



CS 587 Database Implementation

Winter 2021

Database Benchmarking Project

Team:
Likhitha Vanga
Jaya Bhargavi Vengala

Systems Selected



System 1: PostgreSQL

- PostgreSQL is a row- based database.
- we wanted this as an opportunity to learn more about postgresql and gain expertise in this database.

Systems Selected



System 2: Big Query

- BigQuery is Google's serverless data warehouse.
- BigQuery prioritizes scalability and quickness in queries, means that you can easily scale and perform ad hoc analyses much faster than you would on cloud-based server structures.

Key differences



PostgreSQL	Big Query
Row based Database	Column Based database
Uses Indices	No Indices
Uses Keys	No primary or Unique Constraints

Benchmark Goals



- Compare 2 systems(postgreSQL, Big Query)
- Performance Analysis based on following experiments.
 - ◆ Full Scan
 - ◆ Selectivity
 - ◆ Insertion
 - ◆ Aggregates

Experiment - 1(Full table scan)



Table Size: 1 Million Tuples

Goal: Measure performance of scanning full table in 2 systems.

Result:


PostgreSQL	BigQuery
0.319 sec	5 sec

Discussion of Experiment-1



- The results of our first query1 (full table scan) shows that the Cloud PostgreSQL has performed better than Cloud BigQuery which was quite faster and took much less time.
- Our expectation was that cloud Postgres will perform better than Bigquery and the results matched .

Experiment - 2 Selectivity



Goal: To measure performance on different selectivities

Table size: 1 Million tuples

Selectivity	PostgreSQL	Big Query
10%	0.532 sec	2.2 sec
20%	0.48 sec	3.8 sec
40%	1.21 sec	4.9 sec

Discussion of Experiment-2



- Results from the experiment shows that postgres performs better than BigQuery on different selectivities.
- Bigquery cache management is better than postgresSQL. (Cloud BigQuery has a very good cache management system. It takes '0' Second to run the same query again)

Lessons learnt



- Aggregate queries and rowbase database may run slower than column base database.
- Cloud BigQuery has a very good cache management system. It takes ‘0’ Second to run the same query again.
- Due to the fact that a column-based database has each column’s data in a separate file, it is less than ideal for table scan which we observed in Experiment-1.

Lessons Learnt



- We learned that postgres provides us ways to turn on and off various index options that are available from which we can actually understand the usage of each of the indexes being present.
- We noticed that PostgreSQL incur some overheads on the execution time maybe due to its display time and interface requirements which result in more time than Command line.
- According to our research, column-based database works better than row-based database but our experiments are in contradiction.



Thank You !!!

Have a great day :)

Appendix



Experiment-3, 4

New experiments

Experiment - 3 Insertion



Goal: To measure the performance of a single insertion in two different systems

Table Size: 1 Million Tuples

PostgreSQL	BigQuery
0.87 sec	1.5 sec

Discussion of Experiment-3



- Results from the experiment shows that a single insertion performance is a little slower in BigQuery, The reason might be the architecture of cloud deployment which stores data on different node and since, the insertion happens at the end of the table it search the entire node to get to the end and then insert the new one.
- The result match with our experiment where we predict that the insertion in Cloud PostgreSQL will take less time than Cloud BigQuery.

Experiment - 4



Goal: To compare performances using aggregates on two systems

Table Size: 1 Million Tuples

PostgreSQL	BigQuery
0.59 sec	0.5 sec

Discussion of Experiment-4



- The results from experiments shows that there is no much difference in performance in aggregation.
- The performance order we expect is:
 - With Indices: PostgreSQL will perform better than BigQuery since BigQuery doesn't support indices.
 - Without Indices: The performance will depend on the size of the table.