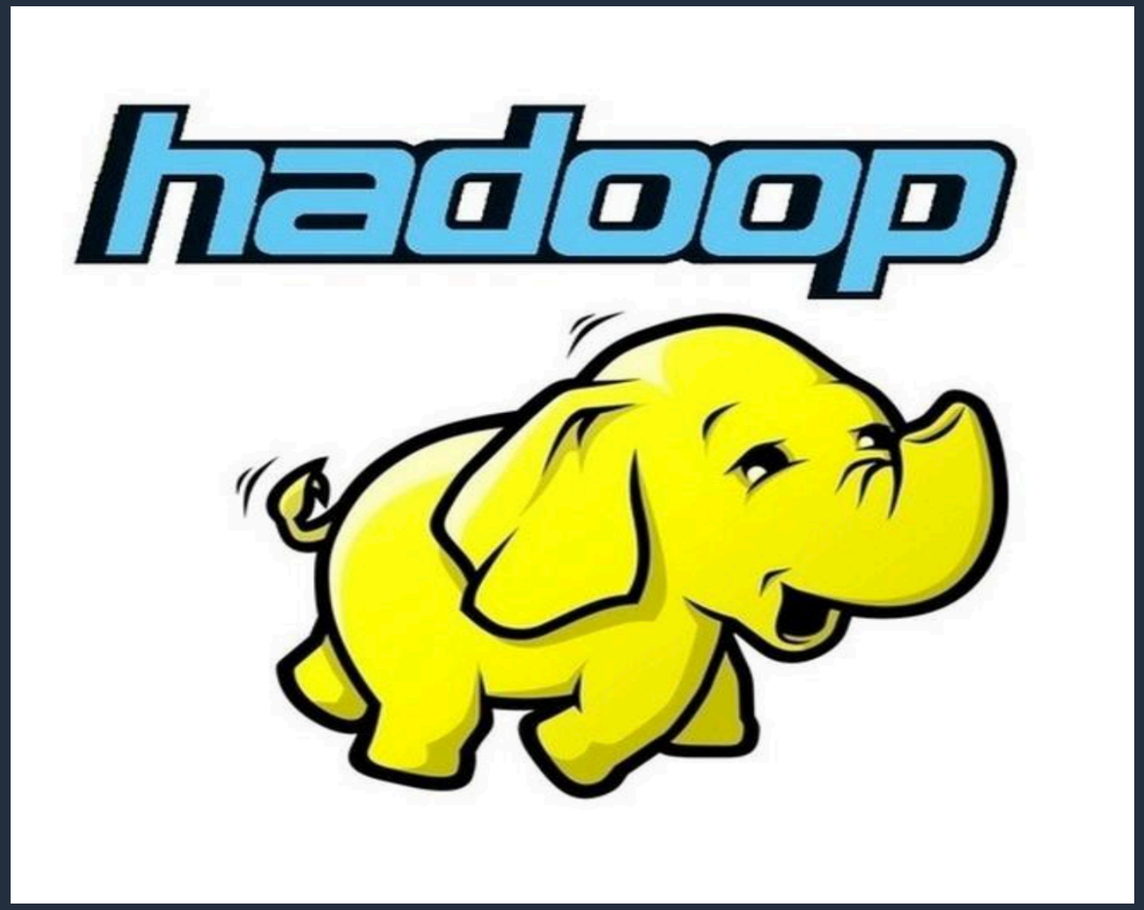
Introduction to Big Data

- Big data can be described in terms of data management challenges that – due to increasing volume, velocity and variety of data – cannot be solved with traditional databases. While there are plenty of definitions for big data, most of them include the concept of what's commonly known as **"three V's" of big data**:

- ✓ **Volume:** Ranges from terabytes to petabytes of data
- ✓ **Variety:** Includes data from a wide range of sources and formats (e.g. web logs, social media interactions, ecommerce and online transactions, financial transactions, etc)
- ✓ **Velocity:** Increasingly, businesses have stringent requirements from the time data is generated, to the time actionable insights are delivered to the users. Therefore, data needs to be collected, stored, processed, and analysed within relatively short windows – ranging from daily to real-time

aws

Introduction to Big Data



aws

Introduction to Hadoop

- Apache Hadoop is an open source project and it allows for the distributed processing of large data sets across clusters of nodes/instances using simple programming model.

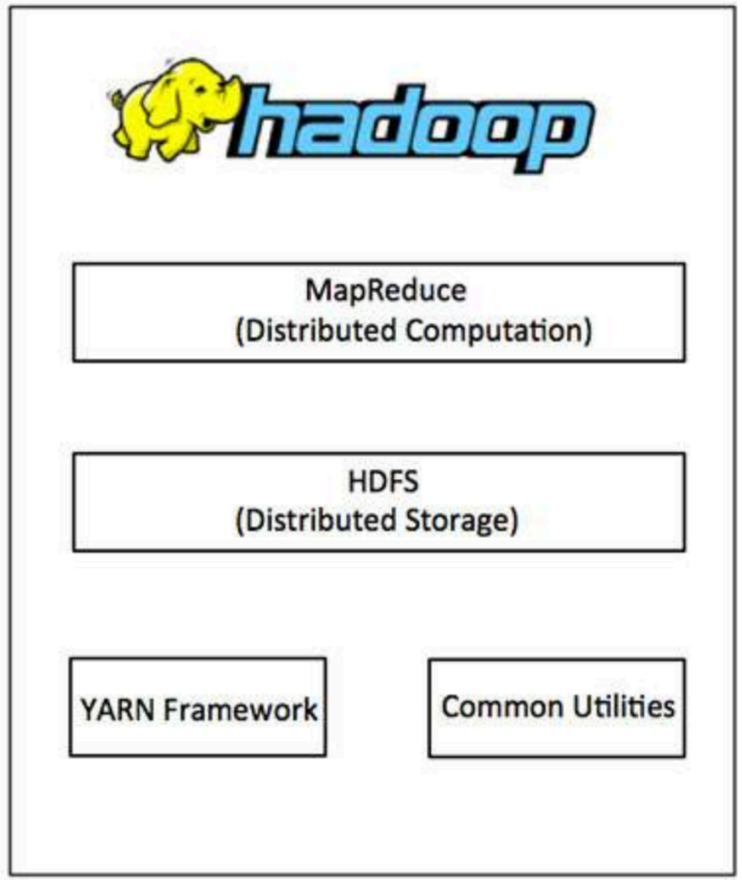
- Hadoop ecosystem is a scalable, fault tolerant and distributed system for data storage and processing.

