**EMR 🡪 ELASTIC MAP REDUCES.**

* Cloud based big DATA processing service provided by AWS
* EMR – scale and run big servers on demand on the cloud.
* Pre Installed and pre configured with Spark, Hadoop, Hive, tensflow etc.,
* Use ec2 instance as worker node and s3 bucket as its file system.
* Can Integrate with other AWS services : kensis, Dynamo DB, etc.,
* It allows to easily provision and manages **Hadoop Clusters, Sparks Clusters** and other **Big Data** Processing frameworks on AWS infrastructure.
* Here we can Analyze and process the vast amounts of data using a variety of Big data Tools and Technologies
* Supports range of Big data Processing Engines Such as Hadoop, Apache Spark and Apache Hive as well as other variety of open source Big Data Tools.
* EMR is a flexible platform that can be customized to meet the specific needs of different types of data processing and analysis tasks.

**Best Practice:**

* **Choose the right instance type:**

Select the instance type that meets the requirements of your use case, workload, and budget. EMR offers a wide range of instance types, and you should choose the one that best fits your needs.

* **Choose the right number of instances:**

It's important to select the right number of instances to meet your performance and cost goals. You can start with a small number of instances and scale up or down as needed.

* **Use Spot instances:**

Spot instances are a cost-effective option for running EMR clusters. You can save up to 90% of the cost of on-demand instances by using Spot instances.

* **Use EMRFS:**

EMRFS (EMR File System) is a Hadoop-compatible file system that allows you to store data in Amazon S3. EMRFS is optimized for EMR and offers better performance and durability than HDFS (Hadoop Distributed File System) when used with S3.

* **Enable EMR cluster termination protection:**

EMR cluster termination protection prevents accidental termination of clusters. This can be particularly useful if you have long-running or critical clusters.

* **Use EMR release labels:**

EMR release labels are specific versions of EMR that are tested and recommended for different types of workloads. You should choose the release label that best fits your needs.

* **Monitor your clusters:**

It's important to monitor your EMR clusters to ensure that they are running as expected. You can use Amazon CloudWatch to monitor metrics such as CPU utilization, memory usage, and disk I/O.

* **Use encryption:**

EMR supports encryption of data in transit and at rest. You should use encryption to protect sensitive data.

* **Use EMR security configurations:**

EMR provides preconfigured security configurations that you can use to secure your clusters. You should choose the configuration that best fits your security requirements.

* **Use EMR step failures:**

EMR step failures can help you troubleshoot issues with your clusters. You should configure your clusters to send step failure notifications to a specified email address or SNS topic

