
aws

AWS EMR

Grow Data Skills

AWS EMR (Elastic MapReduce) is a cloud big data platform for processing massive amounts of data using open-source tools such as Apache Hadoop, Apache Spark, HBase, Presto, and Flink, among others.

Key Features and Properties of AWS EMR:

- **Scalability**: Automatically scales the number of instances up or down to meet the workload demands.
- **Cost-Effectiveness**: Offers the option to bid for spare Amazon EC2 computing capacity at reduced rates with Spot Instances.
- **Flexibility**: Supports multiple big data frameworks and allows you to run custom applications and code.
- **Data Processing Capabilities**: Capable of processing data from diverse AWS data sources, including Amazon S3, Amazon RDS, and Amazon DynamoDB.
- **Security**: Provides multiple layers of security, including AWS Identity and Access Management (IAM), network isolation using Amazon VPC, data encryption at rest and in transit using AWS Key Management Service (KMS).

AWS EMR

Grow Data Skills

Components and Tools:

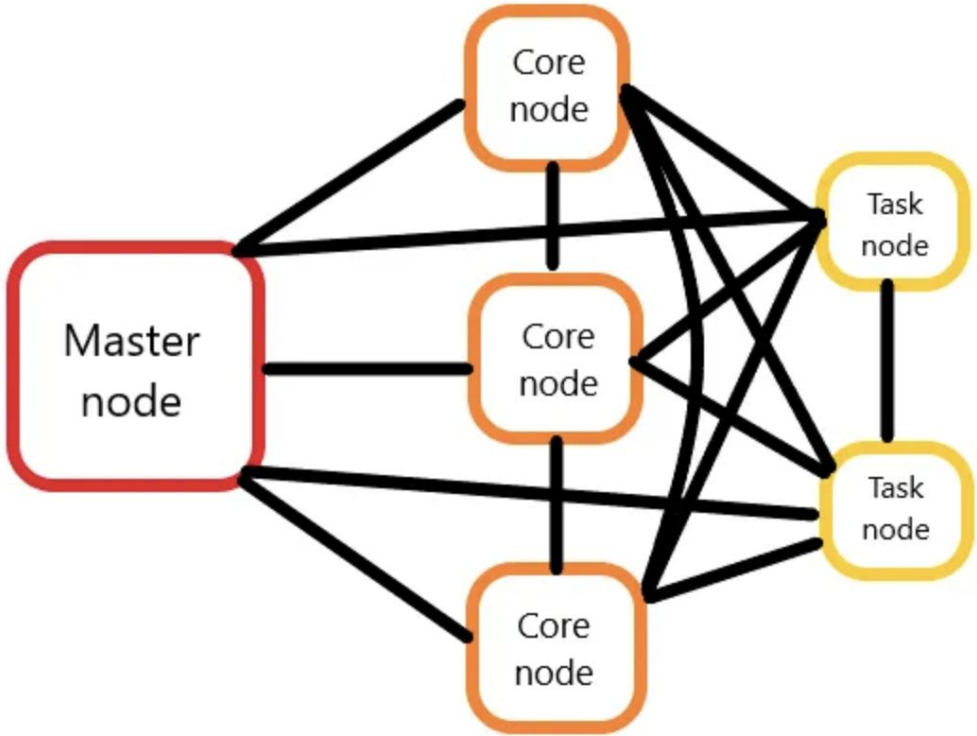
- **Hadoop**: A framework that allows for the distributed processing of large data sets across clusters of computers.
- **Spark**: An engine for large-scale data processing that is faster than Hadoop for certain applications.
- **HBase**: A non-relational, distributed database that runs on top of HDFS, providing real-time read/write access to large datasets.
- **Presto and Hive**: Tools for querying and managing large datasets residing in distributed storage.
- **Flink**: A framework and distributed processing engine for stateful computations over unbounded and bounded data streams.

Advantages of Using AWS EMR:

- **Reduced operational overhead**: AWS handles provisioning, configuration, and tuning of Hadoop clusters.
- **Integration with AWS ecosystem**: Seamless integration with AWS storage, database, and analytics services.
- **Cost savings**: Pay-as-you-go pricing model ensures you only pay for what you use, with additional savings from Spot Instances and Reserved Instances.

AWS EMR Architecture

Grow Data Skills



AWS EMR Architecture

Grow Data Skills

The node types in Amazon EMR are as follows:

- **Master Node:** It manages the clusters, can be referred to as Primary node or Leader Node.
    - It manages the cluster resources. It essentially coordinates the distribution of the parallel execution for the various Map-Reduce tasks. We can think about it as the leader that's handing out tasks to its various employees.
    - It tracks and directs the HDFS. Therefore, the master node knows the way to lookup files and tracks the info that runs on the core nodes.
    - With 5.23.0+ versions we have the ability to select three master nodes. Multiple master nodes are for mitigating the risk of a single point of failure. So, if one master node fails, the cluster uses the other two master nodes to run without any interruptions and what EMR does is automatically replaces the master node and provisions it with any configurations or bootstrap actions that need to happen.
    - The master node is also responsible for the YARN resource management. Its job is to centrally manage the cluster resources for multiple data processing frameworks. So, it's the master node's job to allocate to manage all of these data processing frameworks that the cluster uses.
    - It also performs monitoring and health on the core and task nodes. So, its job is to make sure that the status of the jobs that are submitted should be in good health, and that the core and tasks nodes are up and running.

AWS EMR Architecture

Grow Data Skills

- **Core Nodes:** It hosts HDFS data and runs tasks
    - They run tasks for the primary node. So, the primary node manages all of the tasks that need to be run on the core nodes and these can be things like Map Reduce tasks, Hive scripts, or Spark applications.
    - The core node is also responsible for coordinating data storage. So, it knows about all of the data that's stored on the EMR cluster and it runs the data node Daemon. This means that it breaks apart all of the files within the HDFS file system into blocks and distributes that across the core nodes
    - We know that we can have multiple core nodes, but we can only have one core instance group and we'll talk more about what instance groups are or what instance fleets are and just a little while, but just remember, and just keep it in your brain and you can have multiple core nodes, but you can only have one core instance group.
- **Task Nodes:** Runs tasks, but doesn't host data
    - These nodes are optional helpers, meaning that you don't have to actually spin up any tasks nodes whenever you spin up your EMR cluster, or whenever you run your EMR jobs, they're optional and they can be used to provide parallel computing power for tasks like Map-Reduce jobs or spark applications or the other job that you simply might run on your EMR cluster.
    - It does not store any data in HDFS. So there is no risk of data loss on removing. It's not used as a data store and doesn't run data Node Daemon.
    - They are often added or removed on the fly from the cluster. So this will help scale up any extra CPU or memory for compute-intensive applications.
    - It can cut down the all-over cost in an effective way if we choose spot instances for extra processing.