**Amazon Elastic MapReduce (Amazon EMR)**

fully managed, ondemand Hadoop framework.

Spin up large Hadoop clusters instantly and start processing within minutes

Storage types:

**Hadoop Distributed File System**

**EMR File System (EMRFS) -** implementation of HDFS that allows clusters to store data on Amazon S3.

**Use Cases**

**Log Processing – meaningful insights**

**Clickstream Analysis –**

analyze *clickstream* data in order to segment users and understand user preferences

Advertisers can also analyze clickstreams and advertising impression logs to deliver more effective ads.

**Genomics and Life Sciences**

Amazon EMR can be used to process vast amounts of genomic data and other large scientific datasets quickly and efficiently

**Amazon EMR Architecture**

## Storage

There are several different types of storage options as follows.

### Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is a distributed, scalable file system for Hadoop. HDFS distributes the data it stores across instances in the cluster, storing multiple copies of data on different instances to ensure that no data is lost if an individual instance fails.

HDFS is ephemeral storage that is reclaimed when you terminate a cluster. HDFS is useful for caching intermediate results during MapReduce processing or for workloads that have significant random I/O.

### EMR File System (EMRFS)

Using the EMR File System (EMRFS), Amazon EMR extends Hadoop to add the ability to directly access data stored in Amazon S3 as if it were a file system like HDFS.

Most often, Amazon S3 is used to store input and output data and intermediate results are stored in HDFS.

## Cluster Resource Management

The resource management layer is responsible for managing cluster resources and scheduling the jobs for processing data.

Amazon EMR uses YARN (Yet Another Resource Negotiator)

## Data Processing Frameworks

The data processing framework layer is the engine used to process and analyze data.

Different frameworks are available for different kinds of processing needs, such as batch, interactive, in-memory, streaming, and so on

### Hadoop MapReduce - Bacth

### Apache Spark

Spark is a cluster framework and programming model for processing big data workloads

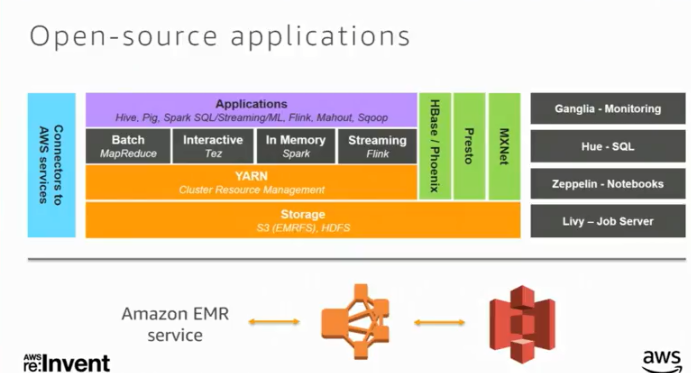
in-memory caching for datasets.

## Applications and Programs

Amazon EMR supports many applications, such as Hive, Pig, and the Spark Streaming library to provide capabilities such as using higher-level languages to create processing workloads, leveraging machine learning algorithms, making stream processing applications, and building data warehouses

you can use Java, Hive, or Pig with MapReduce or Spark Streaming, Spark SQL, MLlib, and GraphX with Spark.



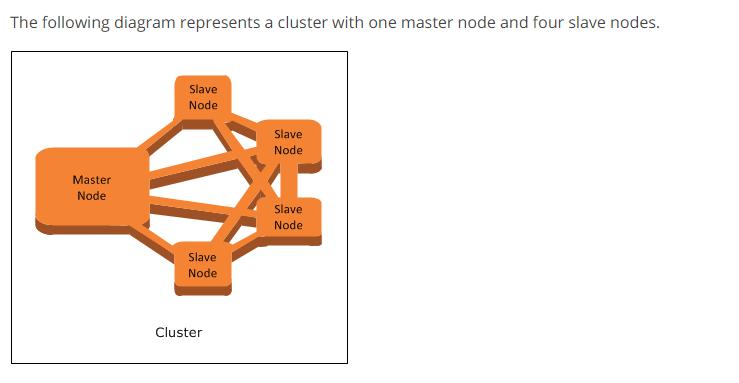


## Understanding Clusters and Nodes

The central component of Amazon EMR is the cluster. 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:

* **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 tasks and monitors the health of the cluster. Every cluster has a master node, and it's possible to create a single-node cluster with only the master node.
* **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.



Thanks & Regards,

Shantaram Vernekar