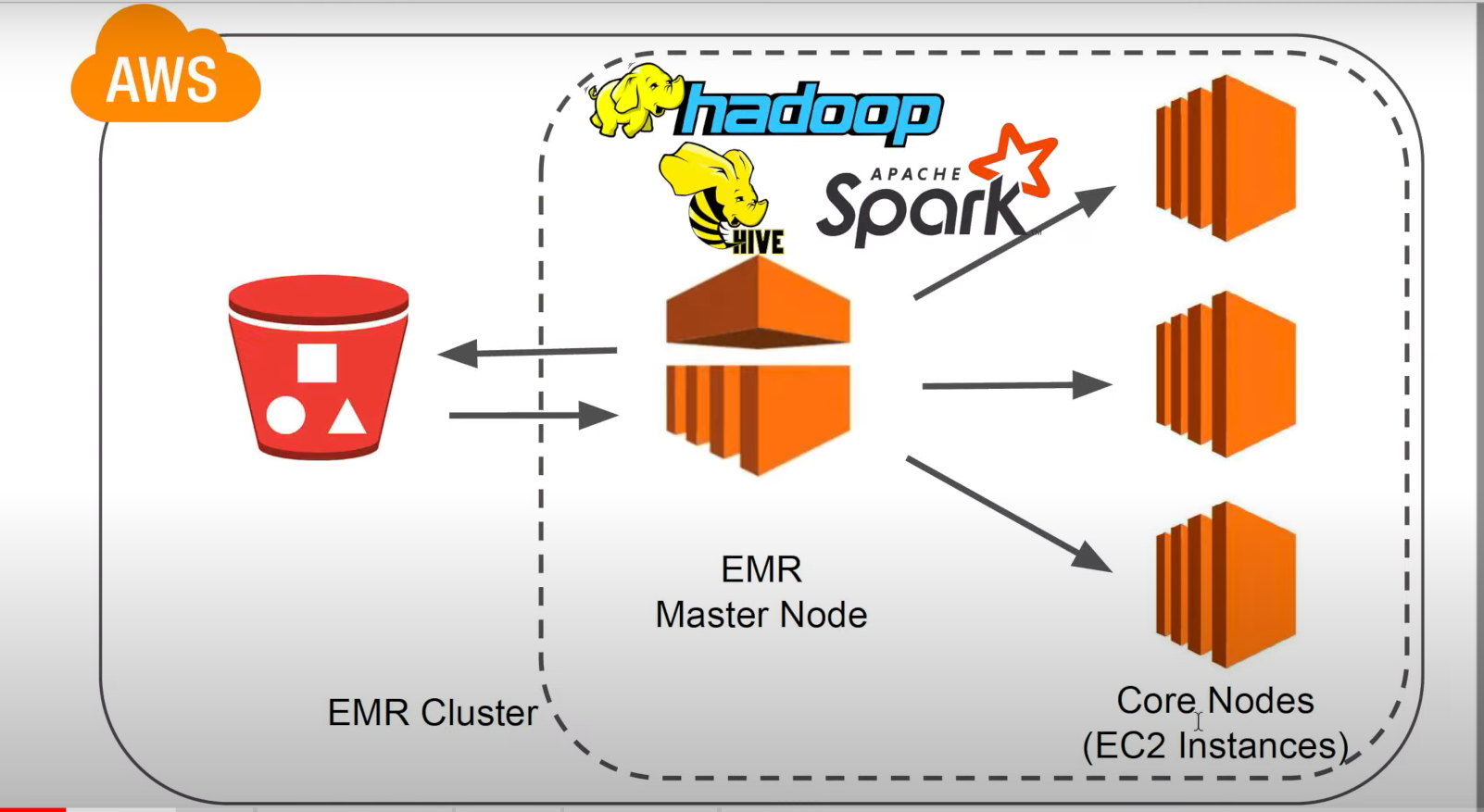
RELIABLE TO DEVOPS :

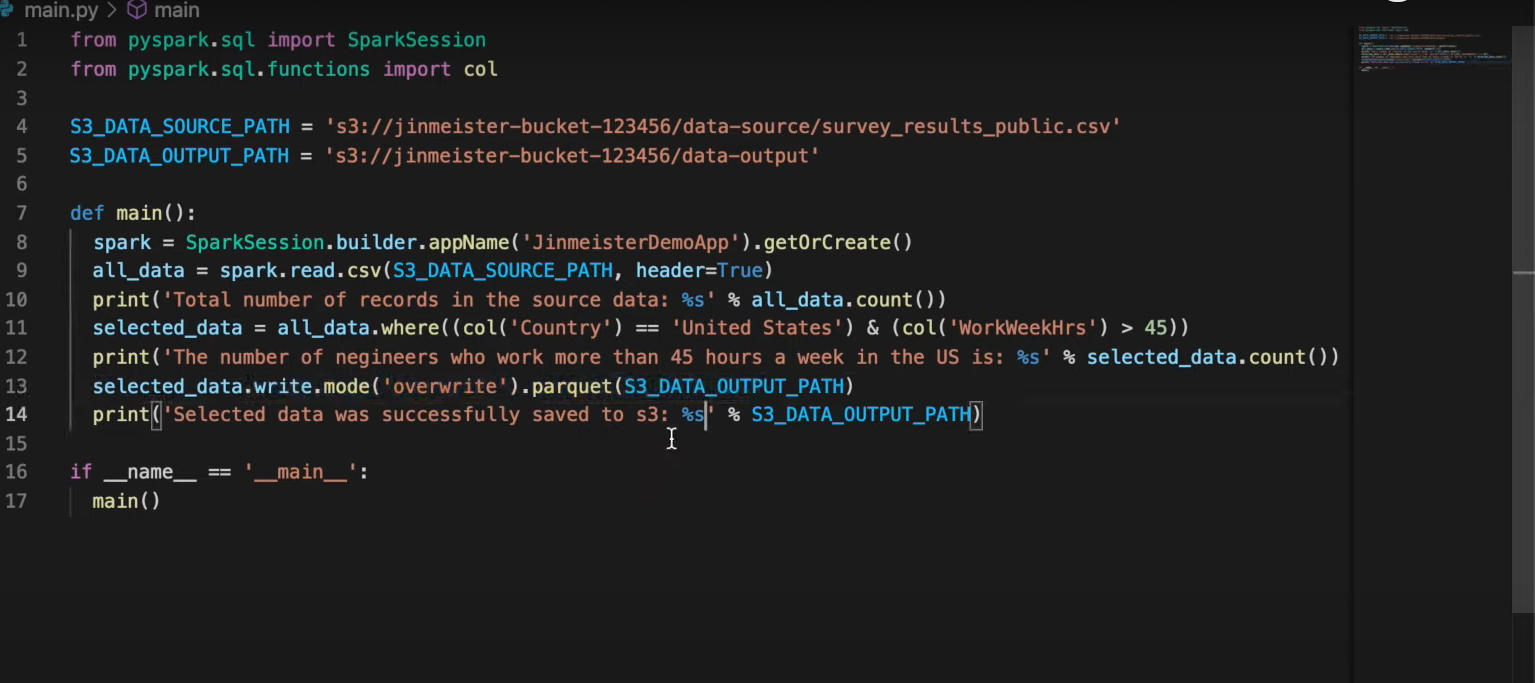
EMR (Elastic MapReduce) can be related to DevOps in several ways, as it is a tool that can help DevOps teams manage and automate big data processing and analysis workflows. Here are some ways in which EMR is related to DevOps:

