Processing, Moving, and Querying Data

Andru Estes
CLOUD SOLUTIONS ARCHITECT

@andru_estes

Module Overview

Targeted learning objectives:

- How to process DynamoDB data with Hive
- Move data between EMR and Redshift
- Get familiar with HBase on EMR
- Use Presto to query within EMR

Apache Hive is a software layer that you can use to query map reduce clusters.

Connecting to DynamoDB with Hive

EMR utilizes built-in connectors for accessing other AWS Services, like DynamoDB and Kinesis streams

We can customize Hive within EMR to perform DynamoDB operations, such as loading data into an EMR cluster, or exporting data to S3

Be sure to accurately plan your read and write capacity units for DynamoDB

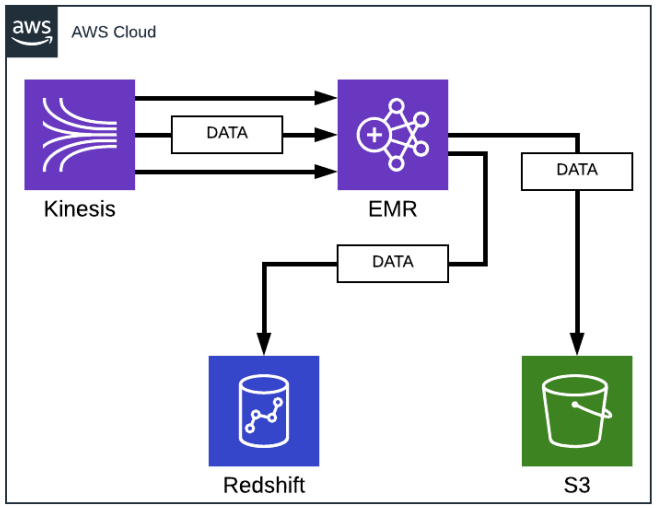
Test out the Pay-Per-Request mode!

Demo

Create a Hive compatible table

Run Hive commands on the created table

Exporting Data from EMR to Redshift



We can copy data from an EMR cluster into Redshift

We use the COPY command

COPY enables reads from multiple sources at the same time

Utilizes SSH with Master node

Requires correct IAM permissions

Demo

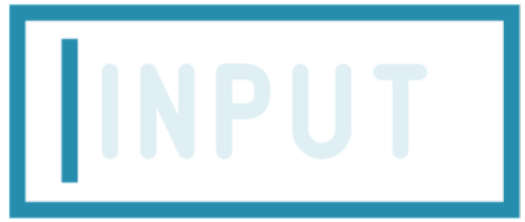
Copy data from EMR cluster steps into Redshift

HBase is an open source, non-relational, distributed database... as part of the... Hadoop project.

HBase and EMR

Runs on top of HDFS

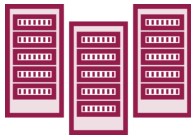
Provides non-relational capabilities



Direct input and output to the MapReduce framework

HBase Capabilities

HBase can store a cluster's HBase root directory and metadata directly in Amazon S3, known as HBase on S3.



HBase with EMR supports read-replica clusters which provides read-only access to a primary cluster's store files and metadata.

HBase snapshots are an alternative to HBase on S3. They create snapshots directly to S3 for recovery.

Check version requirements!

What is Presto?

Open source, distributed SQL query engine

Queries data where it is stored

Allows for fast analytic queries

Compliments Hadoop

Works well with HDFS, S3, and HBase

Used by Netflix and Facebook

Module Summary

- Introduced to Apache Hive
- Imported DynamoDB data into EMR
- Utilized a COPY command to load EMR data into Redshift
- Discovered HBase additional options within EMR
- Used Presto to query within EMR