



# **BENCHMARKING**

CLICKHOUSE VS BIGQUERY

BY DIFINATIVE TECHNOLOGIES

*PUBLISHED ON JUNE 2023*

# Contents

- Executive summary ..... 3
- 1. Introduction ..... 4
  - 1.1 Background and context..... 4
- 2. Methodology ..... 5
  - 2.1 Evaluation criteria..... 5
  - 2.2 Benchmarking process and approach..... 5
- 3. ClickHouse and BigQuery overview ..... 6
  - 3.1 Description of each solution ..... 6
- 4. Features, strengths, and weakness..... 6
- 5. Use Case Examples ..... 7
  - 5.1 Real-world scenarios showcasing the applicability. .... 7
  - 5.2 Relevance of the benchmarking study for addressing challenges ..... 8
- 6. Benchmarking Results..... 8
  - 6.1 Data ingestion performance ..... 8
  - 6.2 Mutation ..... 9
  - 6.3 Storage efficiency ..... 10
  - 6.4 Ease of use and integration..... 10
  - 6.5 Query performance ..... 11
- 7. Recommendations ..... 13
  - 7.1 Tailored recommendations for different scenarios..... 13
  - 7.2 Consideration for OpenTelemetry Data Storage..... 13
- 8. Cost benefit analysis..... 14
  - 8.1 Cost Analysis ..... 14
- 9. Conclusion..... 15
  - 9.1 Summary of key findings ..... 15
- 10. About the contributors ..... 16
- 11. Note of thanks..... 17
- 12. Acknowledgements..... 18

# Executive summary

## Brief overview of the benchmarking study

This report presents the findings of a comprehensive benchmarking study comparing two leading columnar databases: ClickHouse, supported by ChistaData, and Google's BigQuery. The objective was to evaluate their capabilities in the context of managing OpenTelemetry data for observability and performance engineering, thereby identifying the most efficient solution.

Our analysis revealed that ChistaData's ClickHouse demonstrated superior performance in several critical aspects of the evaluation. Notably, it exhibited impressive read performance, which can be attributed to the flexibility it offers in tuning to custom data types. Furthermore, ClickHouse excelled in areas like data compression, scalability, and storage efficiency, making it a strong contender for handling large volumes of OpenTelemetry data.

On the other hand, BigQuery demonstrated its strength in specific areas such as ease of onboarding and general write and ingestion performance. Its user-friendly interface and robust data ingestion capabilities make it a viable choice for organizations looking for a straightforward, efficient solution. Taken together, these findings suggest that while both ClickHouse and BigQuery have their respective strengths, the choice between them should hinge on the specific needs and priorities of the organization. If customization and high read performance are key criteria, ClickHouse's flexible data type tuning, backed by ChistaData's support, makes it a compelling choice. Alternatively, for organizations that prioritize ease of onboarding and robust data ingestion, BigQuery stands out as an effective solution.

In conclusion, this benchmarking study provides valuable insights to help organizations make an informed decision when choosing between ClickHouse and BigQuery for managing OpenTelemetry data. The subsequent sections of this report provide a detailed analysis of the benchmarking results and offer recommendations to guide the decision-making process.

# 1. Introduction

## 1.1 Background and context

As organizations continue to leverage data-driven insights for decision-making, the importance of robust, efficient, and secure data storage solutions cannot be overstated. Cloud-based solutions, like Google's BigQuery, have been widely adopted due to their scalability, ease of use, and the advantages inherent in Software as a Service (SaaS) models. However, for enterprises with strict compliance and security requirements, the need for an on-premises solution in their private cloud environment that provides control over its system design and integration is paramount.

This benchmarking study was designed to aid such enterprises, providing an in-depth comparison between two leading columnar database solutions: ClickHouse, supported by ChistaData, and Google's BigQuery. Both databases were evaluated on multiple parameters, such as data ingestion performance, query performance, scalability, storage efficiency, and ease of use and integration. The goal was to provide an independent view of the strengths and weaknesses of these options, guiding enterprises in their decision-making process when an alternate to the cloud-based solution is required.

There are other columnar databases in the market such as Apache Arrow, MariaDB's ColumnStore, MonetDB, and Greenplum. While these databases have their unique offerings, for the purpose of this study, we have focused on ClickHouse and BigQuery due to their prominence, community support, and wide-scale use in the industry.

