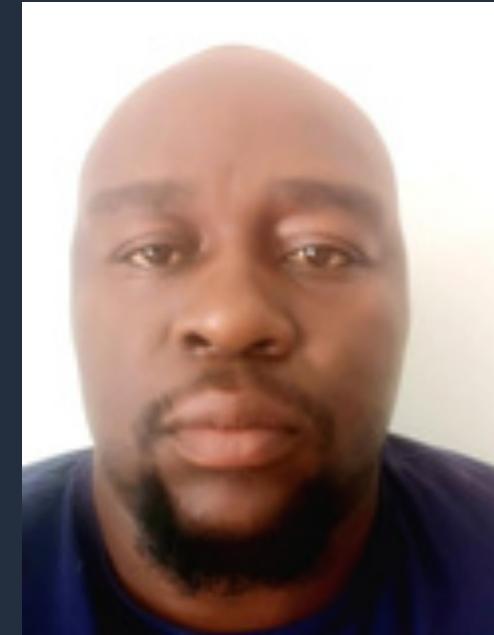


House Rules

- 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

About AWS Premium Support

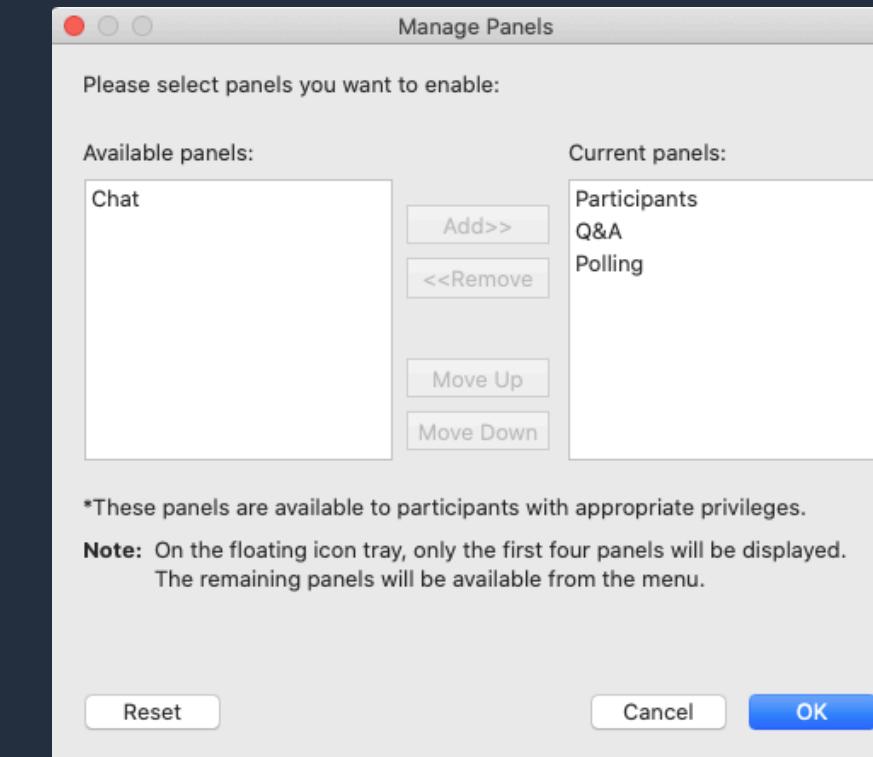
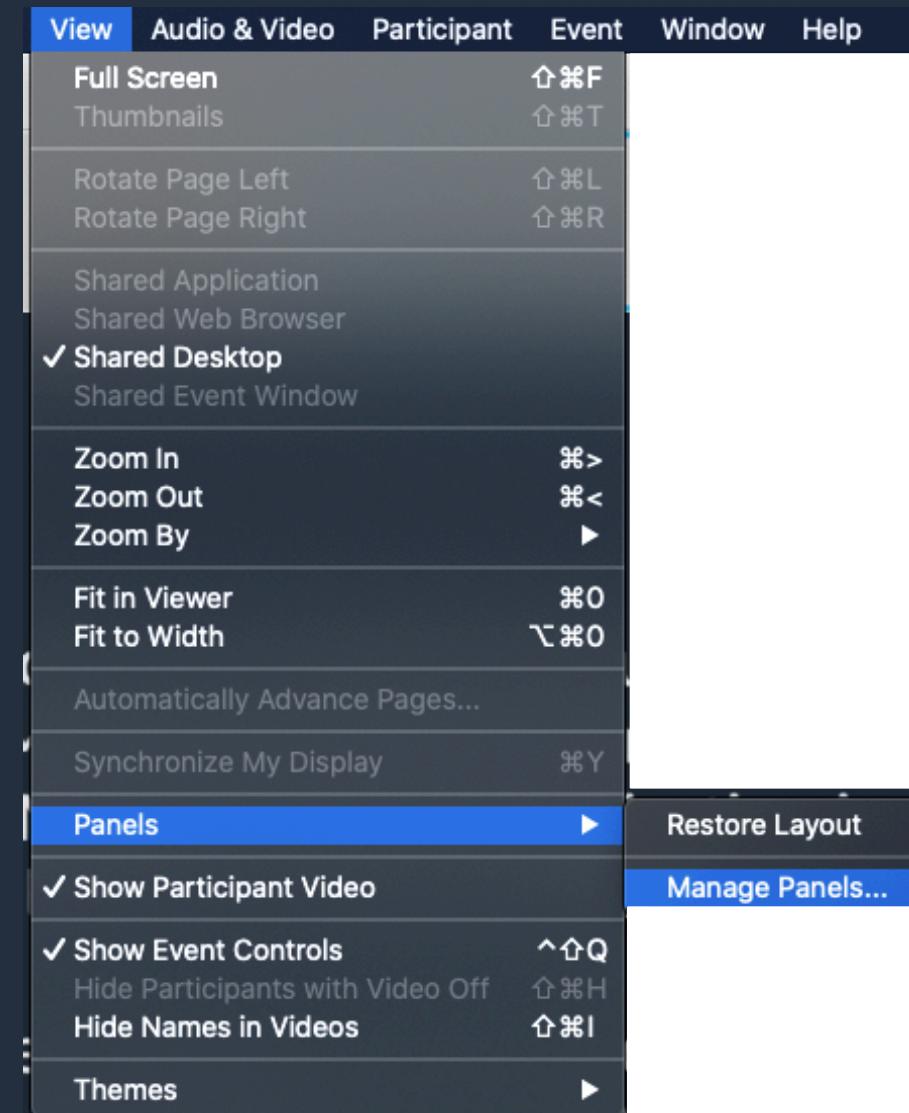
AWS Premium Support is one-on-one, fast-response support from experienced technical support engineers. The service helps customers use AWS products and features. The **Big Data Cloud Support Engineering** team is one of our teams within Premium Support. Our Engineers support our customers around the 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.



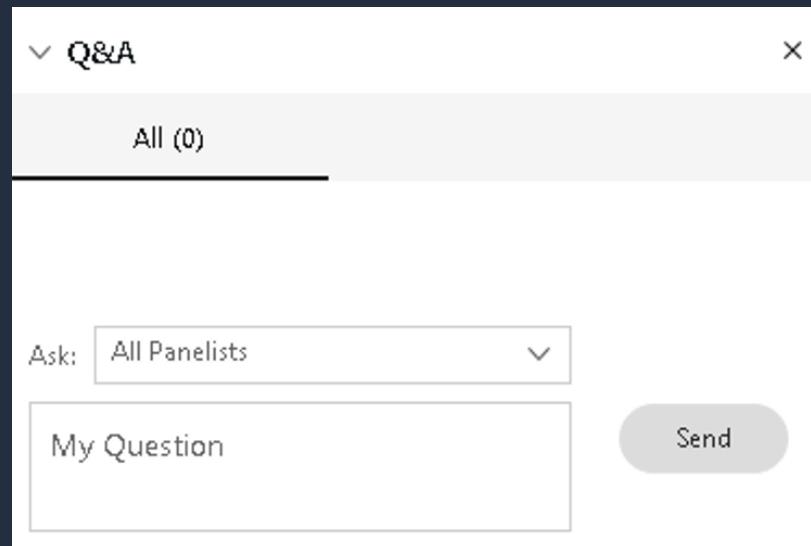
Ops Manager - Brian

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.

