Streaming and Analyzing Data with Apache Products

Andru Estes
CLOUD SOLUTIONS ARCHITECT

@andru_estes

Module Overview

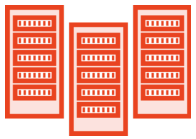
What are we going to learn?

- How to stream data with Apache Flink
- How to stream data with Apache Spark
- Real-time data analyzation with Apache Spark

Get to Know Apache Flink

Apache Flink is a streaming dataflow engine used for real-time stream processing

Supports event time semantics for out-of-order events and exactly-once semantics using third-party connectors



EMR supports Flink as a YARN application

Demo

Enable long-running Flink job on EMR

Apache Spark and EMR

Spark is a distributed processing framework and programming model

Enables machine learning, stream processing, and graph analytics

Optimized DAG engine that caches active data in-memory

Natively supports Scala, Python, and Java

Demo

Adding an EMR Apache Spark step

Module Summary

Interacted with Apache Spark and Flink

Submitted EMR steps utilizing both of the above Apache products

Reviewed some optimization best practices when working with EMR, Spark, and Flink