- Core Hadoop has two main components
  - I. HDFS
  - II. MapReduce

Hadoop Architecture

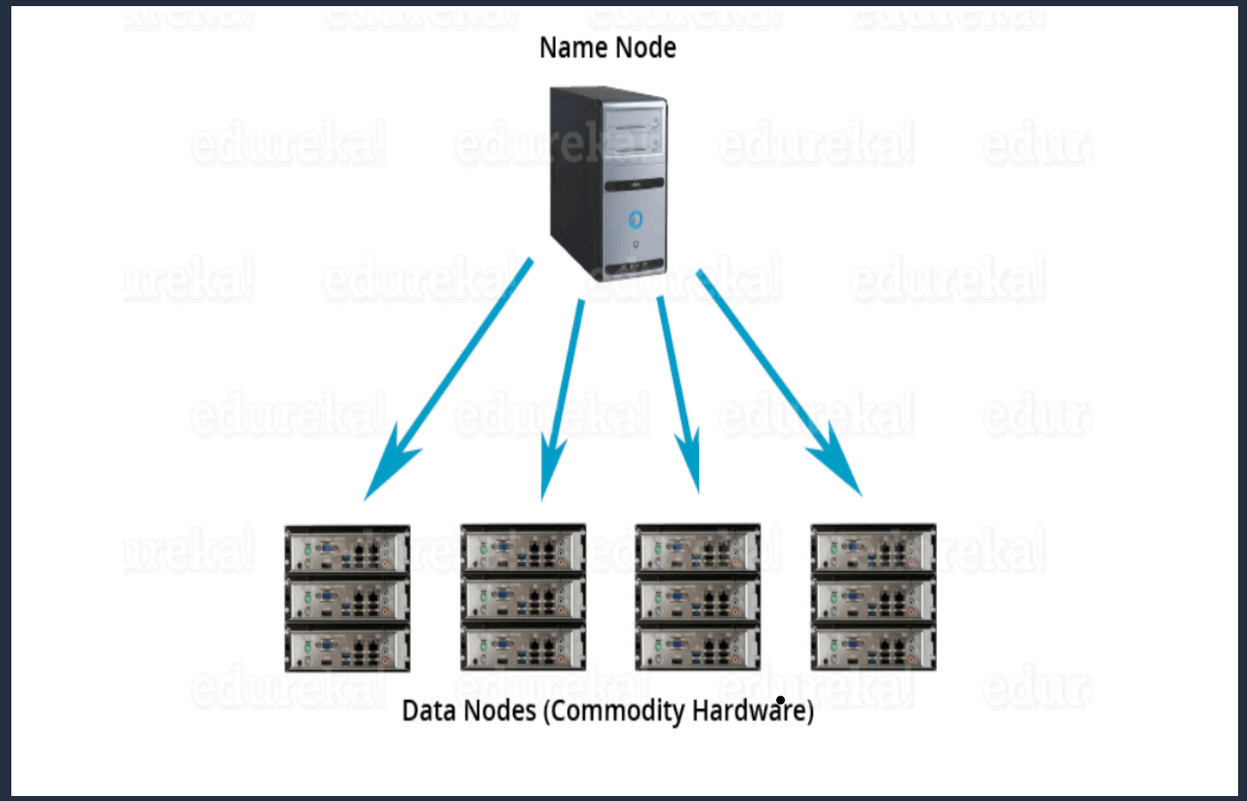
At its core, Hadoop has two major layers namely −

- Processing/Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed File System).



HDFS Storage on Disks

- Let’s understand the example provided here. Imagine we have 4 machines or computers with a hard drive of 1TB on each machine.
- When you install Hadoop as a platform on top of these machines, you will get HDFS as a storage service.
- Hadoop distribution File System is distributed in a such a way that every machine can access Hadoop Distributed file system from any of the four machines in the Hadoop cluster, you will feel as if you have logged into a single large machine which has a storage capacity of 4 TB (total storage over 4 machines).