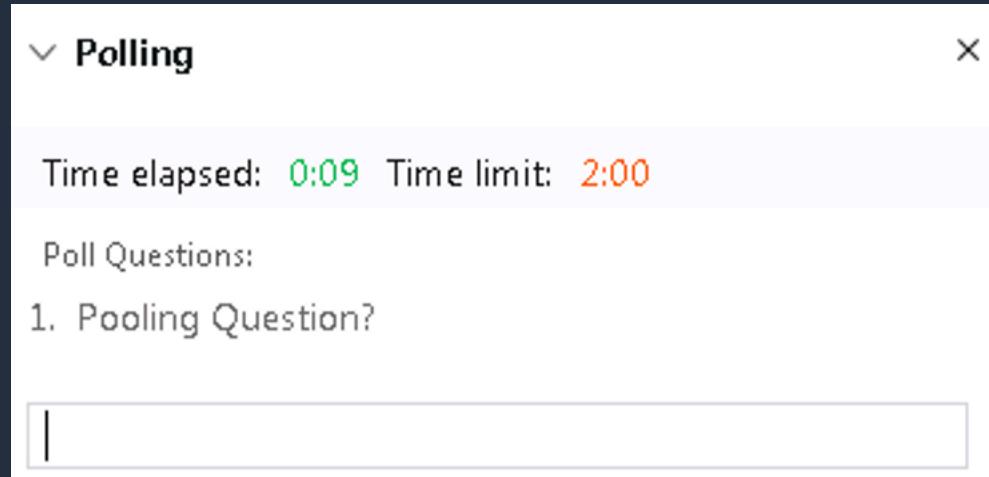
Cisco Webex



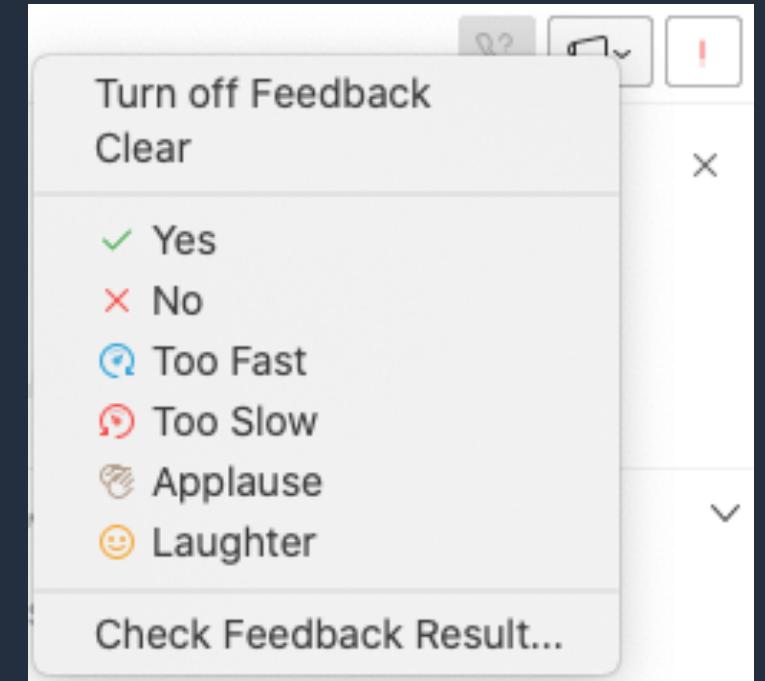
Q&A



Pooling Questions



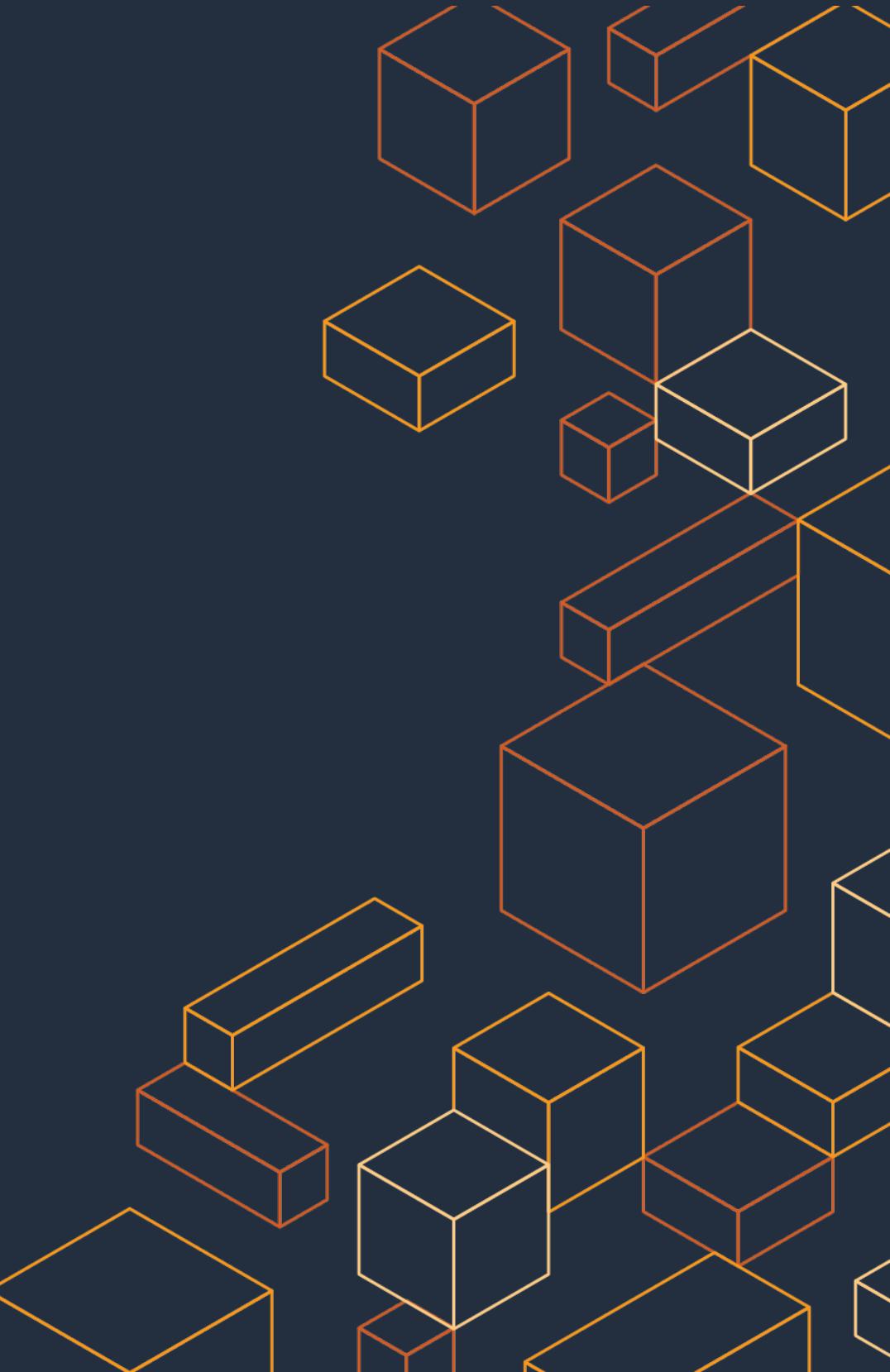
Interaction





Getting started with Amazon EMR

Feb 18th August 2021



Introduction

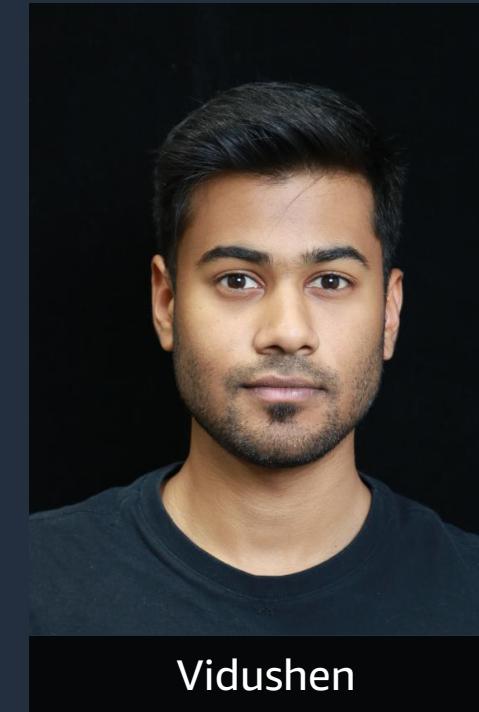
- Instructors



Patrick



Peter



Vidushen

Introduction

Team



Vidushen



Patrick



Peter



Emmanuel



Umesh



Amit



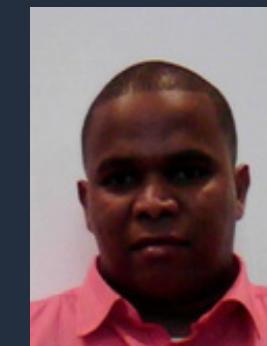
David



Ridick



Andrew



Neil



Riyaad

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



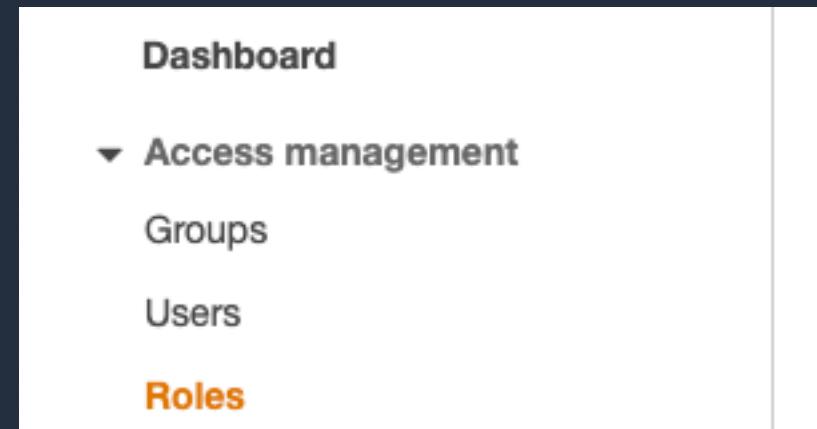
Cluster Setup

How to launch an EMR cluster

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

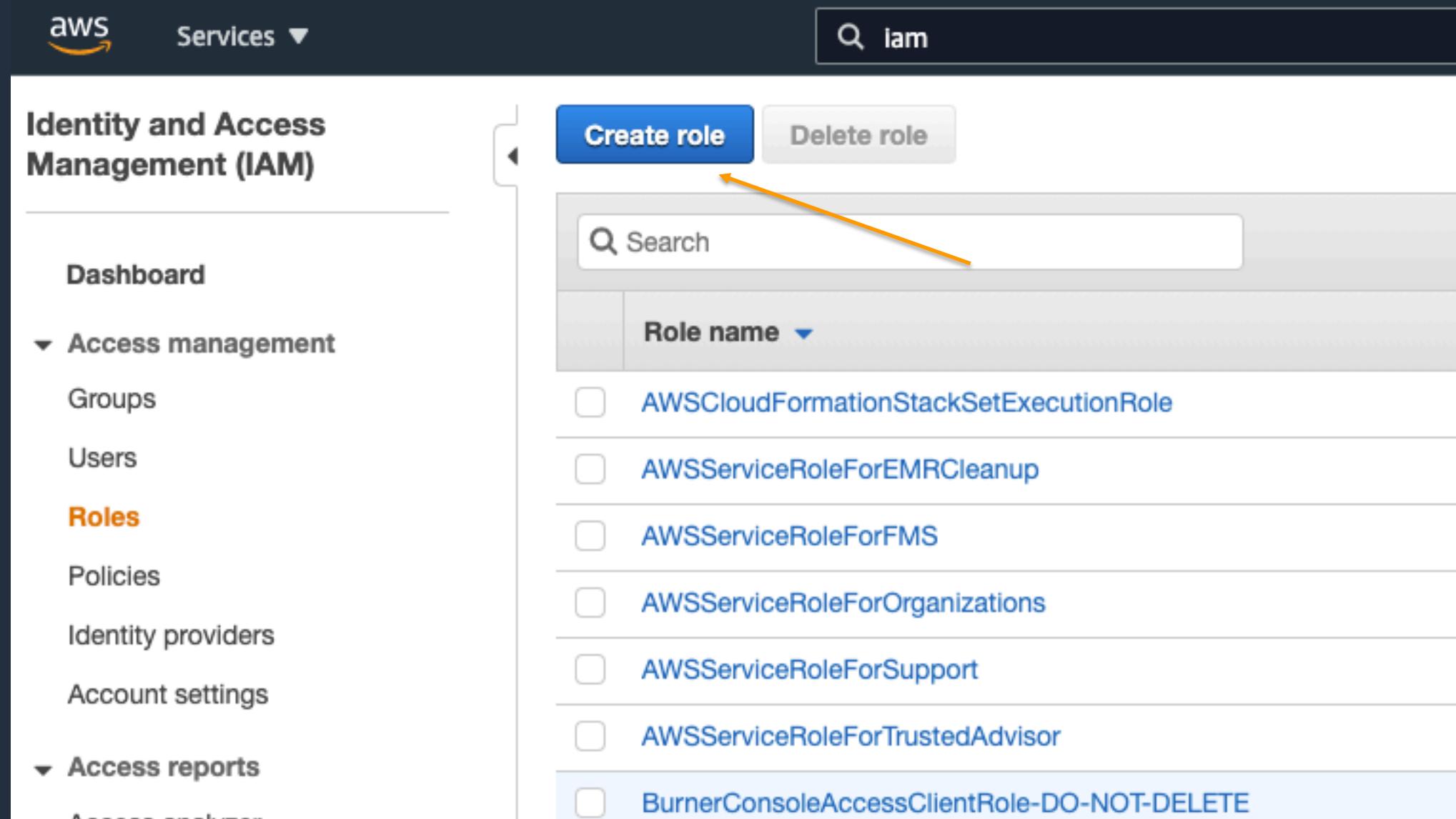
The screenshot shows the AWS search interface with a search bar containing 'iam'. Below the search bar, the text 'Search results for 'iam'' is displayed. On the left, there's a sidebar with categories: 'Services (1)', 'Features (10)', 'Documentation (55,458)', and 'Marketplace (192)'. The main area shows a 'Services' section with a box for 'IAM' which is described as 'Manage access to AWS resources'. Below this is a 'Features' section with a link to 'See all 10 results'.