Our intention is to offer a comprehensive, unbiased perspective that will serve as a valuable resource for enterprises seeking to make an informed choice. This benchmark should be seen as a guidepost, providing insights into how ClickHouse and BigQuery perform under various circumstances, with a particular emphasis on their application in managing OpenTelemetry data for observability and performance engineering.

In the following sections, we delve into the specifics of the benchmarking results, providing a detailed comparative analysis and actionable recommendations.

## 2. Methodology

We followed a structured and rigorous methodology to conduct this benchmarking study, which ensures the validity and reliability of the results. The study aimed to provide a comprehensive view of how ClickHouse and BigQuery perform and compare in handling OpenTelemetry data.

### 2.1 Evaluation criteria

Our evaluation was based on several key criteria that are critical to the performance and efficiency of a columnar database. These include:

- **Data Ingestion Performance:** The speed and latency of data ingestion were measured, assessing the databases' capabilities in handling large volumes of data efficiently.
- **Query Performance:** The speed and efficiency of simple, complex, and resource-intensive queries were evaluated to understand how each database handles data retrieval under different circumstances.
- **Scalability:** The ability of each database to scale, both in terms of handling increasing data volumes and accommodating more concurrent users, was assessed.
- **Storage Efficiency:** Factors such as data compression capabilities, storage cost per GB, data retention options, and support for data partitioning were evaluated.

### 2.2 Benchmarking process and approach

The benchmarking process was undertaken in a controlled environment to ensure that the performance of both databases could be accurately measured and compared. The following steps were followed:

- **Data Preparation:** A representative sample of OpenTelemetry data amounting to 100 million rows, encompassing traces, logs, and metrics, was prepared for ingestion into both databases.
- **Database Configuration:** ClickHouse and BigQuery were set up and configured according to their respective best practices to ensure optimal performance.
- **Benchmarking Execution:** A series of tests were run to evaluate each database according to the criteria outlined above. Tests were repeated multiple times to confirm consistency of results.
- **Data Analysis:** The results of the tests were compiled and analyzed to draw comparisons between the performance of ClickHouse and BigQuery.
- **Reporting:** The findings were documented in this report, providing a detailed comparison of the two databases, along with insights and recommendations for enterprises considering these solutions.

By following this structured approach, we aimed to provide an independent and objective comparison of ClickHouse and BigQuery, highlighting their respective strengths and weaknesses in handling OpenTelemetry data.

### 3. ClickHouse and BigQuery overview

This section compares two columnar databases: ClickHouse, supported by ChistaData, and Google’s BigQuery. We explore their main features, advantages, and disadvantages to set the stage for the following benchmarking analysis.

#### 3.1 Description of each solution

**ClickHouse:** ClickHouse is an open-source columnar DBMS that enables real-time analytical data reporting. It supports standard SQL syntax and can handle large amounts of data with fast query processing. ChistaData helps enterprises use ClickHouse to perform real-time data analysis for various applications and services.

**BigQuery:** BigQuery is a fully-managed, serverless data warehouse that offers super-fast SQL queries using Google’s infrastructure. It supports standard SQL syntax and provides easy scalability and flexibility. It’s a good option for enterprises looking for a cloud-based data analytics solution with low maintenance.

### 4. Features, strengths, and weakness

ClickHouse:

Key Features	Strengths	Weaknesses
<ul style="list-style-type: none"><li>• Real-time query processing</li><li>• Support for standard SQL syntax</li><li>• High data compression ratio</li><li>• Scalable architecture</li><li>• Flexible deployment models for high secure environments</li></ul>	<ul style="list-style-type: none"><li>• Offers high-speed reads and lower query latency.</li><li>• Provides flexibility in tuning to custom data types, enhancing read performance.</li><li>• Excellent scalability to handle increasing data volumes.</li><li>• Ability to deploy in on-prem data centers and edges.</li></ul>	<ul style="list-style-type: none"><li>• Steeper learning curve for setup and optimization</li><li>• Requires manual management, which could be complex compared to SaaS solutions.</li><li>• SQL limitations in update, join and delete operations.</li></ul>