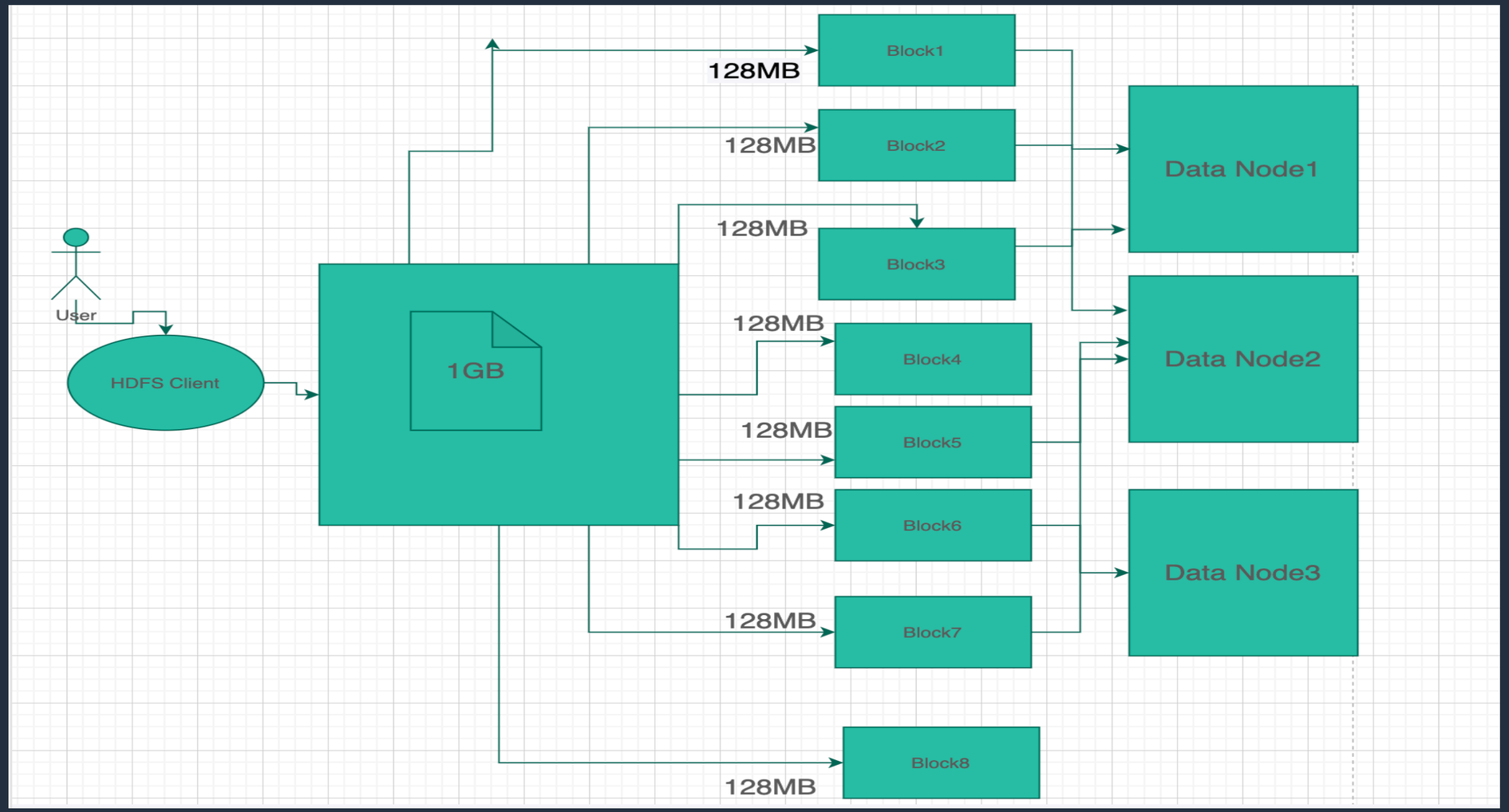
aws

Name Node Vs Data Node

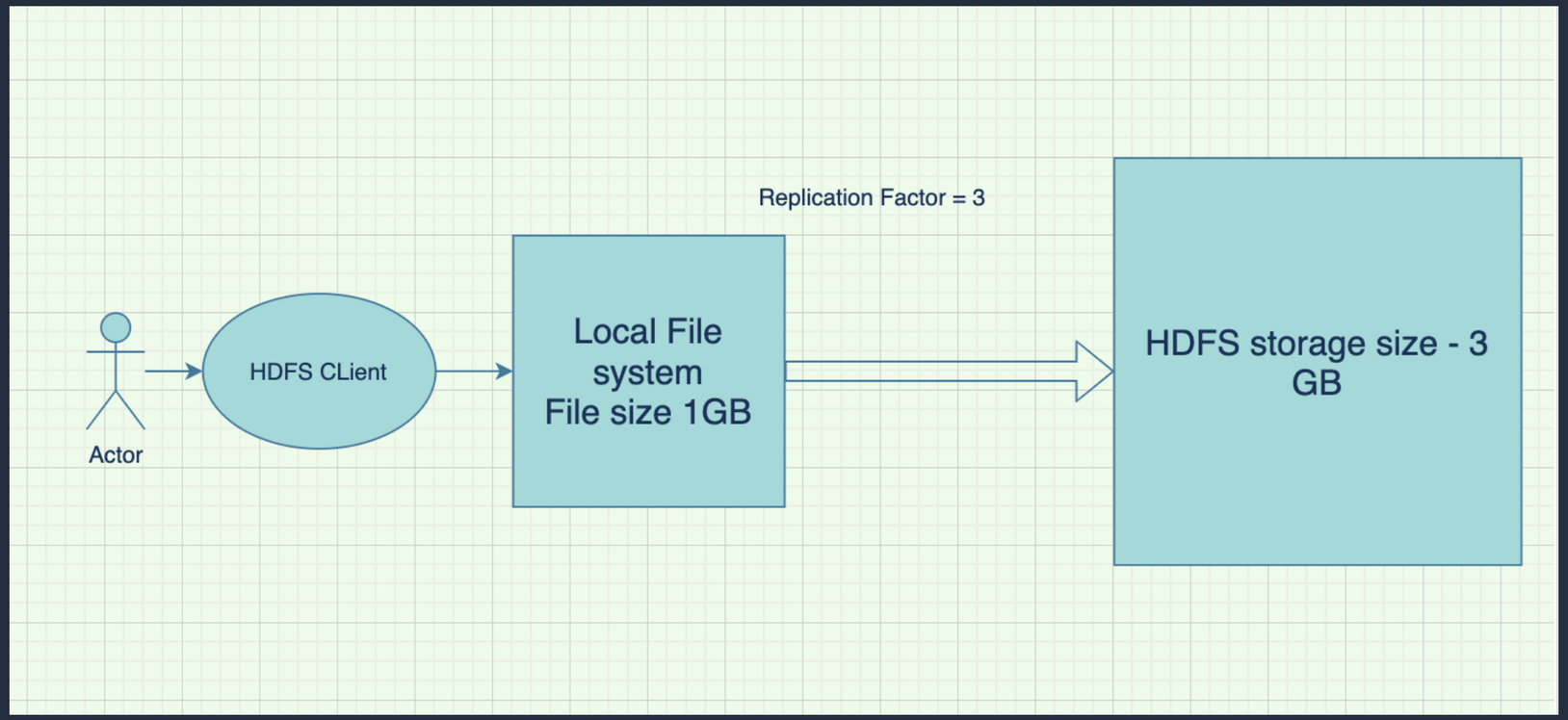
- The Name node (Stores metadata) is the centre piece of an HDFS file system. It keeps the directory tree of all files in the. file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself

- A Data node stores data in the HDFS. A functional filesystem has more than one Data node, with data replicated across them. Data in data nodes are stored as block. The block size In HDFS is 128MB.