And select the Roles



How to launch an EMR cluster

Click in Create role



How to launch an EMR cluster

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

Create role

1 2 3 4

Select type of trusted entity

AWS service EC2, Lambda and others

Another AWS account Belonging to you or 3rd party

Web identity Cognito or any OpenID provider

SAML 2.0 federation Your corporate directory

Allows AWS services to perform actions on your behalf. [Learn more](#)

Choose a use case

Common use cases

EC2

Lambda

Allows EC2 instances to call AWS services on your behalf.

Allows Lambda functions to call AWS services on your behalf.

* Required

Cancel

Next: Permissions

© 2021, Amazon Web Services, Inc. or its Affiliates.

aws

How to launch an EMR cluster

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

Create role

1 2 3 4

▼ Attach permissions policies

Choose one or more policies to attach to your new role.

Create policy

Filter policies ▾ Q AmazonEC2RoleforSSM Showing 1 result

Policy name ▾	Used as
<input checked="" type="checkbox"/> ➔  AmazonEC2RoleforSSM	Permissions policy (1)

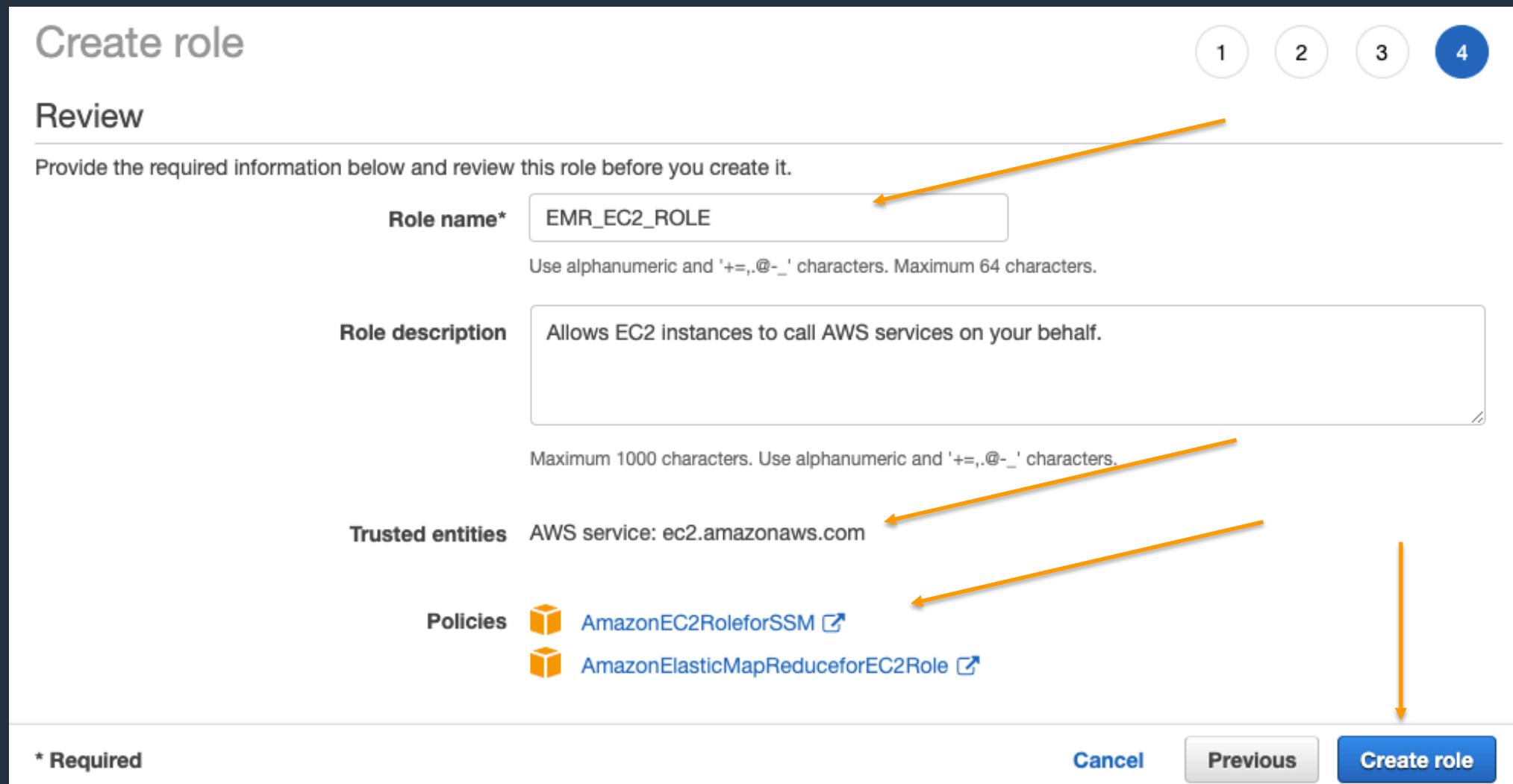
Filter policies ▾ Q AmazonElasticMapReduceforEC2Role

Policy name ▾
<input checked="" type="checkbox"/> ➔  AmazonElasticMapReduceforEC2Role

How to launch an EMR cluster

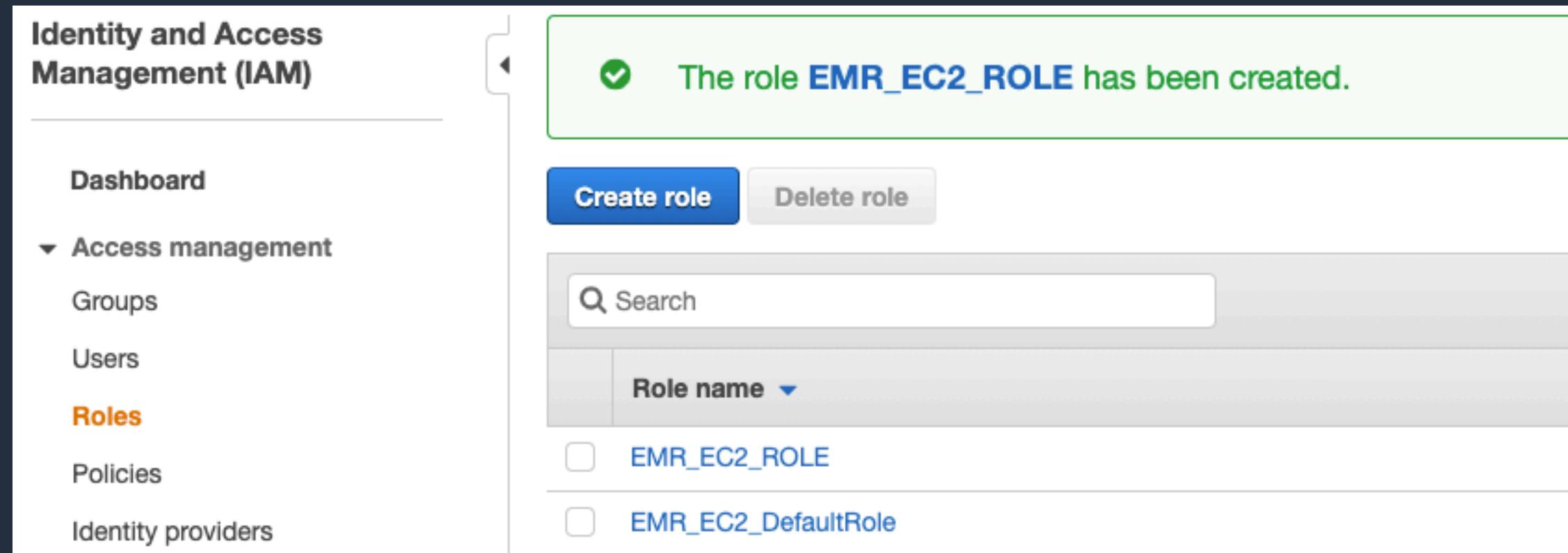
In the Tags session click in Next, In the Review add EMR_EC2_ROLE as Role name.

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



How to launch an EMR cluster

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



How to launch an EMR cluster

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

The screenshot shows the AWS console interface. On the left, there is a sidebar with the AWS logo and a 'Services' dropdown. Below it, under 'Identity and Access Management (IAM)', are links for 'Dashboard', 'Access management' (with a downward arrow), 'Groups', 'Users', and 'Policies'. The main area has a search bar at the top with the query 'emr' and a clear button 'X'. The search results are titled 'Search results for 'emr'' and show two items: 'EMR Managed Hadoop Framework' and 'AWS Glue DataBrew'.

aws Services ▾

Identity and Access Management (IAM)

Dashboard

Access management

Groups

Users

Policies

Search results for 'emr'

Services (2)

Documentation (27,372)

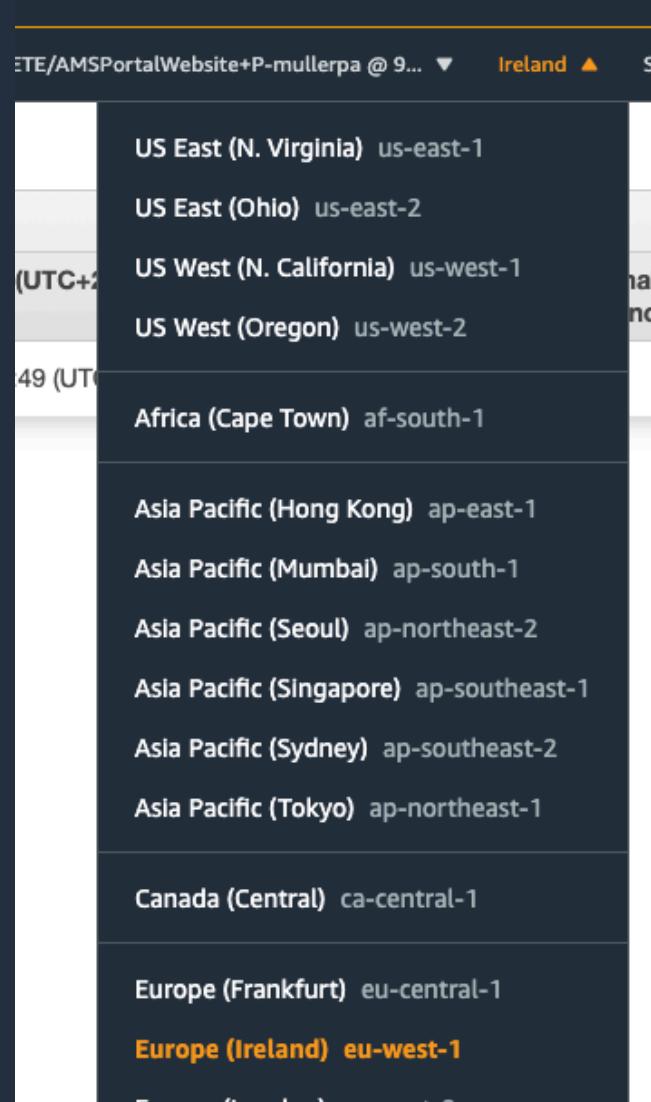
Marketplace (87)