* **Infrastructure as Code (IaC):** EMR can be managed using infrastructure as code tools such as AWS CloudFormation or AWS CDK. This allows DevOps teams to automate the deployment, scaling, and management of EMR clusters using code.
* **Continuous Integration and Continuous Delivery (CI/CD):** EMR can be integrated with CI/CD pipelines to automate the build, test, and deployment of big data processing and analysis workflows. This allows DevOps teams to streamline the release process and ensure that changes are deployed quickly and reliably.
* **Monitoring and Alerting:** EMR can be monitored using AWS CloudWatch, which provides metrics and alarms that can be used to monitor the health and performance of EMR clusters. This allows DevOps teams to proactively identify and address issues before they impact the system.
* **Security and Compliance:** EMR provides features such as encryption, access control, and compliance with industry standards such as HIPAA and SOC 2. This allows DevOps teams to ensure that their big data processing and analysis workflows meet security and compliance requirements.
* **Scalability and Resilience:** EMR is designed to be scalable and resilient, allowing it to handle large volumes of data and to recover from failures automatically. This helps DevOps teams to build and maintain robust big data processing and analysis workflows that can handle the demands of their organization.
* EMR can be a valuable tool for DevOps teams looking to manage and automate big data processing and analysis workflows. By using EMR alongside other DevOps tools and practices, teams can build and maintain scalable, resilient, and secure big data systems that meet the needs of their organization.

Getting Stated With EMR 🡪 Link : <https://www.youtube.com/watch?v=ulQdA6MUnTk>

**SAVED 16/03/2023**





### [AWS EMR Big Data Processing with Spark and Hadoop | Python, PySpark, Step by Step Instructions](https://www.youtube.com/watch?v=a_3Md9nV2Mk&t=512s)

Link : <https://www.youtube.com/watch?v=a_3Md9nV2Mk&t=315s> 🡪 YouTube link for EMR Spark data check using python@08:00.