## BigQuery:

Key Features	Strengths	Weaknesses
<ul style="list-style-type: none"><li>• Fully-managed, serverless architecture</li><li>• Built-in machine learning capabilities</li><li>• Seamless scalability</li><li>• Support for standard SQL syntax</li></ul>	<ul style="list-style-type: none"><li>• Offers ease of onboarding with a user-friendly interface</li><li>• Robust write and ingestion performance</li><li>• Lower administrative overhead due to its fully-managed nature</li><li>• Relatively lower initial cost for small data volume, free tier option</li></ul>	<ul style="list-style-type: none"><li>• No on-premise deployment available</li><li>• Cost can be higher for large-scale data analysis, depending on usage patterns</li><li>• Limited SQL functionality</li><li>• Data export limitation</li></ul>

## 5. Use Case Examples

This section outlines several real-world scenarios where the use of ClickHouse and BigQuery could provide significant benefits. These examples, drawn from various sectors such as banking, telecom, and e-commerce, illustrate the versatility and applicability of these columnar databases. The benchmarking study using OpenTelemetry data provides valuable insights for these sectors, given the structural similarity of this data to many enterprise scenarios.

### 5.1 Real-world scenarios showcasing the applicability.

**Banking:** Banks deal with an enormous volume of data related to transactions, customer interactions, and risk assessments. These institutions need to process and analyze this data in real-time for fraud detection, customer service optimization, and regulatory compliance. Both ClickHouse and BigQuery can handle such large data volumes and deliver quick, actionable insights.

**Telecom:** Telecommunication companies generate and collect vast amounts of data from network usage, customer data, and system logs. Analyzing this data can improve network performance, optimize resource allocation, and enhance customer experience. The high-speed processing and real-time query capabilities of ClickHouse and BigQuery are highly beneficial for these tasks.

**E-commerce:** E-commerce platforms have a diverse range of data, from customer interactions and transaction history to website analytics and supply chain data. Real-time analysis of this data can drive personalized customer experiences, efficient inventory management, and strategic decision-making. ClickHouse and BigQuery's ability to handle large data volumes and deliver real-time analysis make them well-suited to this sector.

**Generative AI use cases:** Data is crucial for organizations, and it comes in different forms: structured, unstructured, and semi-structured. Structured data is organized and fits well in databases, while unstructured data lacks organization and includes text-heavy content like emails



and documents. Semi-structured data is a mix, often tagged for searchability. Columnar databases like ClickHouse store data by columns instead of rows, making them efficient for analytical queries and handling big data. In the era of Generative AI, data quality affects the capabilities of AI models like GPT-3, which generate content like humans. ClickHouse is valuable for storing, retrieving, and analyzing the data generated by these models. Therefore, understanding and effectively managing the various forms of data within an organization, along with efficient data handling systems like ClickHouse, are not just beneficial, but necessary in the era of Generative AI.

## 5.2 Relevance of the benchmarking study for addressing challenges

The benchmarking study, conducted with OpenTelemetry data, is highly relevant for these sectors. OpenTelemetry, with its increasing adoption as a standard for observability, generates data structures that closely mirror those found in many enterprise scenarios.

The study provides insights into how effectively ClickHouse and BigQuery can manage such data, considering factors like ingestion performance, query speed, and scalability. This information can guide enterprises in these sectors when they're building or optimizing their own OpenTelemetry-based observability stacks.

Moreover, as these industries increasingly embrace open-source solutions for their technology stacks, the insights gained from this benchmarking study will be crucial in driving informed decisions about the best fit for their specific needs.

# 6. Benchmarking Results

This section provides a detailed overview of the benchmarking results, which encompass key performance areas such as data ingestion, query performance, scalability, storage efficiency, and ease of use and integration.

