CRD: ReportDataSource AWSBilling

CRD: ReportDataSource Prometheus Metrics Importer

High level overview of interactions between reporting-operator, Presto, and Hive

1. Scan CRDs in namespace

reporting-operator

2. Get what the ReportDataSource said to gather

3. Create or update tables as needed

S3

Prometheus

Presto

Hive Server & Metastore

metastore(PV or SQL db)

CRD: storage loc

Storage of ORC files (S3,S3-like,GCS, Azure, NFS, HDFS*)

Flow for ReportDataSource of type PrometheusMetricsImporter

Newly created CRD: ReportDataSource

Check if table already created for this ReportDataSource

Apply Hivetable name status.tableRef

reporting-operator

No table? Create a HiveTable

Send table name back

Hive Server & Metastore

reporting-operator

Get list of ReportDataSource

Does the table for this ReportDataSource exist?

Yes

No

Execute Prometheus query contained in spec.prometheusMetricsImporter.query

Skip for now

Presto

Store the data received into a Presto Table Currently done by Executing INSERT query using Presto

Flow for ReportDataSource of type AWSBilling

Newly created CRD: ReportDataSource

reporting-operator

Check if table already created for this ReportDataSource

Apply Hivetable name status.tableRef

No table? Create a HiveTable

Configured to point at the S3 bucket configured in spec.awsBilling

Hive Server & Metastore

reporting-operator

Get list of ReportDataSource

Does the table for this ReportDataSource exist?

Yes

No

Skip for now

Scan configured S3 bucket for the cost usage report JSON manifests.

For each manifest, determine which S3 directory contains the most up to date report and store it

Presto

Table with multiple partitions. One partition per AWS billing period pointing to an S3 directory