**Configuration In AWS EMR Cluster:**

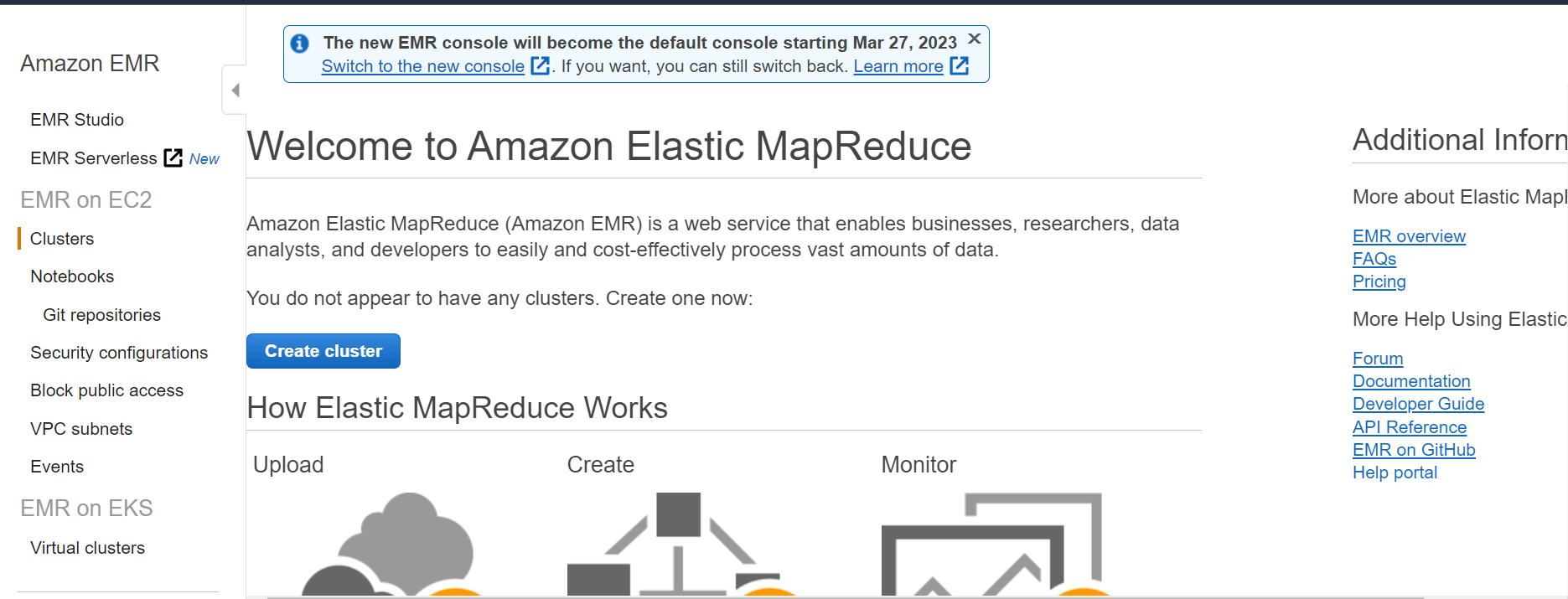
**Login to AWS Console :**

**Prequisite :**

VPC, Subnet (Minimum 2 in two different region), Security Group (allowing SSH connect and required Port numbers to be allowed ), Keypair (To launch the instance and running spark-shell in VM).

**BELOW STEPS TO BE FOLLOWED IN CREATING OF CLUSTER:**

Navigate to EMR 🡪 Create a cluster.  
choosing Advance Options for Configurations



**SOFTWARE CONFIGURATION :**

Here we configure Software requirements and Master Servers.

**🡪**Select the EMR(Elastic Map Reduce) Version  
Add Spark 2.4.5 along with [Hadoop 2.8.5,PG 0.17.0, Hive 2.3.6, Hive 4.6.0, ]

🡪Multiple master Node (If Required can be added, based on requirement)

🡪Edit Software Settings

To add the service of all the tools that are using for Hive and Spark.

🡪Optional setup of Master & Nodes

A step is a unit of work you submit to the cluster. For instance, a step might contain one or more Hadoop or Spark jobs. You can also submit additional steps to a cluster after it is runnin

* Clusters enter waiting mode.
* Terminate the Cluster if Not used for 1 hour.
* Cluster Auto terminates.

🡪Click on NEXT Button for Jumping into Hardware Configuration.

**HARDWARE CONFIGURATIONS:**

Here we Configure Network & the Compute Configurations of Cluster. Request Spot Instance to reduce cost of Expenses.

Types of Instances available in this C, D, G, H, I, M, P, R & Z.

🡪Instance Group Configuration   
 Uniform Instance group

Instance Fleet.

(Specify target Capacity and Amazon Fulfills it for each node type. Mix Instance type and purchase Options. )

🡪Networking

Select the Pre created VPC’s and Subnet if not created can Create here Option available.