EMR
Managed Hadoop Framework

AWS Glue DataBrew

How to launch an EMR cluster

Select the Ireland Region



How to launch an EMR cluster

Click on
"Create cluster"

Amazon EMR

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)

How Elastic MapReduce Works

<p>Upload</p> 	<p>Create</p> 	<p>Monitor</p> 
--	--	---

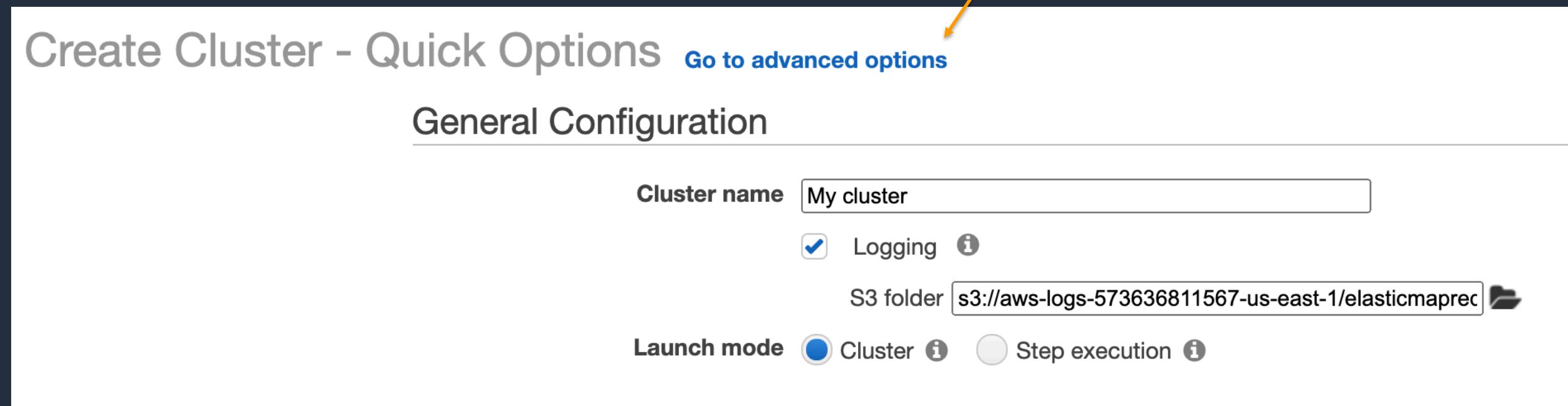
Upload your data and processing application to S3.

Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

Monitor the health and progress of your cluster. Retrieve the output in S3.

How to launch an EMR cluster

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



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”

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release **emr-5.30.1**

<input checked="" type="checkbox"/> Hadoop 2.8.5	<input type="checkbox"/> Zeppelin 0.8.2
<input type="checkbox"/> JupyterHub 1.1.0	<input checked="" type="checkbox"/> Tez 0.9.2
<input checked="" type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 1.4.13
<input checked="" type="checkbox"/> Hive 2.3.6	<input type="checkbox"/> Presto 0.232
<input type="checkbox"/> MXNet 1.5.1	<input type="checkbox"/> Sqoop 1.4.7
<input type="checkbox"/> Hue 4.6.0	<input type="checkbox"/> Phoenix 4.14.3
<input checked="" type="checkbox"/> Spark 2.4.5	<input type="checkbox"/> HCatalog 2.3.6

Multiple master nodes (optional)

Use multiple master nodes to improve cluster availability. [Learn more](#)

AWS Glue Data Catalog settings (optional)

Use for Hive table metadata i

Use for Spark table metadata i

Edit software settings i

Enter configuration Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

How to launch an EMR cluster

In the Hardware Configuration just click in Next

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration i

Specify the networking and hardware configuration for your cluster. Request Spot instances (unused EC2 capacity) to save money.

Cluster Composition

Specify the configuration of the master, core and task nodes as an instances group or instance fleet. This choice applies to all nodes for the lifetime of the cluster. Instance fleets and instance groups cannot coexist in a cluster. [see this topic](#) ↗

Instance group configuration

Uniform instance groups
Specify a single instance type and purchasing option for each node type.

Instance fleets
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#) ↗

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”.

The screenshot shows the "Create Cluster - Advanced Options" page in the AWS EMR console. The URL in the browser is <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster>. The page is titled "Step 3: General Cluster Settings".

General Options:

- Cluster name: My cluster
- Logging: S3 folder: s3://aws-logs-837864701453-us-east-1/elasticmapreduce/
- Log encryption
- Debugging
- Termination protection

Tags:

Key	Value (optional)
Add a key to create a tag	

Additional Options:

- EMRFS consistent view
- Custom AMI ID: None

Bootstrap Actions:

At the bottom right, there are "Cancel", "Previous", and "Next" buttons. The "Next" button is highlighted in blue.

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

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair ⓘ

Cluster visible to all IAM users in account ⓘ

Permissions ⓘ

Default Custom

Select custom roles to tailor permissions for your cluster.

EMR role ⓘ

EC2 instance profile ⓘ

Auto Scaling role ⓘ

► Security Configuration
► EC2 security groups

ⓘ No EC2 key pair has been selected, so you will not be able to SSH to this cluster or connect to HUE (unless you are using a VPN). [Learn how to create an EC2 Key Pair.](#)

[Cancel](#) [Previous](#) [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.

The screenshot shows the AWS EMR Cluster Overview page for a cluster named "My cluster". The cluster is currently in the "Starting" state, which is highlighted in green. The page includes tabs for "Summary", "Application user interfaces", "Monitoring", "Hardware", "Configurations", "Events", "Steps", and "Bootstrap actions". The "Summary" tab is active. Key details shown include:

- ID:** j-Q2S1VU993QDY
- Creation date:** 2021-02-11 15:49 (UTC+2)
- Elapsed time:** 5 minutes
- After last step completes:** Cluster waits
- Termination protection:** On [Change](#)
- Tags:** -- [View All / Edit](#)
- Master public DNS:** ec2-54-246-247-154.eu-west-1.compute.amazonaws.com [Copy](#)
[Connect to the Master Node Using SSH](#)

The "Configuration details" section lists:

- Release label:** emr-5.32.0
- Hadoop distribution:** Amazon 2.10.1
- Applications:** Hive 2.3.7, Pig 0.17.0, Hue 4.8.0
- Log URI:** s3://aws-logs-144364850537-eu-west-1/elasticmapreduce/ [File](#)
- EMRFS consistent view:** Disabled
- Custom AMI ID:** --

The "Network and hardware" section shows:

- Availability zone:** eu-west-1a
- Subnet ID:** [subnet-16c78d4c](#)
- Master:** Bootstrapping 1 m5.xlarge
- Core:** Provisioning 2 m5.xlarge
- Task:** --

Two warning messages are displayed in callout boxes:

- Core - 2:** Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..
- Master - 1:** Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..

How to launch an EMR cluster

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

Cluster: My cluster Running Running step

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Add task instance group

Instance groups

Filter: Filter instance groups ... 2 instance groups (all loaded) C

ID	Status	Node type & name	Instance type
ig-KOD8UM2RQ9IE	Running	CORE Core - 2	m5.xlarge 4 vCore, 16 GiB memory EBS Storage: 64 GiB
ig-35KA9BFUMQJOO	Running	MASTER Master - 1	m5.xlarge 4 vCore, 16 GiB memory EBS Storage: 64 GiB

How to launch an EMR cluster

Click in the EC2 Instance ID Link

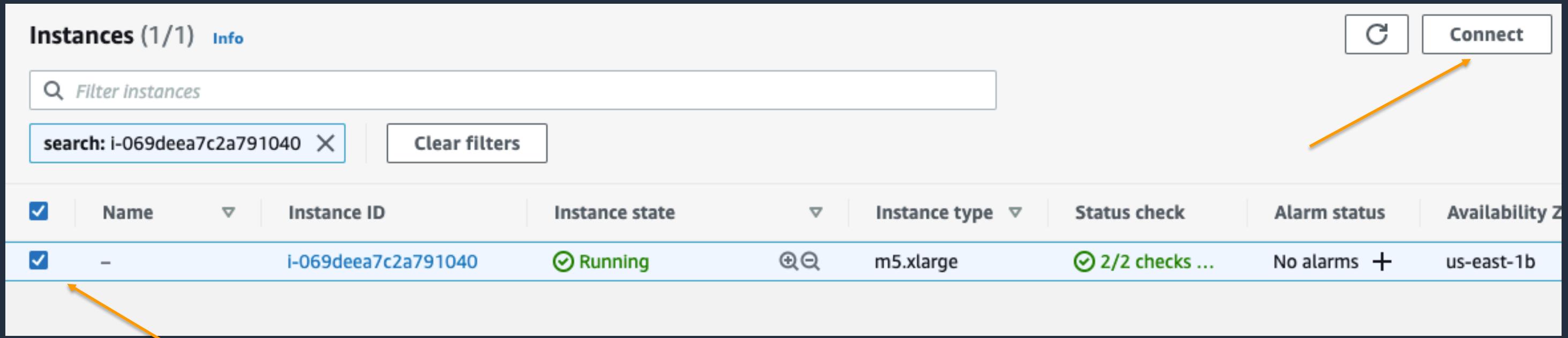
The screenshot shows the AWS EMR Cluster Instances page for a cluster named "My cluster". The status is "Running". The "Hardware" tab is selected. A blue button labeled "Add task instance group" is visible. The table below shows one instance:

ID	EC2 instance ID	EBS volumes per instance	Status	Pub...
ci-1VNT8JFM5IBX8	i-0ef1a6ccecb4334c6	2 View All	Running	ec2...

An orange arrow points to the EC2 instance ID link ([i-0ef1a6ccecb4334c6](#)).