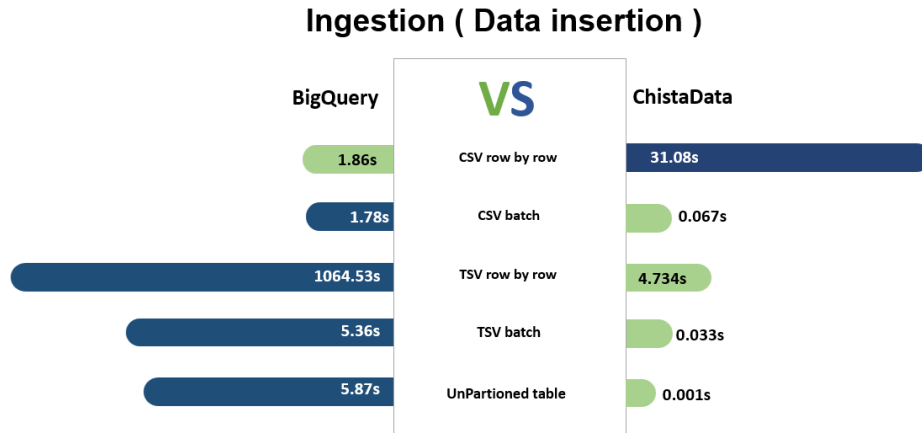
This section offers a side-by-side comparison of ClickHouse and BigQuery, based on our benchmarking evaluation criteria. It identifies the key strengths and weaknesses of each solution, providing a comprehensive picture for enterprises seeking to choose the best fit for their needs.

## 6.1 Data ingestion performance

Data ingestion performance represents the speed and efficiency with which a system can ingest large volumes of data. Both ClickHouse and BigQuery demonstrated strong performance in this area. However, BigQuery showed a slight edge in general write and ingestion performance, owing to its fully-managed, serverless architecture.



Scenario	BigQuery	Chistadata
CSV row by row	1.86s	31.08s
CSV batch	1.78ss	0.067s
TSV row by row	1064.53s	4.734s
TSV batch	5.36s	0.033s
Unpartitioned table (100000 rows)	5.87s	0.001s



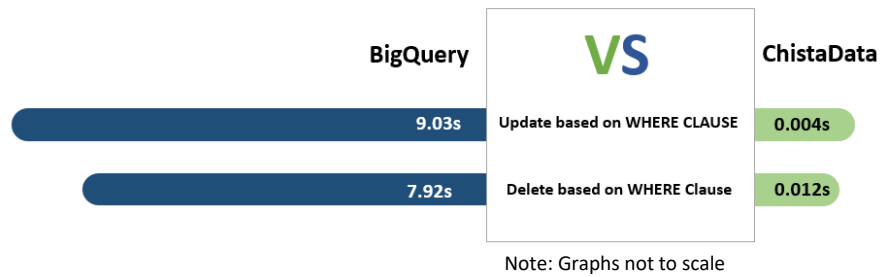
Note: Graphs not to scale

## 6.2 Mutation

Mutation refers to the modification or alteration of data within a database. Clickhouse doesn't support updates and deletes directly. They are implemented using ALTER statements which are known as mutations that are resource intensive and exhibited an excellent scalability. BigQuery, with its inherent cloud-based design, also demonstrated robust scalability capabilities.

Scenario	BigQuery	Chistadata
Update the data based on WHERE clause	9.03s	0.004s
Delete the data based on WHERE clause	7.92s	0.012s

## Mutation ( Update & Delete )

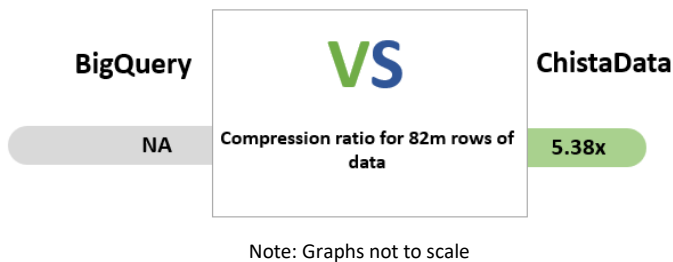


### 6.3 Storage efficiency

In the aspect of storage efficiency, which includes factors such as data compression capabilities, storage cost per GB, data retention options, and support for data partitioning, both ClickHouse and BigQuery performed well. ClickHouse's high data compression ratio offers a significant advantage, while BigQuery's fully managed service provides convenient data storage and management options.

Scenario	BigQuery	Chistadata
The compression ratio for 82 million rows of data	Not clearly disclosed	5.38x

## Data compression



### 6.4 Ease of use and integration

When considering ease of use and integration, BigQuery came out on top due to its user-friendly interface, lower administrative overhead, and comprehensive integration options. However, ClickHouse, with support from ChistaData, also provides extensive integration points and APIs/SDKs, making it a viable option for enterprises that prefer more control over their database management.

