7/25/2022

Rajeev Suravajhula

Spark Notes

Contents

[What is Spark? 2](#_Toc109646542)

# What is Spark?

# Delta Table:

Vacuum:

delta.logRetentionDuration = <interval> : controls how long history for a table is kept. Default value is 30

delta.deletedFileRetentionDuration = "interval <interval>": controls how long ago a file must have been deleted before being a candidate for VACUUM. The default is interval 7 days.

To access 30 days of historical data even if you run VACUUM on the Delta table, set delta.deletedFileRetentionDuration = "interval 30 days". This setting may cause your storage costs to go up

**Dynamic Partition Overwrites**

Configure dynamic partition overwrite mode by setting the Spark session configuration spark.sql.sources.partitionOverwriteMode to dynamic. You can also enable this by setting the DataFrameWriter option partitionOverwriteMode to dynamic. If present, the query-specific option overrides the mode defined in the session configuration. The default for partitionOverwriteMode is static.

Dynamic partition overwrite conflicts with the option replaceWhere for partitioned tables.

* If dynamic partition overwrite is enabled in the Spark session configuration, and replaceWhere is provided as a DataFrameWriter option, then Delta Lake overwrites the data according to the replaceWhere expression (query-specific options override session configurations).
* You’ll receive an error if the DataFrameWriter options have both dynamic partition overwrite and replaceWhere enabled.

**Data skipping**

Data skipping information is collected automatically when you write data into a Delta table. Delta Lake on Databricks takes advantage of this information (minimum and maximum values) at query time to provide faster queries. You do not need to configure data skipping; the feature is activated whenever applicable. However, its effectiveness depends on the layout of your data. For best results, apply [Z-Ordering](https://docs.databricks.com/delta/optimizations/file-mgmt.html#delta-zorder).

Vacuum:

SET spark.databricks.delta.retentionDurationCheck.enabled = false;

SET spark.databricks.delta.vacuum.logging.enabled = true;