PIM Training Program

SQL

Querying with Spectrum tables and best practices

Learning Objective

- At the end of the module, you will be able to write queries to extract data from Spectrum tables

Agenda

- Spectrum enabled VS Normal tables
- Onboarding process
- Best Practices
  - Redshift WLM property change
  - Long Running Queries and Costly Datanet jobs notification mail

Spectrum enabled Vs Normal tables

Amazon Redshift Spectrum enables you to run Redshift SQL queries to extract data that is stored in S3.

The processing that is done in the Redshift Spectrum layer (the S3 scan, projection, filtering, and aggregation) is independent from any individual Amazon Redshift cluster. In general, any operation that can be pushed down to Redshift Spectrum experiences a performance boost because of the powerful infrastructure that supports Redshift Spectrum. To understand the difference between Redshift and Redshift Spectrum, Please refer here.

Note:

- Normal redshift tables do not incur any cost
- RS Spectrum Scan Pricing - For every 1 TB scan, RBS will be charged with $5

Onboarding frequently used external tables locally

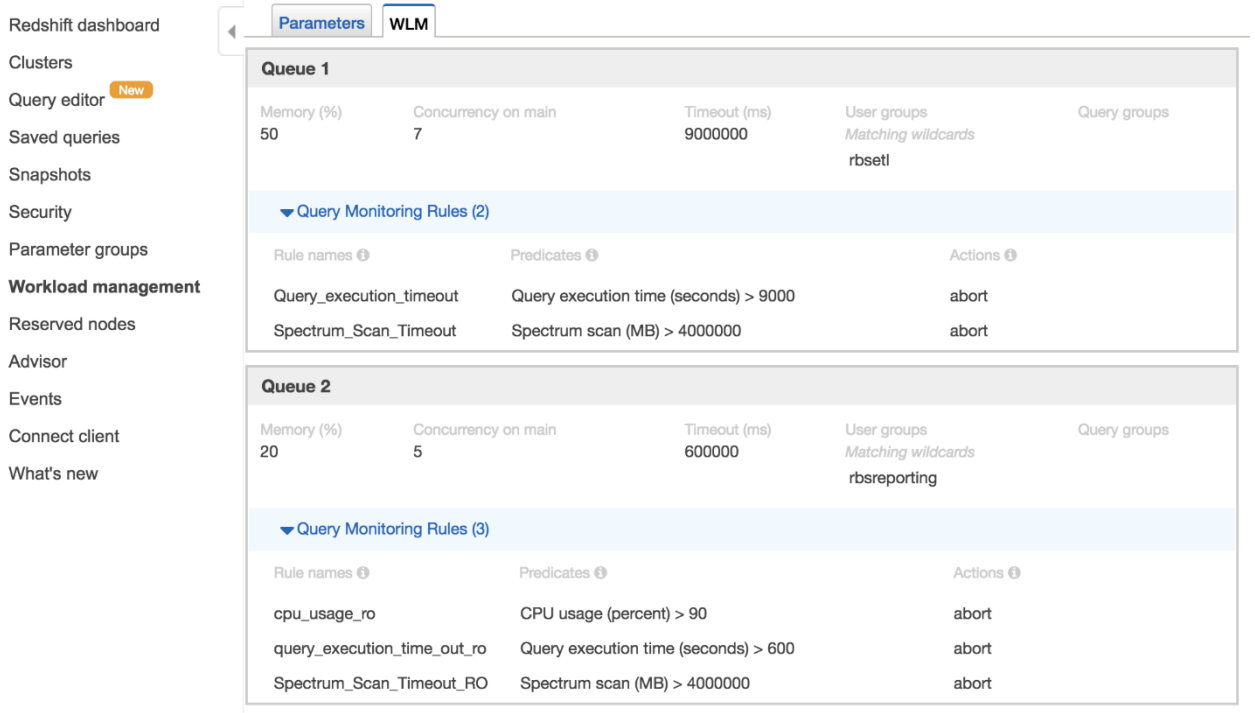
In order to cut the Redshift spectrum scan cost, we have on-boarded few of the frequently used external (spectrum) tables to our cluster locally. With the help of Spectrum Cost Dashboard and the spectrum analysis sheet , we can identify the tables which were used most.

Recently, we have on-boarded D_MP_ASINS & O_REMEDY_TICKETS to RBSETL cluster locally. By on-boarding them we are able to reduce the spectrum cost by $5000 every month. We will continue to monitor the usage of tables and onboard them if required.

Best Practices

Redshift WLM property change

RBS Tech team has implemented WLM property change in our redshift clusters to bring stability & better usage of clusters with optimal performance. A datanet job or SQL query will be automatically aborted if it exceeds the threshold limit of 150 mins runtime or the spectrum scan cost of $20



In the Queue 1 (rbsetl user group), we have defined two rules to abort that particular process. This user group is used by datanet jobs & SQL Workbench

1. Query execution time more than 9000 seconds (150 mins)
2. Spectrum scan size is more than 4,000,000 MB (4 TB) i.e. scan cost of $20

In the Queue 2 (rbsreporting user group), we have defined three rules to abort that particular process. This user group is used by Tableau

1. CPU usage is more than 90%
2. Query execution time more than 600 seconds (10 mins)
3. Spectrum scan size is more than 4,000,000 MB (4 TB) i.e. cost of $20

If a job or sql satisfies any one of the above rule, it will be terminated by Redshift automatically.

Long Running Queries & Costly datanet jobs notification mail

A Tableau dashboard is created to capture all the jobs running at that time in each cluster. It is running for every two hours and will generate a mail to RBS Data Engineering team with the list of long running queries/jobs (Queries running for more than 25 mins) and costly jobs (spectrum scan cost more than $5) in each RBS cluster. It will help the team to take action against the jobs which needs to be optimized.

END