HDFS Blocks and Split

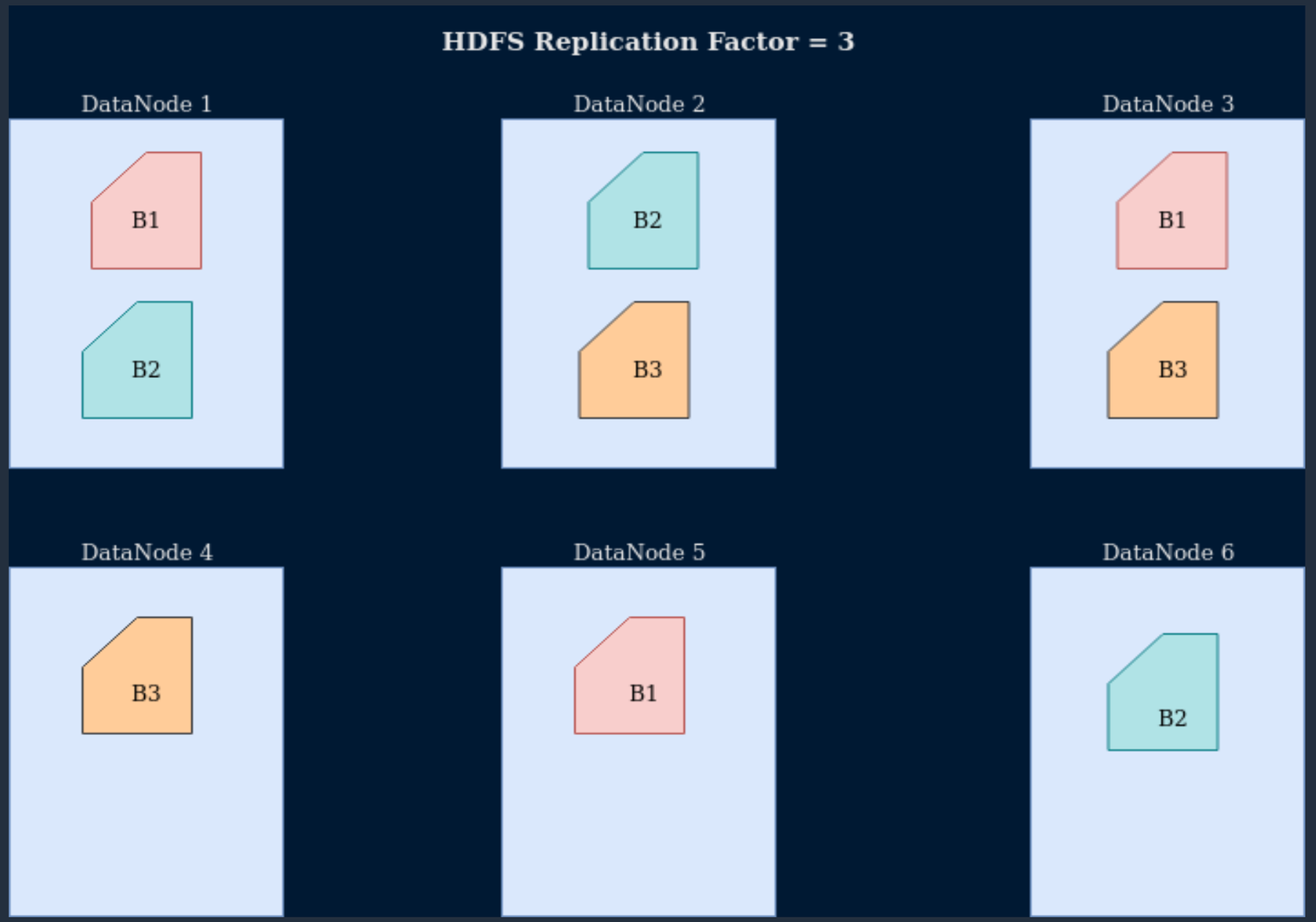


HDFS Replication Factor



aws

HDFS Replication Factor



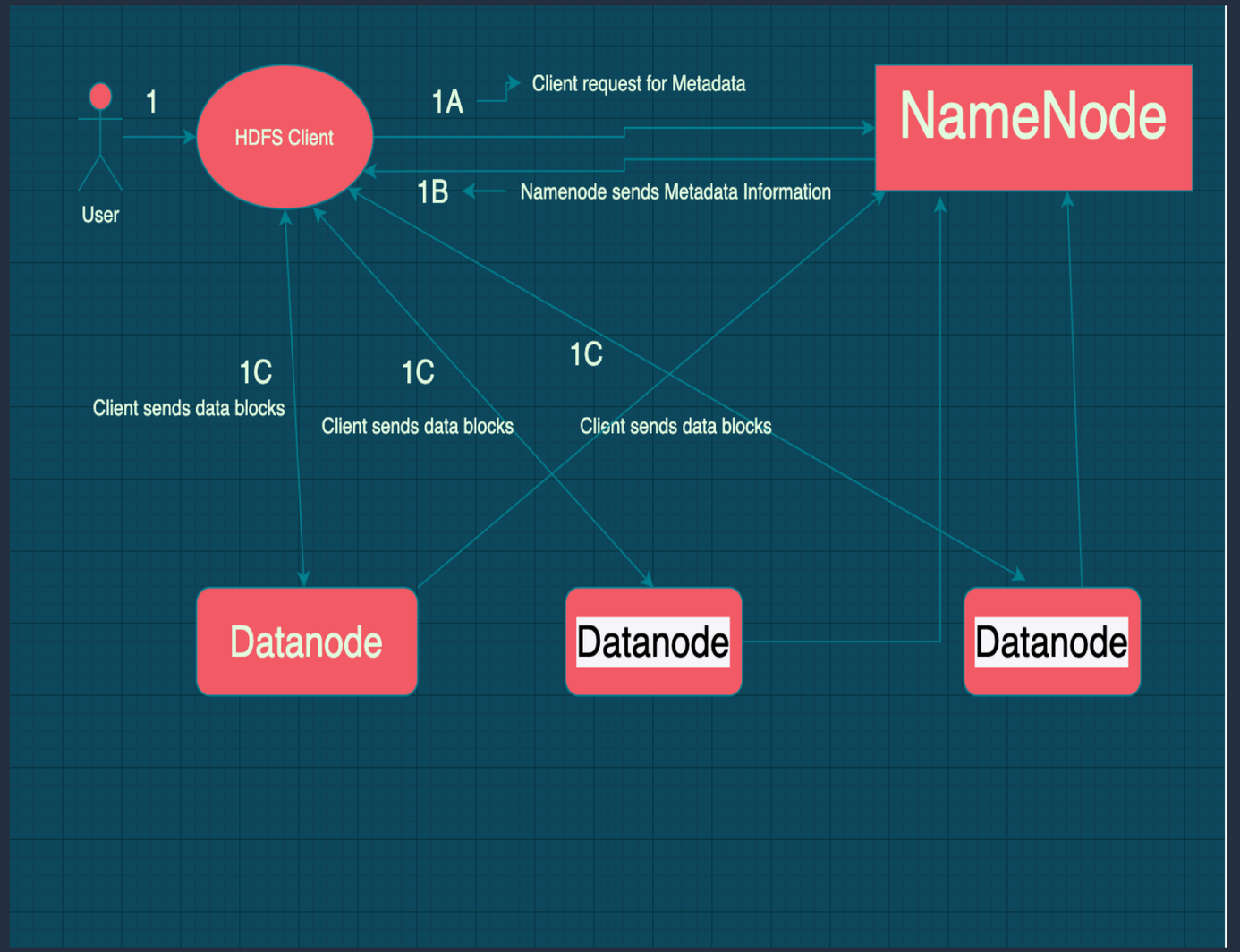
aws

How client read/write into HDFS

1. Client applications talk to the Name node whenever they wish to locate a file, or when they want to add/copy/move/delete a file.

2. The Name node responds the successful requests by returning a list of relevant Data node servers where the data lives.

3. Client applications can talk directly to a Data Node, once the Name node has provided the location of the data.



aws

Hands-on Lab on HDFS

- Login to the EMR master Node
- Run HDFS DFS and Admin commands
- Increase the replication factor
- Check the Name Node UI

aws

Checks on HDFS

Test below commands.

