Handling and Analyzing Data with AWS Elastic MapReduce

CONFIGURING ELASTIC MAPREDUCE IN A PIPELINE

Andru Estes
CLOUD SOLUTIONS ARCHITECT

@andru_estes in/robertandruestes

Who Is This for and Why?

Who are you?

- Data scientist or engineer
- Solutions architect
- Cloud engineer
- Someone with big data knowledge

Why take this course?

- You want a deeper experience with AWS Elastic MapReduce
- You desire to learn how to utilize big data practices as they relate to AWS

Module Overview

Several learning objectives:

- Understand EMR and its components
- Provision and configure a Hadoop cluster
- Connect to S3 via EMRFS
- Load data into a Hadoop Distributed File System

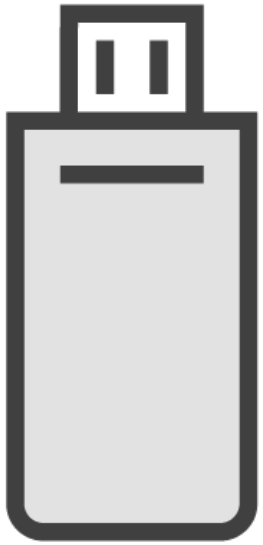
What is AWS EMR?

Managed cluster platform used to simplify using big data frameworks to process and analyze sizeable amounts of data

Can be used to move or transform data in and out of AWS databases and data stores, including S3 and DynamoDB

Uses open-source projects with the provided big data frameworks, such as Apache Hadoop and Apache Hive

EMR Architecture Layers



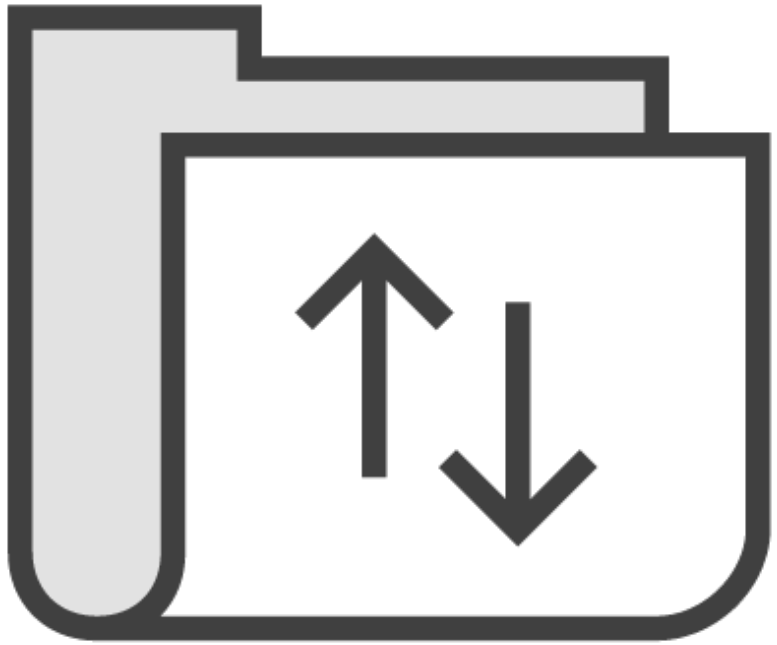
Storage

Cluster Resource Management

Data Processing Frameworks

Applications and Programs

Get to Know EMRFS



AWS implementation of HDFS (Hadoop Distributed File System)

Used by all EMR clusters for reading and writing to S3

Stores persistent data in S3 that is usable by Hadoop

Provides “consistent view”

Enables use of data encryption

Module Summary

What AWS EMR is and how it is used

The architectural components that make up AWS EMR

How to set up our own EMR cluster

What EMRFS is, how it is works, and how we can use it for our needs