**Understanding clusters and nodes**

 A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node. Each node has a role within the cluster, referred to as the node type. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

**The node types in Amazon EMR are as follows:**

* **Primary node**: A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes for processing. The primary node tracks the status of tasks and monitors the health of the cluster.
* **Core node**: A node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node.
* **Task node**: A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional.

A picture containing text, screenshot, font, diagram

Description automatically generated

**EBS root volume – *optional:-***

 EBS volumes are like virtual hard disks that are attached to the server for storage.

**Spot purchasing option:-**

Spot Instances in Amazon EMR provide an option for you to purchase Amazon EC2 instance capacity at a reduced cost as compared to On-Demand purchasing.

# Bootstrap actions:-

Bootstrap action to install additional software or customize the configuration of cluster instances. Bootstrap actions are scripts that run on cluster after Amazon EMR launches the instance using the Amazon Linux Amazon Machine Image (AMI).

EXAMPLE:-copies a file, myfile.jar, from Amazon S3 to a local folder, /mnt1/myfolder, on each cluster node. The script is saved to Amazon S3 with the file name copymyfile.sh with the following contents.