I. hdfs dfs –ls /

II. hdfs dfs –mkdir /user/test

III. hdfs dfsadmin –safemode enter

IV. Hdfs dfs –mkdir /user/test1

Q. Are you able to run the last command. If not, why?

aws

Break

Will resume in 15 minutes

aws

Break

Will resume in 15 minutes

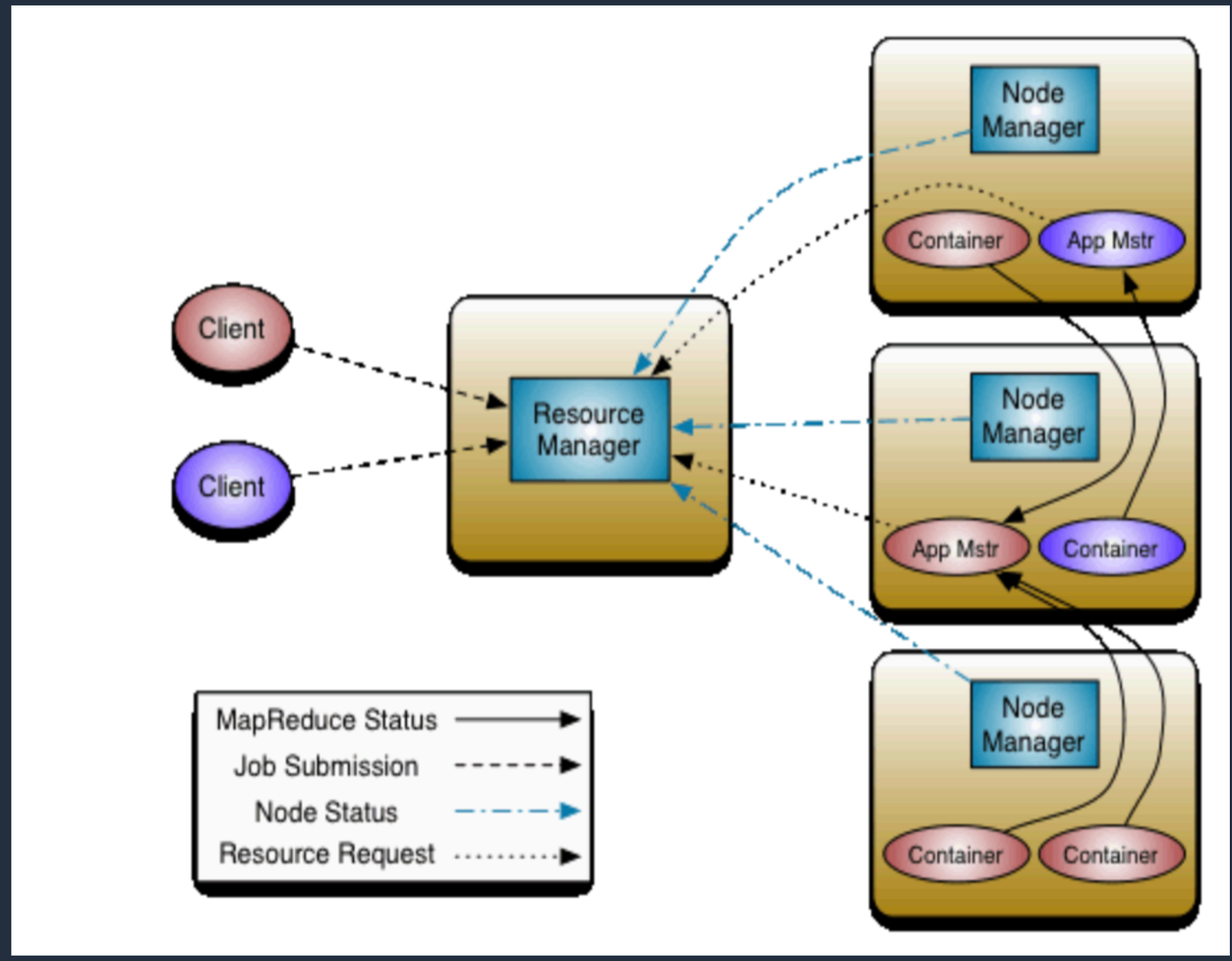
aws

YARN

- The idea is to have a global Resource Manager (*RM*) and per-application Application Master (*AM*). An application is either a single job or a DAG of jobs.

- The Resource Manager and the Node Manager form the data-computation framework.

