Adding Machine Learning to the Pipeline

Andru Estes
CLOUD SOLUTIONS ARCHITECT

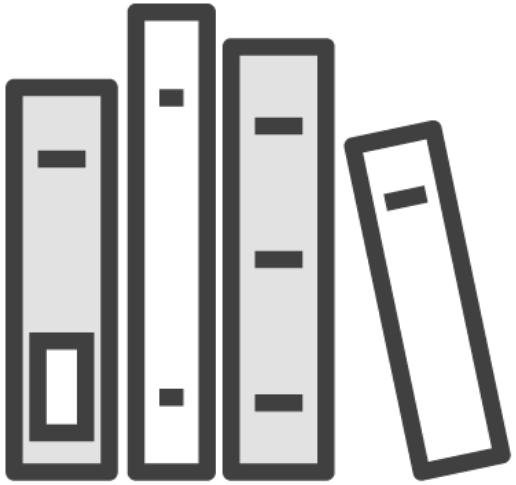
@andru_estes

Module Overview

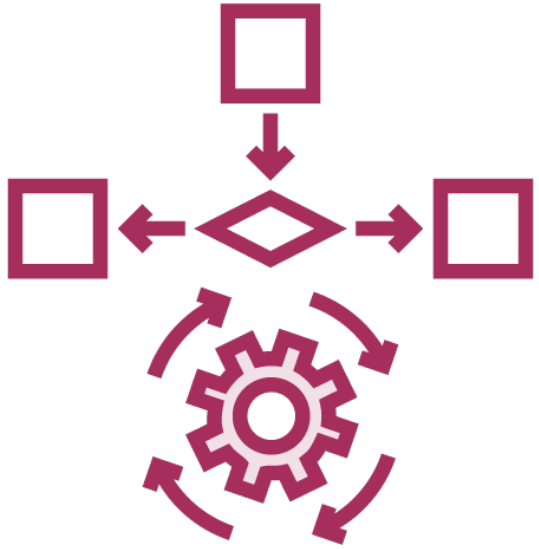
Learning objectives

- Use MLlib to perform machine learning tasks

What is MLlib?



Spark machine learning library



Algorithms, pipelines, persistence, featurization, and utilities

Diverse open-source community

More efficient than MapReduce

Demo

Configuring and using Spark and MLlib in EMR

Module Summary

Overview of what MLlib is

How MLlib is used

Configuring and using MLlib with EMR

Course Summary

Discussed and explored AWS Elastic MapReduce

Processed data using Apache Hive, Redshift, HBase, and Presto

Streamed data with Apache Flink and Spark

Utilized MLlib to optimize our machine learning tasks