Big Query Costs

Breanna Niekamp

2024-09-22

## Assignment Overview

In this project, we will work through estimating the costs of working with BigQuery and come up with strategies to mitigate those costs. We will work with a sample dataset of 50,000 records from the Carbitrage project, using BigQuery queries and estimating the cost for a full year of data (2.5 million records). The table we are working with is umt-msba.carbitrage.raw\_listing\_pages\_50K\_20240913.

For each query we will: - Paste the query - Record the amount of data processed by the pasted query - Estimate how much data would be processed by the query against a full year of data - Estimate the cost for running the query against a full year of data - and finally estimate the cost of running the query every six hours for a year

## Queries

### Query 1

**SELECT COUNT(\*) from the table.** SELECT \* FROM umt-msba.carbitrage.raw\_listing\_pages\_50K\_20240913

Recorded amount of data processed: 1.37 GB (1.275912 GiB) Estimate of how much data would be processed by the query against a full year of data: 63.7956 GiB Estimate of the cost for running the query against a full year of data: $0.31 Estimate the cost of running the query every six hours for a year: $455.79

### Query 2

**SELECT location, COUNT(\*) from the table.** SELECT location, COUNT(\*) FROM umt-msba.carbitrage.raw\_listing\_pages\_50K\_20240913 GROUP BY location ORDER BY 2 DESC

Recorded amount of data processed: 480.44 KB (0.0004474446 GiB) Estimate of how much data would be processed by the query against a full year of data: 0.2237223 GiB Estimate of the cost for running the query against a full year of data: $0.00011 Estimate the cost of running the query every six hours for a year: $0.16

### Query 3

**SELECT url, location from the table.** SELECT url, location FROM umt-msba.carbitrage.raw\_listing\_pages\_50K\_20240913 ORDER BY datetime\_pulled DESC

Recorded amount of data processed: 5.1 MB (0.00474975 GiB) Estimate of how much data would be processed by the query against a full year of data: 0.2374875 GiB Estimate of the cost for running the query against a full year of data: $0.00116 Estimate the cost of running the query every six hours for a year: $1.69

### Query 4

**SELECT url, raw\_html, location for just the “missoula” location from the table.** SELECT url, raw\_html, location FROM umt-msba.carbitrage.raw\_listing\_pages\_50K\_20240913 WHERE location = ‘missoula’ ORDER BY datetime\_pulled DESC

Recorded amount of data processed: 1.37 GB (1.275912 GiB) Estimate of how much data would be processed by the query against a full year of data: 63.7956 GiB Estimate of the cost for running the query against a full year of data: $0.31 Estimate the cost of running the query every six hours for a year: $455.79

## Conclusion

Based on the data gathered above, it would seem that the queries that drive costs include those that query the full dataset and queries that include bigger columns like the raw\_html. It would be good to include more filtration when querying such as selecting only the needed columns, and using more clauses like “WHERE”. The more precise you are with what is needed, the cheaper it would be.