- The per-application Application Master is, in effect, a framework specific library and is tasked with negotiating resources from the Resource Manager and working with the Node Manager(s) to execute and monitor the tasks.



aws

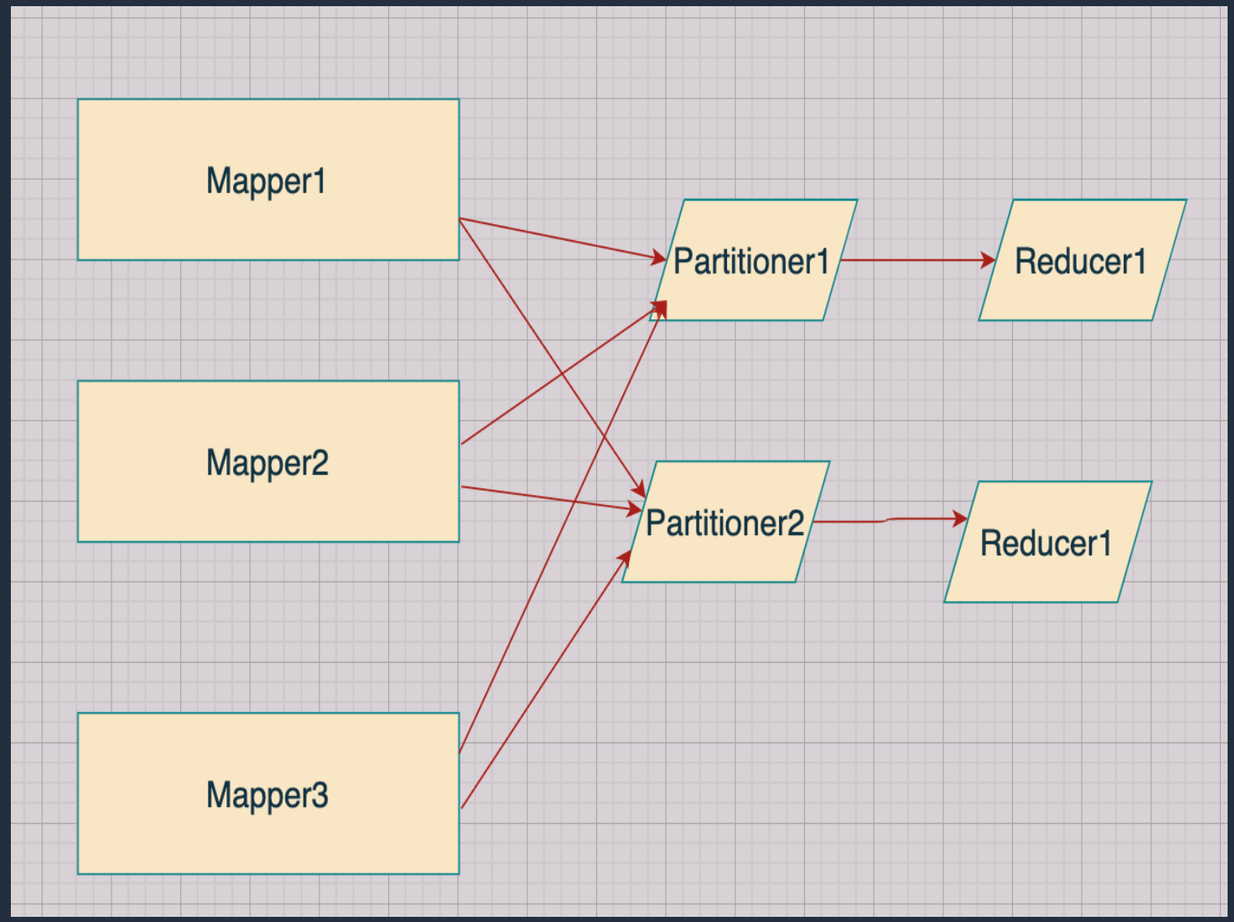
MapReduce

1. A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner.

2. The input determines number of map task. The output of the map is written into the disk.

3. Reducers pull the map output from the disk into memory(shuffle stage)

4.The final output is written into HDFS.



