Module 22 Challenge

Justin Bein

UC Berkeley Data Bootcamp

January 2024 cohort

In this challenge I will use my knowledge of SparkSQL to determine key metrics about home sales data. Then, I will use Spark to:

1. Create temporary views
2. Partition the data
3. Cache a temporary table
4. Uncache a temporary table
5. Verify the table has been uncached

For this challenge I will use Google Colab

Applications Used

1. spark-3.5.1
2. java 11

Data

The source data was access via the following URL:

"https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-classroom/v1.2/22-big-data/home\_sales\_revised.csv"

The Approach

Part I

A temporary table containing the source data was created.

This table was queried, and the resulting runtime was recorded.

Part II

The temporary table from Part I was cached

This cached table was queried, and the resulting runtime was recorded.

Part III

The source data was converted partitioned.

The partitioned data was saved in parquet format.

This partitioned parquet data was queried, and the resulting runtime was recorded.

Results

Querying cached data has a much faster runtime than querying non-cached data.

Querying partitioned data stored in parquet format has a faster runtime than querying cached data.