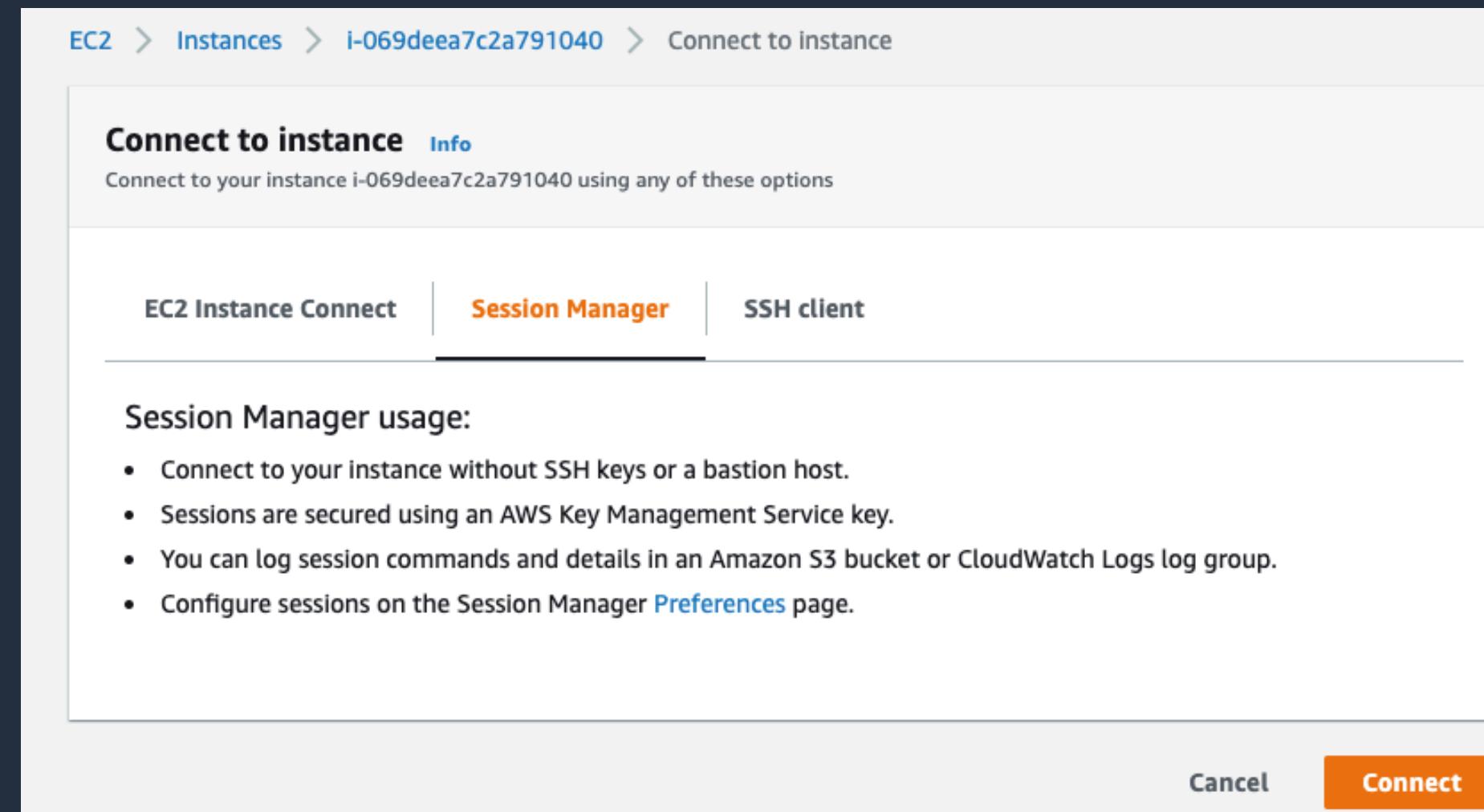
How to launch an EMR cluster

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



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

```
sh-4.2$ whoami
ssm-user
sh-4.2$ sudo su - hadoop
Last login: Thu Feb 11 10:05:37 UTC 2021 on pts/0

          EEEEEEEEEE      MMMMMMM   MBBBBBBB RRRRRRRRRRRRRR
          E:::::::::::E M:::::M   M:::::M R:::::R
          EE:::::EEEEE:::E M:::::M   M:::::M R:::::RRRRR:::R
            E:::E     EEEEE M:::::M   M:::::M RR:::R   R:::R
            E:::E           M:::::M:::M   M:::M::::M   R:::R
            E:::::EEEEE   M:::::M M:::M M:::M M:::::M   R::::RRRRR:::R
            E:::::::::::E M:::::M   M:::M:::M   M:::::M   R::::::::::RR
            E:::::EEEEE   M:::::M   M:::::M   M:::::M   R:::::RRRRR:::R
            E:::E         M:::::M   M:::M   M:::::M   R:::::R   R:::R
            E:::E         EEEEE M:::::M   M:::M   M:::::M   R:::::R
          EE:::::EEEEE:::E M:::::M           M:::M   R:::::R   R:::R
            E:::::::::::E M:::::M           M:::M RR:::R   R:::R
          EEEEEEEEEE      MMMMMMM   MBBBBBBB RRRRRRRR   RRRRRR

[hadoop@ip-172-31-33-10 ~]$ whoami
hadoop
[hadoop@ip-172-31-33-10 ~]$ █
```