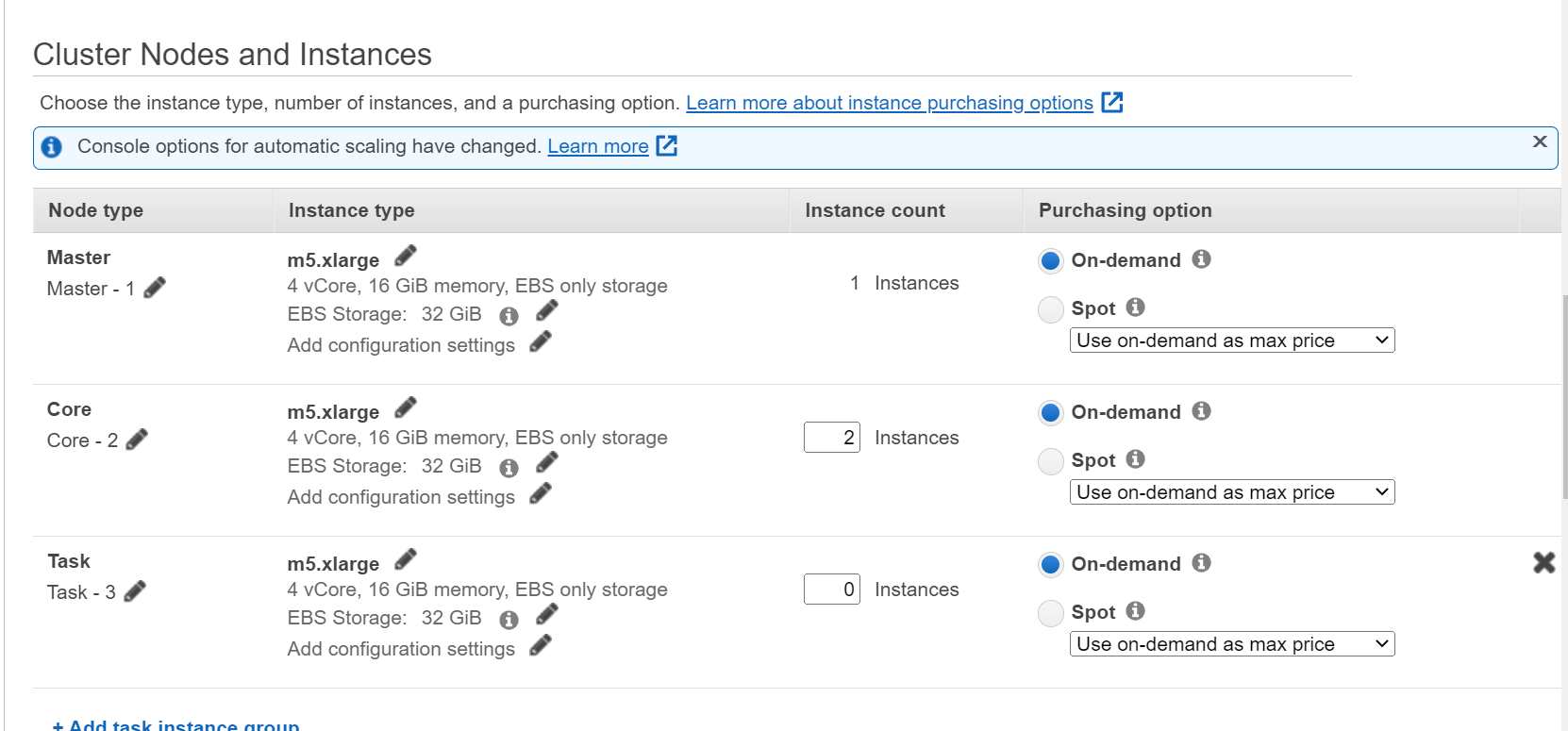
🡪Cluster Nodes and Instances

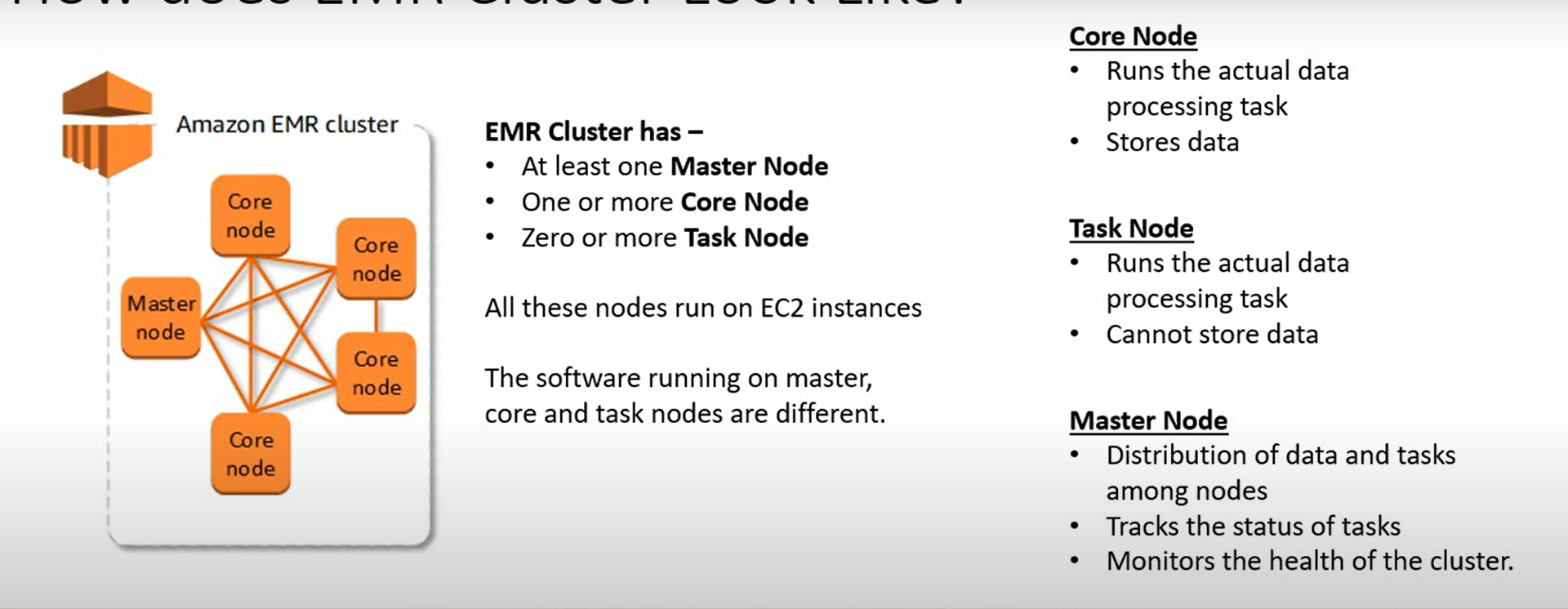
Here it gives the Option to create Master, Task & core

Can select the required Nodes configuraition.

Type of the instance and number of Instance can be defined here.

Here we can customize the purchasing option by selecting the **On-Demand** or **Spot.**

****

****

**🡪**Cluster scalling

If requied can enable Cluster scalling if not leave it by Default.

🡪Auto termination

Enabling Auto termination by giving the time in Hours and Minutes format.

🡪EBS Root Volume

Here we can increase the Elastic Block Storage Volume size upto 100gib.

🡪click on Next Button after completed.

**GENERAL CLUSTER SETTINGS:**

🡪Name : <Name\_of\_Cluster.>

🡪Looging -> enabling this for S3 Bucket (This will create a file in S3 Bucker with Logs.)

🡪 Log Encryption : Can Encrypt the logs stored in S3 Bucket by using KMS (Key Managed Service).

🡪Termination Protection: Prevents accidental Termination of Cluster. To shutdown you need to disable the Termination protection

TAGS : These are used if requied for identification.

ADDITIONAL OPTIONS:

🡪EMRFS (Elastic Map reduce)

OPERATING SYSTEM OPERATIONS:

🡪This creates a new Amazon Linux release that Amazon EMR validates.

🡪if any Custom AMI ID are available you can use them.

Update all installed packages on reboot (Recommedned.)

BOOTSTRAP ACTONS:

These are the scripts that are executed when the Hadoop starts on every cluster node. We can use them to install additional software and customize our applications.

**SECURITY OPTIONS**

EC2 KEY\_PAIR 🡪 Created Key\_pair to be configure here/ selected.

PERMISSIONS:

🡪DEFAULT : If Don’t have any customized permissions.

🡪CUSTOM : In case of customized permissions this to be enabled and add the permissions

🡪EMR Role : Default.

🡪Use EMR\_DefaultroleV2 : calls AWS other services on behalf of users./Ours.

SECURITY CONFIGURATION:

🡪Choosing security Configuration to specify authentication and encryption.

EC2 SECURITY GROUPS:

🡪MASTER NODE : Configuring Security groups by allowing required port.

🡪CORE AND TASK : Configuring Security groups by allowing required port.

* **CLICK ON CREATE A CLUSTER TO LAUNCH A CLUSTER IT WILL TAKE 5-7 MINS.**

**SAVED 17/03/2023**