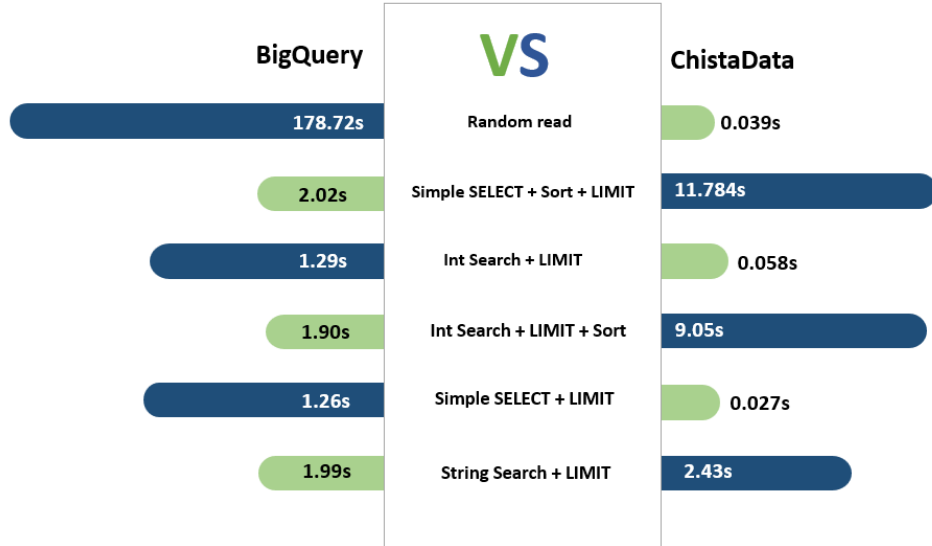
Overall, these benchmarking results provide a detailed picture of the strengths and weaknesses of both ClickHouse and BigQuery, enabling enterprises to make an informed decision based on their specific needs and requirements.

## 6.5 Query performance

Query performance is critical for real-time data analysis. In our benchmarking tests, ClickHouse showed superior performance, especially in complex and resource-intensive queries, due to its flexibility in tuning to custom data types. BigQuery also showed strong performance, although it was somewhat slower in comparison for more complex queries.

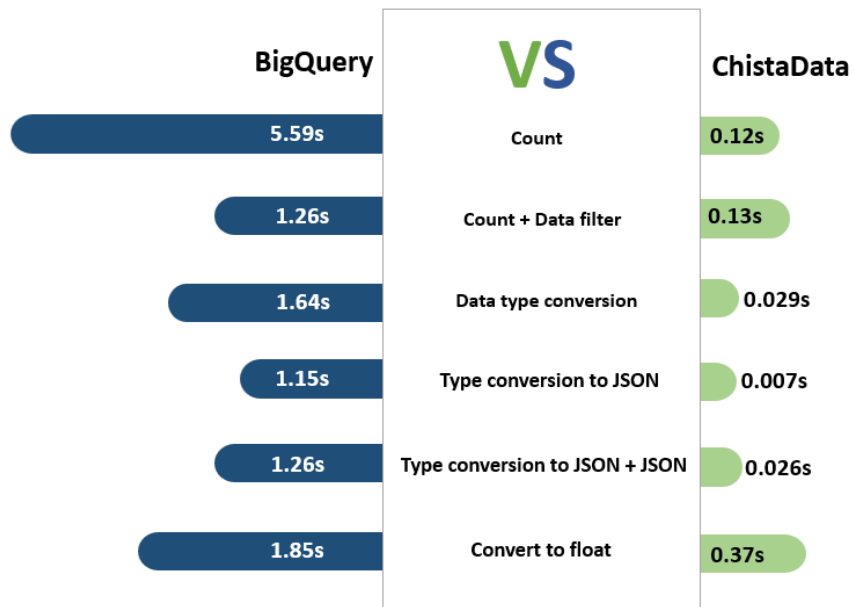
Scenario	BigQuery	Chistadata
Random read	178.72s	0.039s
Query -Simple SELECT + Sort + LIMIT	2.02s	11.784s
Query - Int Search + LIMIT	1.29s	0.058s
Query - Int Search + LIMIT + Sort	1.90s	9.05s
Query -Simple SELECT + Limit	1.26s	0.027s
Query - String Search + LIMIT	1.99s	2.43s
Query - Count	5.59s	0.12s
Query - Count + data filter	1.26s	0.13s
Query - Data type conversion	1.64s	0.029s
Query- Type conversion to JSON	1.15s	0.007s
Query- Type conversion JSON + JSON	1.26s	0.026s
Convert to float	1.85s	0.37s
Aggregate function – 1 + sorting	1.75s	0.005s
Aggregate function – 2 + sorting	1.67s	0.029s
Aggregate function – 3 + sorting	1.57s	0.024s
Aggregate function – 4 + sorting	1.46s	0.448s
Aggregate function – 5 + sorting	1.46s	0.233s