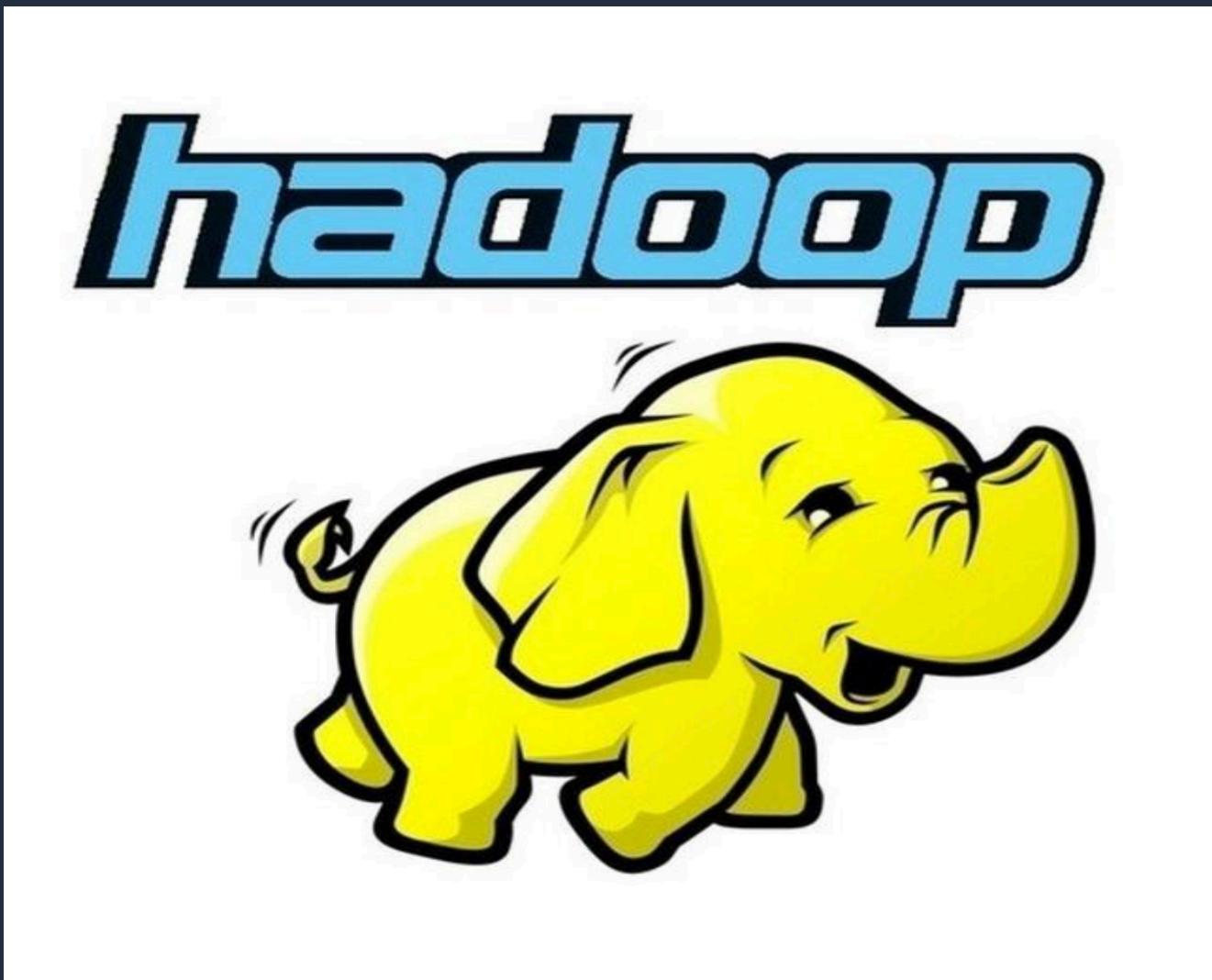


Introduction to Big Data

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

Introduction to Big Data



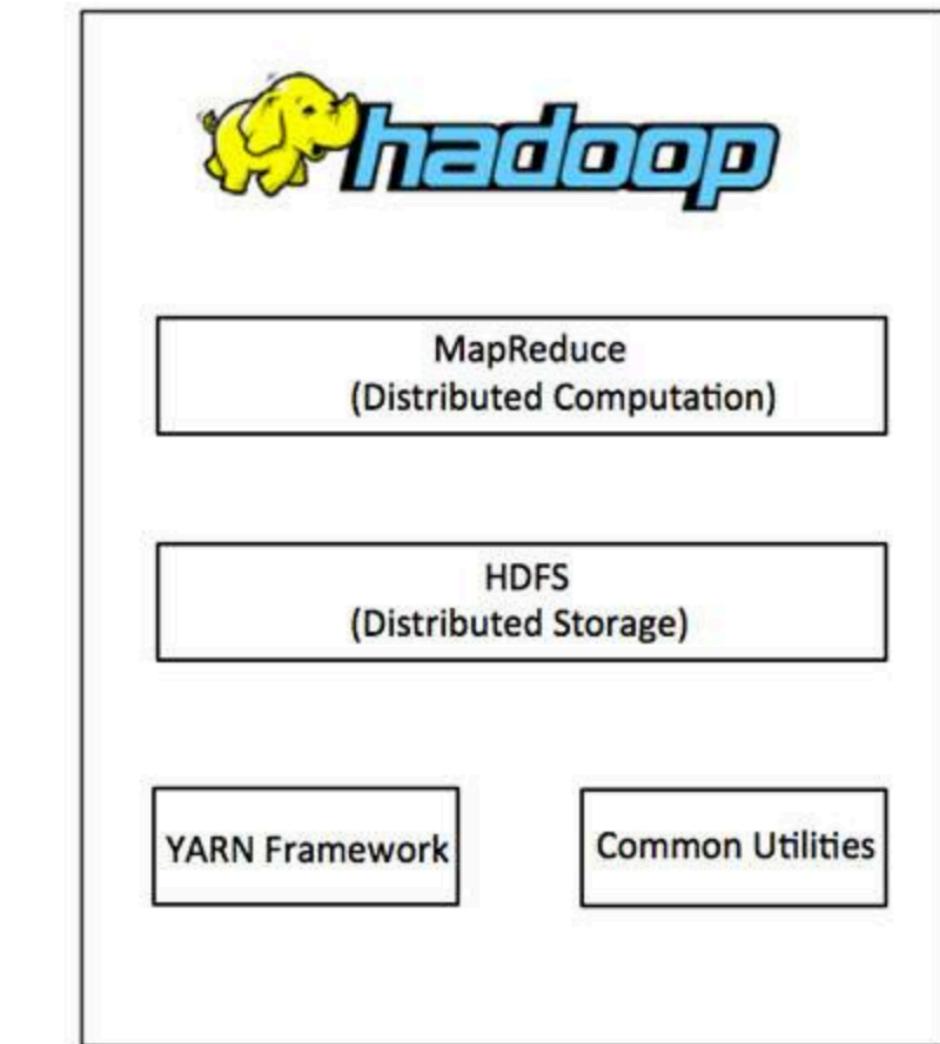
Introduction to Hadoop

- 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

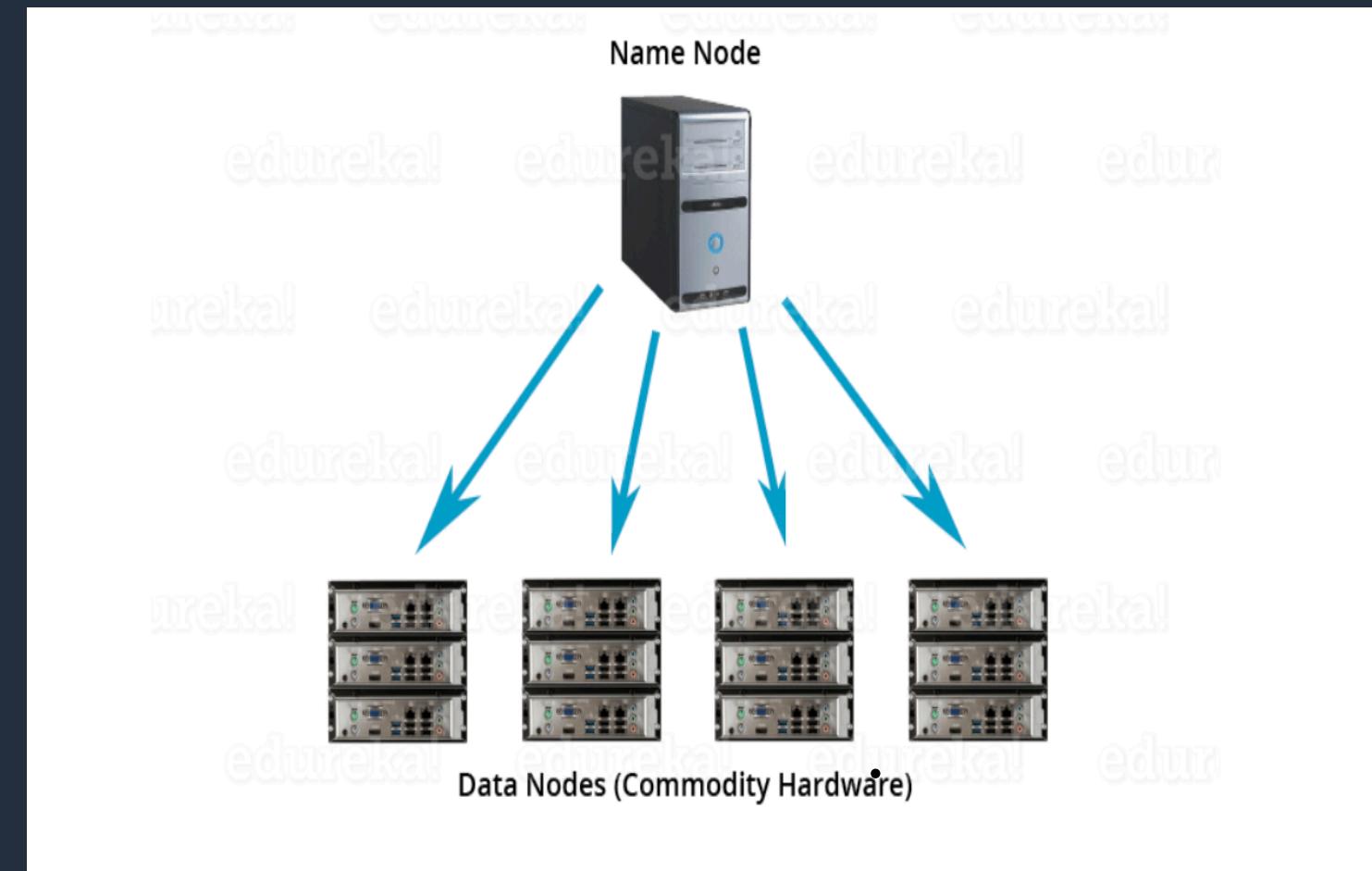
At its core, Hadoop has two major layers namely –

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



HDFS Storage on Disks

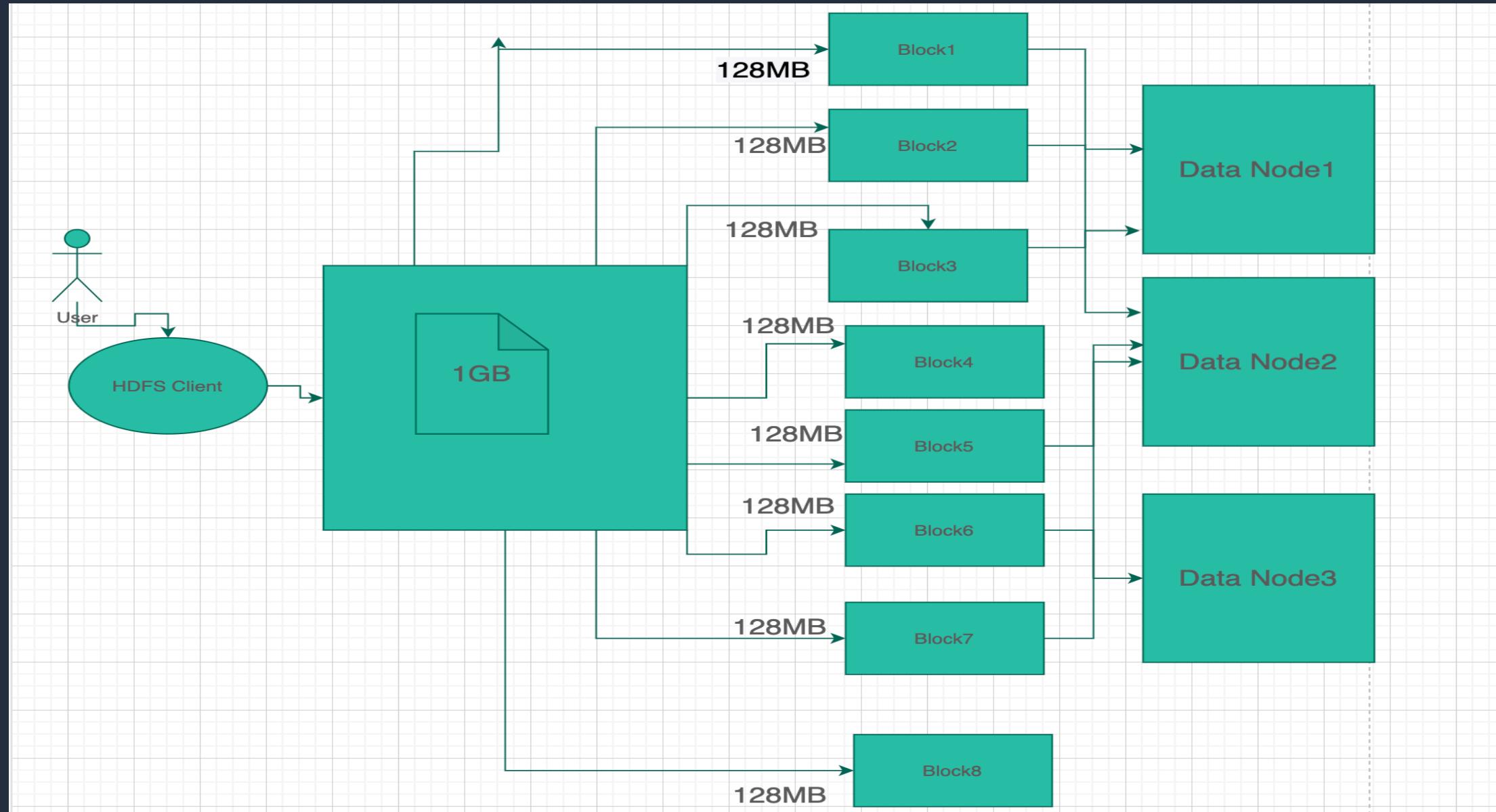
- Let's understand the example provided here. Imagine we have 4 machines or computers with a hard drive of 1TB on each machine.
- When you install Hadoop as a platform on top of these machines, you will get HDFS as a storage service.
- Hadoop distribution File System is distributed in a such a way that 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).



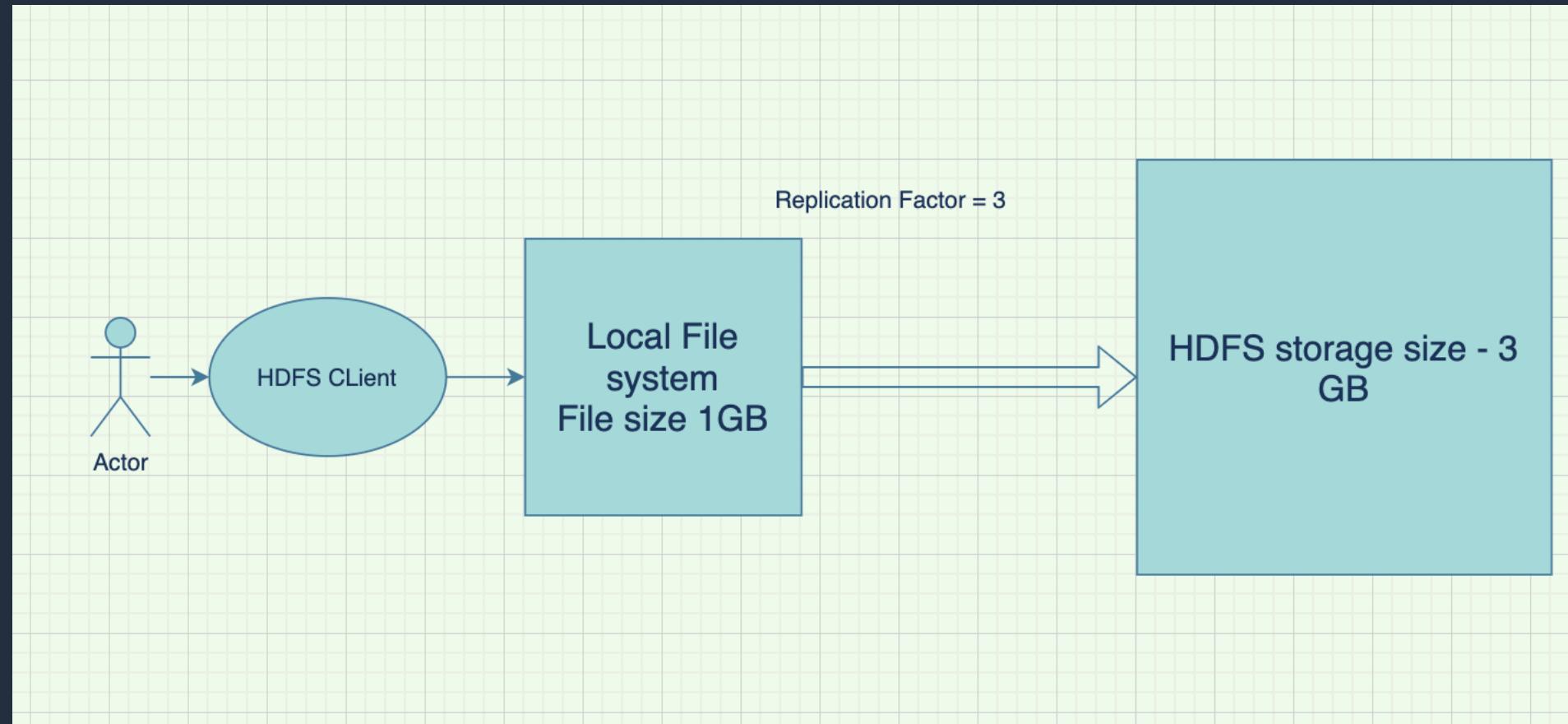
Name Node Vs Data Node

- 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.

