Amazon EMR

is a web service that enables businesses, researchers, data analysts, and developers to process vast amounts of data.

EMR utilizes a hosted Hadoop framework running on Amazon EC2 and Amazon S3.

With EMR you can run petabyte-scale analysis at less than half of the cost of traditional on-premises solutions and over 3x faster than standard Apache Spark.

You can run workloads on Amazon EC2 instances, on Amazon Elastic Kubernetes Service (EKS) clusters, or on-premises using EMR on AWS Outposts.

Runs in one Availability Zone within an Amazon VPC.

Supports Apache Spark, HBase, Presto and Flink.

Most 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 provisioned by EMR to run your Steps.

EMR is a good place to deploy Apache 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 Amazon EC2 Availability Zone.

You can access Amazon EMR by using the AWS Management Console, Command Line Tools, SDKs, or the EMR API.

With EMR you have access to the underlying operating system (you can SSH in).

<https://digitalcloud.training/aws-organizations/>

<https://digitalcloud.training/amazon-cloudfront/>

<https://digitalcloud.training/aws-sysops-administrator-exam-scenarios/>