**EMR**

* EMR is a web service that enables businesses, researchers, data analysts, and developers to easily create cost-effective processes for vast amounts of data
* EMR utilizes a hosted Hadoop framework running on EC2 and S3
* Hadoop frameowrk for processing huge amounts of data
* Also support Spark, Hbase, Presto and Flink
* Most commonly used for log analysis, financial analysis, or extract, translate and loading (ETL) activities
* A Step is a programmatic task for performing some process on the data(e.g. count words)
* A cluster is a collection of EC2 instances provisioined by EMR to run your Steps
* EMR uses Apache Hadoop as its distributed data processing engine which is an open source, Java software framework that supports data-intensive distributed applications running on large clusters of commodity hardware
* EMR is a good place to deploy Spark, an open-source distributed processing used for big data workloads which utilizes in-memory caching and optimized query execution
* You can also launch Presto clusters. Presto is an open-source distributed SQL query engine designed for fast analytic queries against large datasets
* EMR launches all nodes for a given cluster in the same EC2 AZ
* You can access EMR by using the Managemetn Console, Command Line Tools, SDKs or the EMR API
* With EMR you have access to the underlying OS(you can SSH in)