

Hadoop Services on Amazon – EMR

EMR (Elastic MapReduce)

It is used in a variety of applications, including log analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics.

Amazon Elastic MapReduce (EMR) is an Amazon Web Services (AWS) tool for big data processing and analysis. Amazon EMR offers the expandable low-configuration service as an easier alternative to running in-house cluster computing.

Amazon EMR is based on Apache Hadoop, a Java-based programming framework that supports the processing of large data sets in a distributed computing environment. MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. It was developed at Google for indexing web pages and replaced their original indexing algorithms and heuristics in 2004.

Amazon EMR processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). The elastic in EMR's name refers to its dynamic resizing ability, which allows it to ramp up or reduce resource use depending on the demand at any given time.

Amazon EMR can be classified as a tool in the **"Big Data as a Service"** category, while Hadoop is grouped under "Databases".

"On demand processing power" is the top reason why over **13** developers like Amazon EMR, while over **34** developers mention **"Great ecosystem"** as the leading cause for choosing Hadoop.

Pros of Amazon EMR

- On demand processing power
- Don't need to maintain Hadoop Cluster yourself
- Hadoop Tools
- Elastic
- Backed by Amazon
- Flexible
- Economic - pay as you go, easy to use CLI and SDKs
- Don't need a dedicated Ops group

- Massive data handling

Cons

- Cost overhead is a bit high