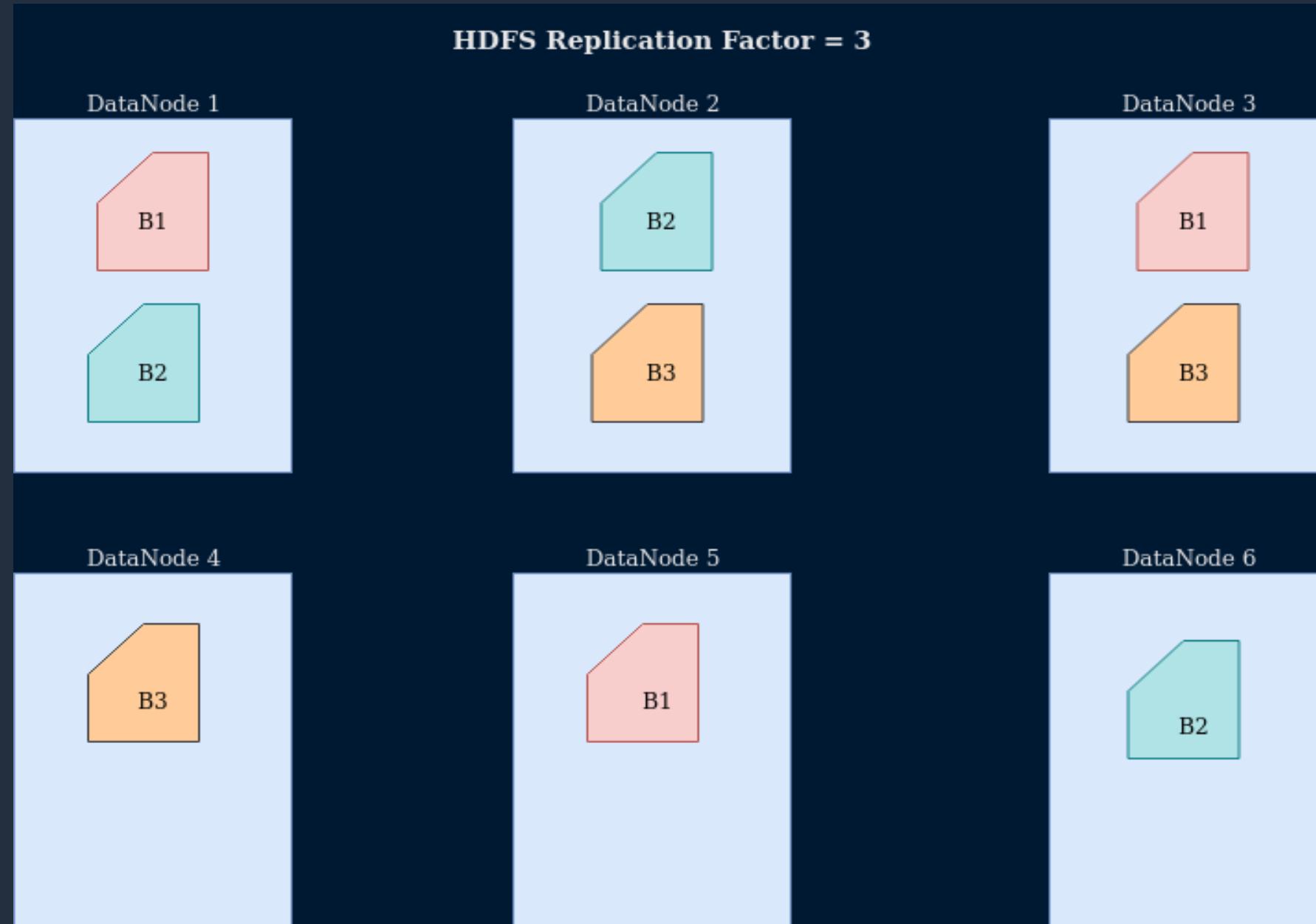
HDFS Blocks and Split



HDFS Replication Factor

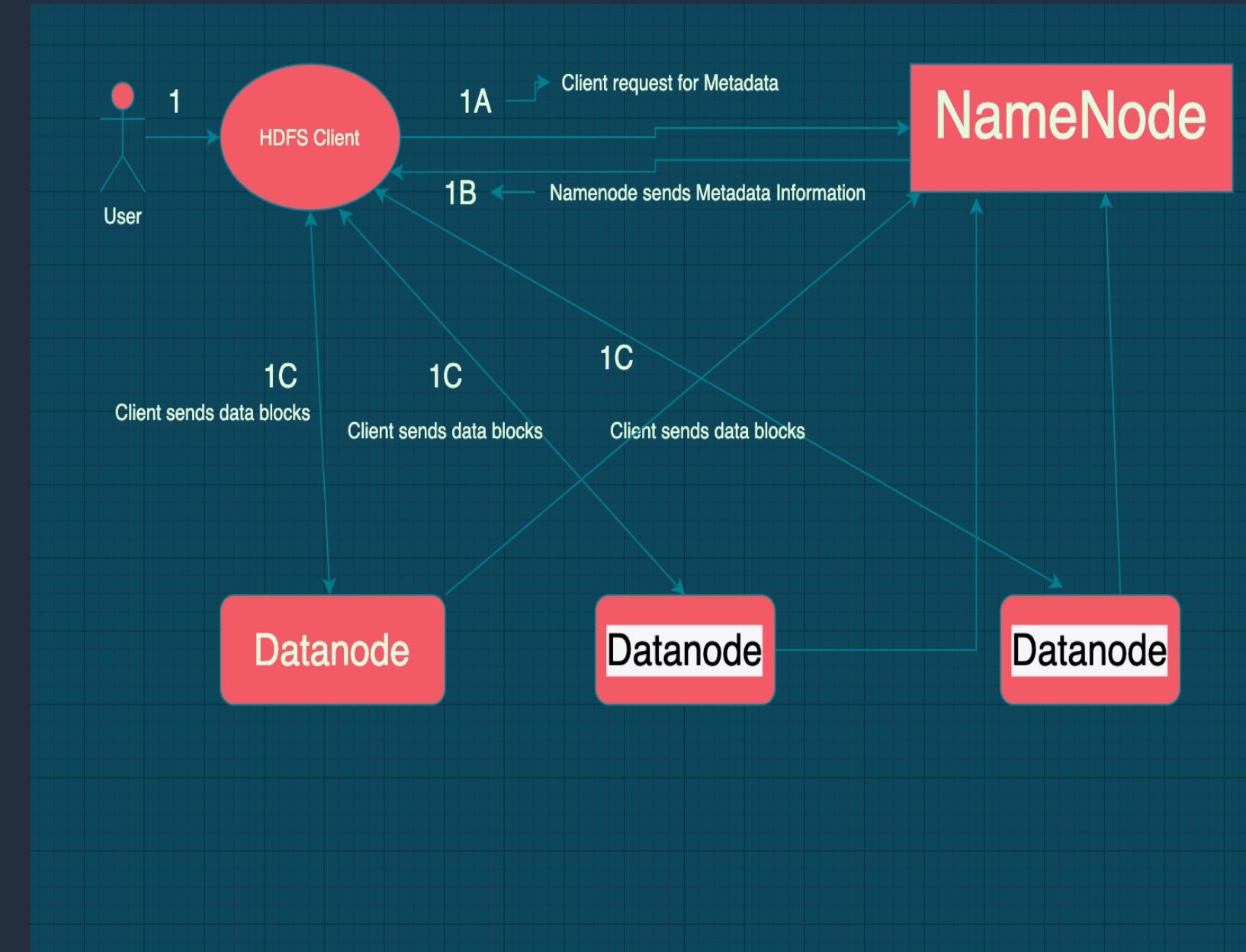


HDFS Replication Factor



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.



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

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?

Break

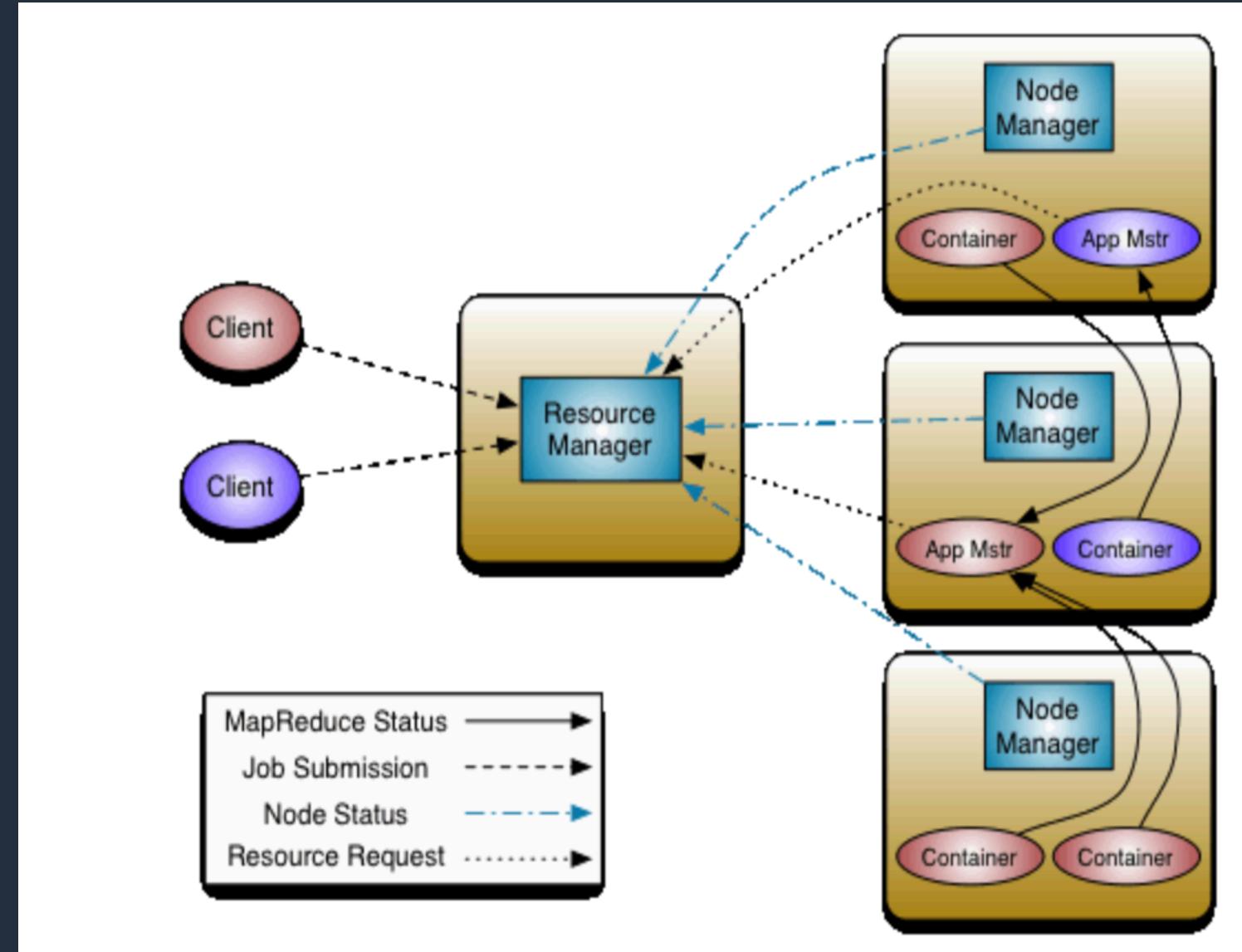
Will resume in 15 minutes

Break

Will resume in 15 minutes

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.