## Query performance

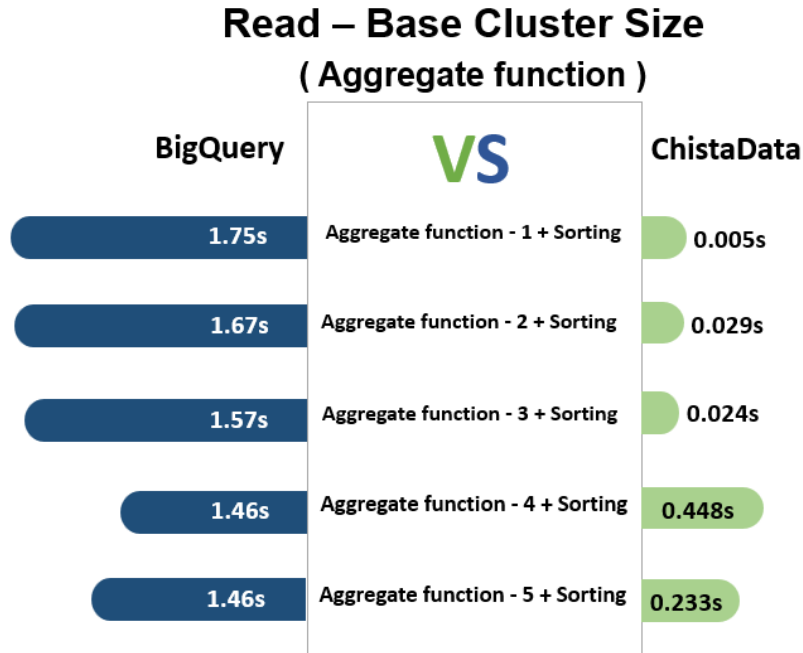


Note: Graphs not to scale

## Query performance



Note: Graphs not to scale



Note: Graphs not to scale

## 7. Recommendations

This section provides tailored recommendations for different scenarios and organizations, along with specific considerations for selecting the most suitable columnar database for columnar data storage and analysis.

### 7.1 Tailored recommendations for different scenarios

Scenario	Recommendation
High-volume data ingestion	Both ClickHouse and BigQuery perform well, but BigQuery has a slight edge due to its serverless architecture
Complex query performance	ClickHouse, given its superior performance in complex and resource-intensive queries

### 7.2 Consideration for OpenTelemetry Data Storage

Choosing the best columnar database for OpenTelemetry data storage and analysis should factor in the following considerations:

- Data ingestion performance:** Consider the volume of OpenTelemetry data that will be generated and how fast it needs to be ingested into the system. BigQuery may have an edge here with its serverless architecture. A batch ingestion is always recommended over a streaming update for large volume of data.

- **Query performance:** If the use case involves complex queries on OpenTelemetry data, ClickHouse's superior query performance could be a deciding factor.
- **Scalability:** As OpenTelemetry data grows, so does the need for a scalable database. Both ClickHouse and BigQuery offer high scalability, but specific scalability requirements might tilt the balance.
- **Storage efficiency:** Depending on the volume of OpenTelemetry data and the storage cost, ClickHouse's high data compression ratio might provide a significant advantage.
- **Ease of use and integration:** If ease of setup, use, and integration with other systems is a priority, BigQuery would be a suitable choice.