No. of Mapper = (total data size)/ (input split size)

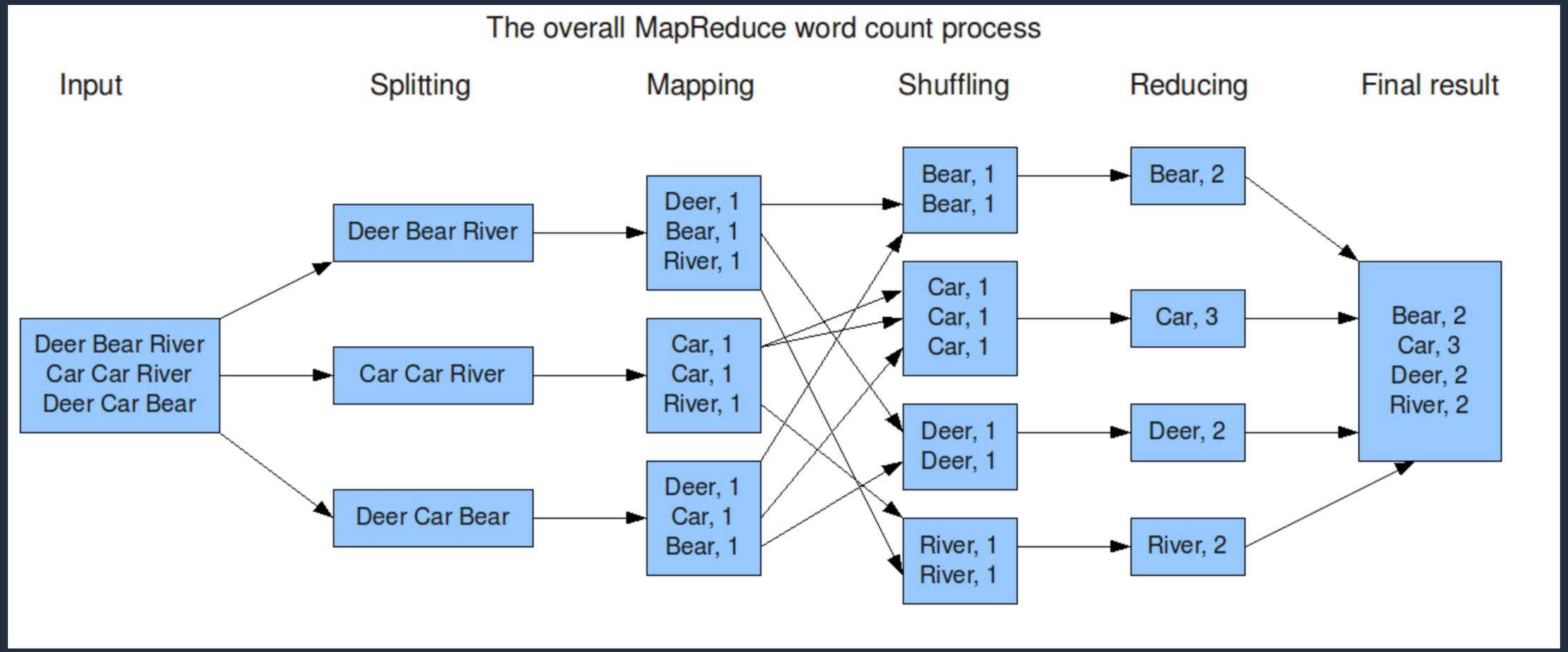
aws

Checks on Map Reduce

Q. Where do you see the final output of Reducer?

aws

MapReduce Flow Diagram



aws

Checks on Map Reduce

Q. True or False

Reducers can run before Map stage completion

aws

Checks on Map Reduce

Q. How many mappers will run for a file of size 1GB using default block size?

Checks on Map Reduce

Q. What is shuffle and sort stage?

aws

Checks on Map Reduce

Q. Can you run Map only job? (no Reducer)

aws

MapReduce Lab

1. Run a simple MapReduce Wordcount example

2. Check how Map and Reduce works (Verify the counter)

3. Open the WEB-UI of the Resource Manager and click on the application master

aws

What is Amazon S3

- Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance.

- This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

- Amazon S3 provides easy-to-use management features so you can organize your data and configure finely-tuned access controls to meet your specific business, organizational, and compliance requirements.

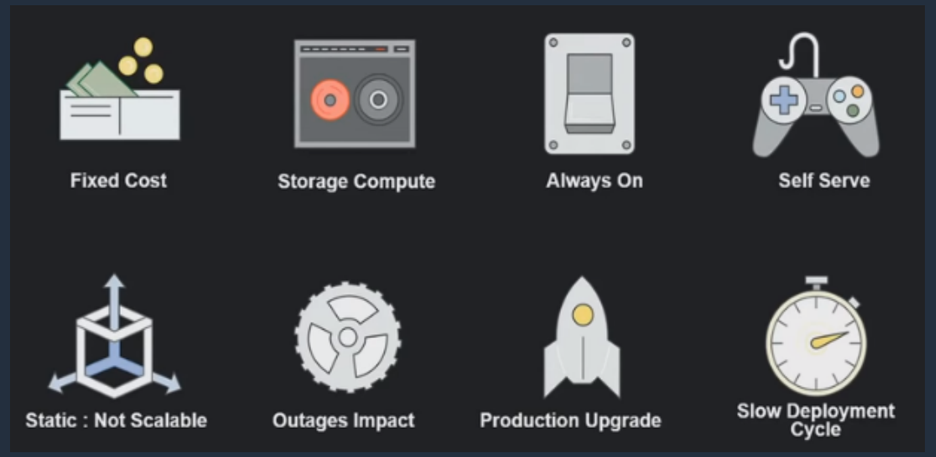
- Amazon S3 is designed for 99.999999999% (11 9's) of durability, and stores data for millions of applications for companies all around the world.

EMR and S3??

aws

Advantages of Hadoop on Amazon EMR

- Increased speed and agility
- Reduced administrative complexity
- Integration with other AWS services
- Pay for clusters only when you need them



aws

Hadoop Ecosystem

Applications and frameworks in the Hadoop ecosystem

- Hadoop commonly refers to the actual Apache Hadoop project, which includes MapReduce (execution framework), YARN (resource manager), and HDFS (distributed storage).

- You can also install Apache TEZ a next-generation framework which can be used instead of Hadoop MapReduce as an execution engine. Amazon EMR also allow Hadoop to use Amazon S3 as a storage layer.

