**Elastic MapReduce (EMR)**

1. A platform that simplifies running big data frameworks such as Apache Hadoop & Spark on AWS to process & analyze large amount of data.
2. It distributes computation of the data over multiple Amazon EC2 instances.
3. You can use EMR to transform and move large amounts of data into and out of other AWS data stores and databases.

**Components:**

**1) Clusters:** A collection of EC2 Instances. You can create two types of clusters.

**a)** **Transient cluster** that auto-terminates after steps complete

**b) Long-running** cluster that continues to run until you terminate it.

**2) Nodes:** Each EC2 instance in a cluster is called a node.

**3) Node Types:** Each node as a role within the cluster referred to as node type:

**a) Master 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 master node tracks the status of the tasks and monitors the health of clusters.
* Every cluster has a master node and it’s possible to create a single-node cluster but it won’t support automatic failover.

**b) Core Node:**

* A node with software components that run tasks and store data in HDFS on your cluster. Multi-node cluster have atleast one core node.
* EMR is fault tolerant for slave failures and continues job execution if a slave node goes down.
* No need of more than 2 instances, as most of the data is expected to be on AWS S3.
* Replication factor is relevant to HDFS, not to S3.

**c) Task Node:**

* A node with software components that only run tasks and does not store data in HDFS. Task nodes are optional.
* Use task node for Auto Scaling, as it has no HDFS local storage, and it is used only to add more compute power dynamically.

**Data Processing:**

**1) Submitting jobs directly to an application:** Submit jobs and interact directly with the software that is installed in your EMR cluster. To do this, you connect to the master node over a secure connection and access the interfaces & tools that are available for the software that runs directly on your cluster.

**2) Running Steps to Process Data:** Submit one or more ordered steps to an EMR cluster. Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

**Scaling:** There are two main options for adding or removing capacity.

**1) Deploy Multiple Clusters:** If you need more capacity you can easily launch a new cluster and terminate it when you no longer need it. There is no limit to how many clusters you can have.

**2) Resize a running cluster:** You may want to scale out a cluster to temporarily add more processing power to the cluster.

**EMR Notebook:**

* A Serverless Jupyter notebook.
* An EMR cluster is required to execute the code and queries within an EMR notebook, but the notebook is not locked to the cluster.
* Runs Apache Spark.

**Pricing:**

* You pay a per-second rate for every second for each node you use with a **one-minute** min.
* EMR price is in addition to the EC2 price (the price for underlying servers) and EBS price (if attaching EBS volumes).