Settings For Large Table Optimization

The AtScale engine features settings that help optimizing aggregates for cloud-based data warehouses. When these features are enabled, system-generated aggregates will leverage distribution keys in the databases (sometimes referred to clustering in certain databases). AtScale will look at the available columns in the system-defined aggregate and select the most appropriate column to distribute the aggregate. This will most likely be a join key, but in the absence of an available key, will use several dimensional value columns, if the database allows for it.

Sample Optimization Results

> The information in this section is based on tests performed by AtScale. Consider that the results in your setup, with your data, and modified settings can be different.

BigQuery

- Tests performed with 27GB aggregate table, and 96MB dimension table.
- Using distribution for the aggregate table results in almost half the query time.
- Adding distribution on the dimension table seems to add minor improvement to the performance.
- With smaller tables, there is no noticeable difference in performance with or without distribution.

Redshift

- Tests performed with aggregate table with 6.7 million rows, and a dimension table with 300 thousand rows.
- For a single node setup, enabling distribution leads to worse query time.
- For a two nodes setup, enabling distribution improves the time.
- Having the dimension table distributed does not impact time for two nodes setup, and worsens the time for a single node setup.

Databricks

Enabling distribution leads to queries that are several times slower.

Enabling And Configuring Optimization

The settings described below can be changed without restarting the AtScale engine. For details on how to update them, see Changing Engine Settings.

Aggregates.largeTableOptimization.enabled

Enables consideration of optimizations for large aggregate tables, such as column-based clustering or distribution. Default value is false.

Aggregates.largeTableOptimization.minimumEstimatedRows

The minimum estimated row count required for the aggregate system to consider applying optimizations such as clustering or distribution. Default value is 100000.

Aggregates.largeTableOptimization.distributionKeyColumn.minimumCardinality

The minimum cardinality required for the highest cardinality dimensional attribute to be used as a table distribution or clustering key. Default value is 30.

More Information

For user defined aggregates, you can specify which columns to be used for distribution. For details, see Defining Aggregates Yourself.