- However, there are also other applications and frameworks in the Hadoop ecosystem, including tools that enable low-latency queries, GUIs for interactive querying, a variety of interfaces like SQL, and distributed NoSQL databases.

- The Hadoop ecosystem includes many open source tools designed to build additional functionality on Hadoop core components, and you can use Amazon EMR to easily install and configure tools such as Hive, Pig, Hue, Ganglia, Oozie, and HBase on your cluster. You can also run other frameworks, like Apache Spark for in-memory processing, or Presto for interactive SQL, in addition to Hadoop on Amazon EMR.

aws

Amazon EMR Applications

Application Index

- Amazon EMR is one of the largest Spark and Hadoop service providers in the world, enabling customers to run ETL, machine learning, real-time processing, data science, and low-latency SQL at petabyte scale.
- See the EMR Release page for up to date schedules. Users can bootstrap applications not listed below.

Administrative

- Ganglia — Cluster monitoring
- Livy — REST API interacting with Spark
- Oozie — Workflow Scheduler
- ZooKeeper — Configure & Sync node

Machine Learning

- Mahout — Machine Learning
- MXNet — Deep Learning
- SparkML — Machine Learning
- TensorFlow — Deep Learning

Data Movement

- Sqoop — Relational DB connector

NoSQL

- HBase — Hadoop's non-relational DB

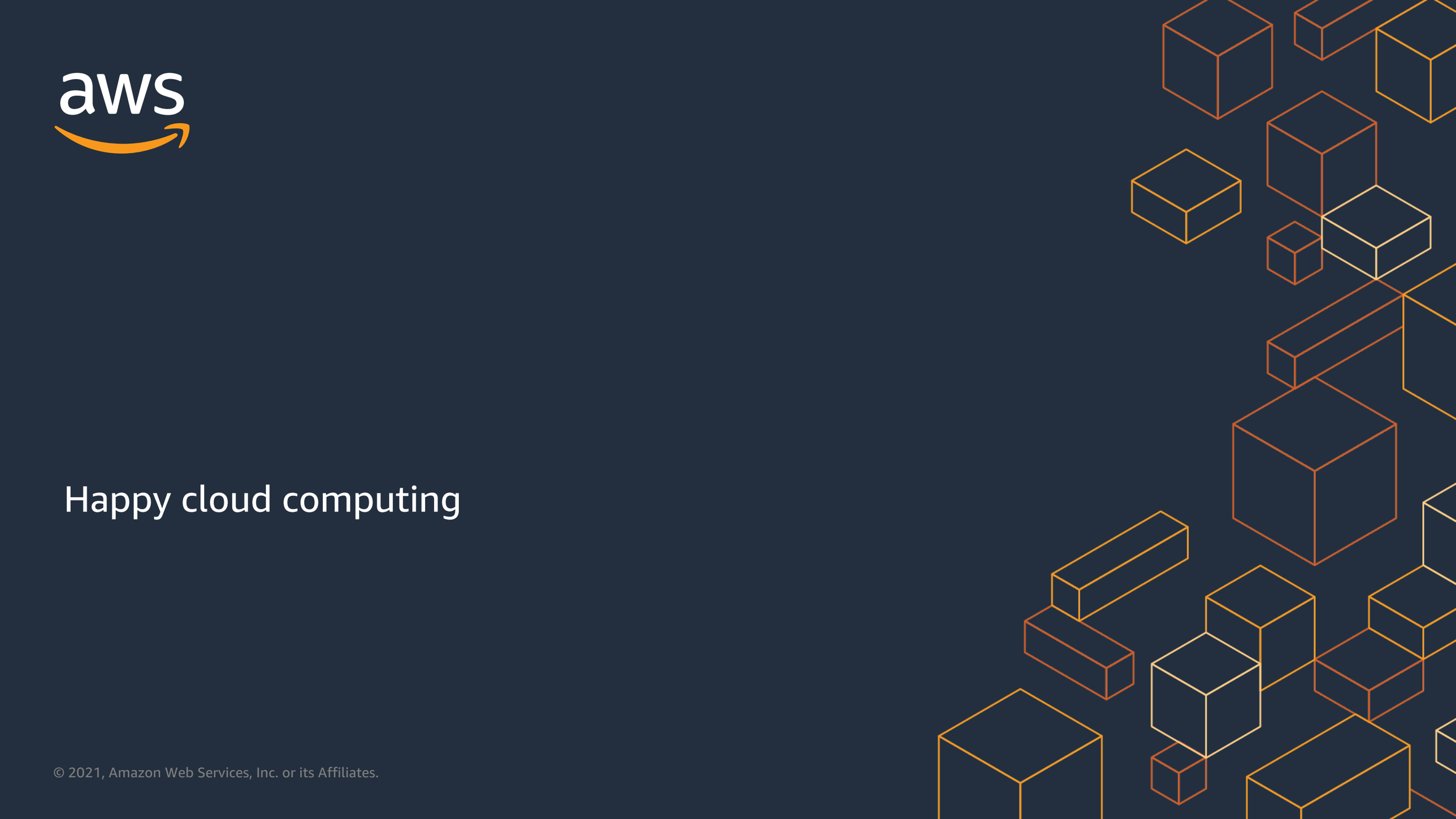
Data Processing

- Flink — Stream Processing
- Hive — Hadoop Data warehouse
- MapReduce — Batch Data Processing
- Presto — Distributed SQL for Big Data
- Spark — In-memory Data Processing & Machine Learning
- Tez — Interactive Data Processing
- Hudi — Update Data Lake Storage

Query Tools

- EMR Notebooks — Serverless notebook
- Hue — Visualization and Querying for Hadoop
- JupyterHub — Multi-user Juptyer Notebook (installed on cluster)
- Phoenix — Querying for HBase
- Pig — Scripting language for MapReduce jobs
- Zeppelin — Data Science notebook

Q&A



Happy cloud computing