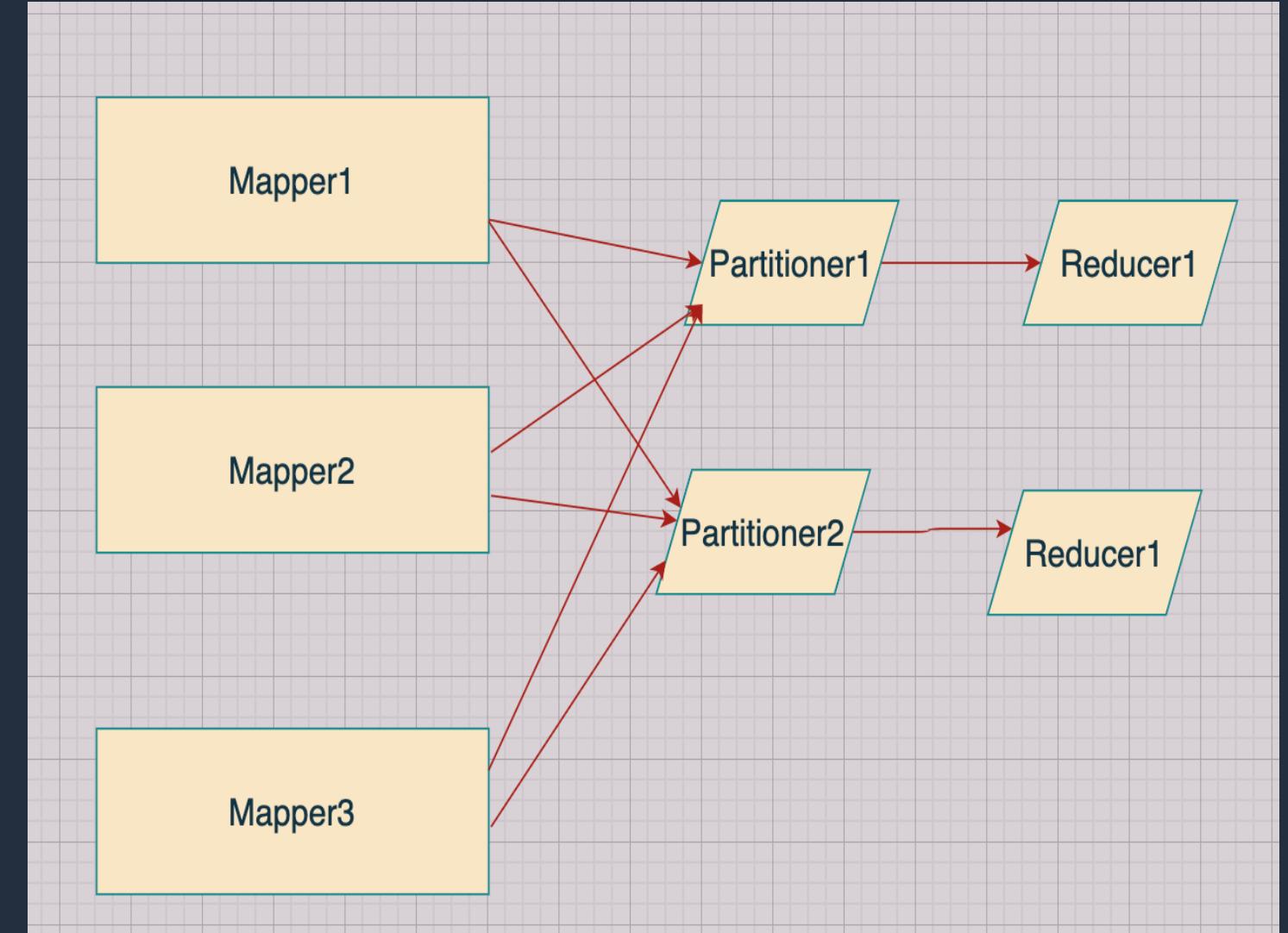
MapReduce

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.

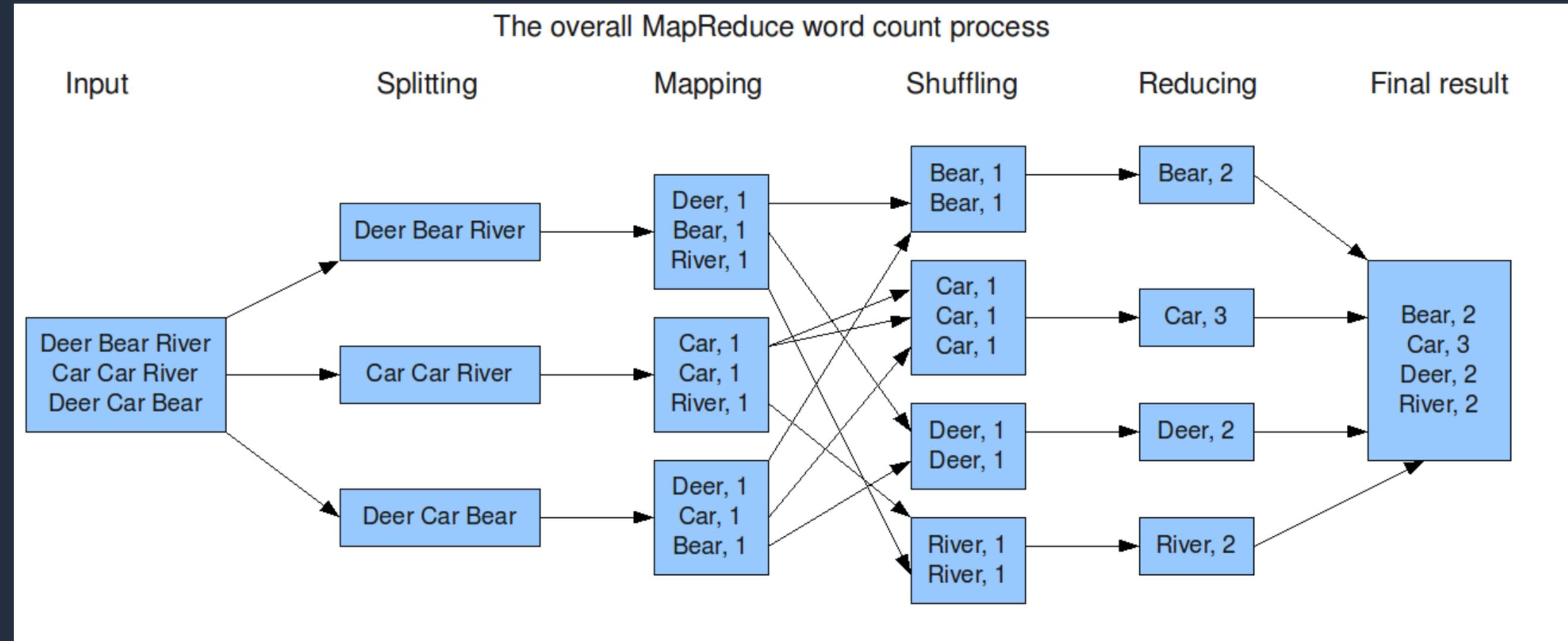


No. of Mapper = (total data size)/ (input split size)

Checks on Map Reduce

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

MapReduce Flow Diagram



Checks on Map Reduce

Q. True or False

Reducers can run before Map stage completion

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?

Checks on Map Reduce

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

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

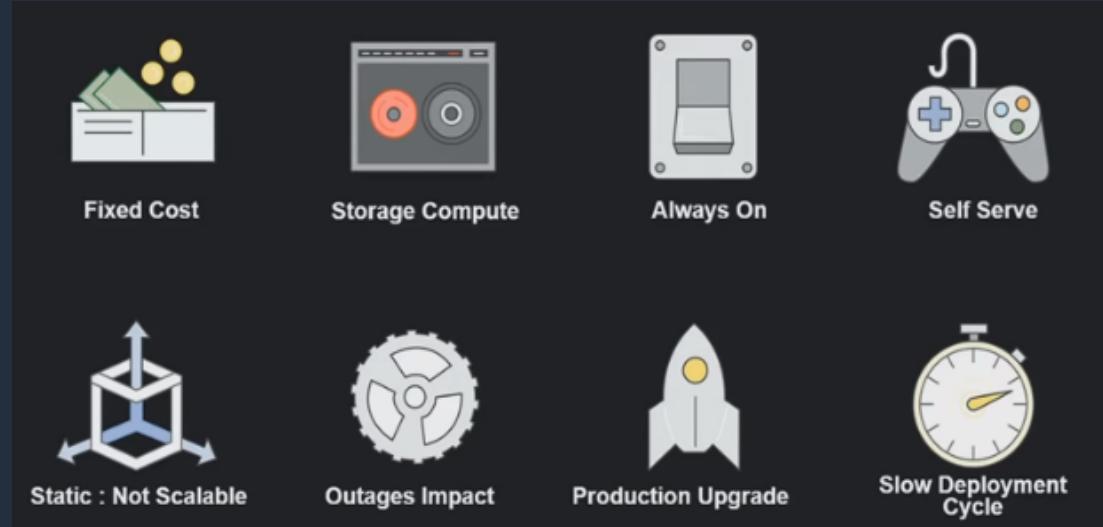
What is Amazon S3

- 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 99.99999999% (11 9's) of durability, and stores data for millions of applications for companies all around the world.

EMR and S3??

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



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.

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

Machine Learning

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

Data Movement

- Sqoop
- Relational DB connector

NoSQL

- HBase
- Hadoop's non-relational DB

Data Processing

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

Query Tools

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

Q&A



Happy cloud computing