These recommendations should be viewed as guidelines. The choice between ClickHouse and BigQuery should ultimately align with the specific needs and requirements of the organization and the nature of the OpenTelemetry data it handles.

## 8. Cost benefit analysis

### 8.1 Cost Analysis

When comparing ClickHouse and BigQuery, it's important to consider the cost aspect. ClickHouse, being an open-source solution supported by ChistaData, offers a cost advantage as there are no licensing fees associated with its usage. Enterprises can leverage ClickHouse on their own infrastructure, making it a cost-effective option for on-premises deployments. However, it's important to note that the overall cost of implementing ClickHouse may vary depending on factors such as hardware infrastructure, maintenance, and support.

On the other hand, BigQuery operates on a pay-as-you-go model within Google Cloud Platform. While this provides scalability and eliminates upfront infrastructure costs, it's crucial to consider the pricing structure based on data storage, queries, and data transfer. Enterprises need to evaluate their data usage patterns and projected costs to determine the most cost-effective option between ClickHouse and BigQuery.

Cost element	BigQuery
Storage	\$0.020 per GB/month for active storage; \$0.010 per GB/month for long-term storage
Streaming inserts	\$0.010 per 200MB
Querying	\$5.00 per TB (beyond 1 TB)
Flat-Rate pricing	Starts at \$10,000 per month for 500 slots
Data transfer	Free up to 1GB/day, then \$10.

Options (For 16GB RAM/month)	ChistaData Clickhouse(Cloud)
Basic ClickHouse Server Nodes (Shared)	\$960 Intel 8 CPU / 320 GB NVMe SSDs / 6 TB Transfer
General Purpose ClickHouse Server Node	\$250 /4 CPUs / 50 GB SSDs / 5 TB Transfer
CPU Optimized ClickHouse Server Node	\$350 /8 CPUs / 100 GB SSDs / 6 TB Transfer
Memory Optimised ClickHouse Server Node	\$200 /2 CPUs / 50 GB SSDs / 4 TB Transfer
Storage Optimised ClickHouse Server Node	\$300 /2 CPUs / 300 GB NVMe SSDs / 4 TB Transfer

*Note : The cost analysis provided here is based on information available as of June 15th, 2023, and may be subject to change. It is recommended to verify the latest pricing and features of ClickHouse and BigQuery before making any decision.*

When evaluating the pricing between BigQuery and ChistaData cloud, one must consider various aspects of the Total Cost of Ownership (TCO) for a solution. These aspects include the number of queries, streaming inserts, compression, and data transfer.

## Key benefits of on-premises deployment

On-premises deployments, like ClickHouse offered through Chistadata, come with several advantages that can be particularly beneficial for certain enterprises. Some of these advantages include:

- Customization: Tailored hardware, software, and configurations.
- Performance: Potentially better performance with local data access.
- Cost Control: Long-term cost-effectiveness for high, predictable requirements.
- Security: Direct control over security systems and protocols.
- Regulatory Compliance: Ensures local data storage for compliance with data sovereignty laws.
- Integration: Better compatibility with local enterprise IT systems.

## 9. Conclusion

### 9.1 Summary of key findings

The benchmarking study undertaken to compare ClickHouse and BigQuery in the context of OpenTelemetry data storage and analysis has led to several key insights. Both ClickHouse and BigQuery offer robust capabilities as columnar databases, excelling in different aspects of data management and analysis.

ClickHouse demonstrated superior performance in complex query processing, scalability, and storage efficiency due to its high data compression ratio. It is a highly suitable option for organizations seeking more control over their database management and who have specific needs



for on-premise deployments. ClickHouse's flexibility and performance, especially in the context of handling OpenTelemetry data, make it an appealing choice for a wide range of businesses, including those in banking, telecom, and e-commerce sectors.

BigQuery, on the other hand, shone in the areas of data ingestion performance and ease of use and integration. Its fully-managed, serverless architecture, user-friendly interface, and comprehensive integration options make it an excellent choice for organizations looking for a cloud-based solution with less administrative overhead.

However, no one-size-fits-all solution exists when it comes to choosing a database for OpenTelemetry data storage and analysis. The choice between ClickHouse and BigQuery should be driven by the specific needs, constraints, and objectives of the organization. Factors like data volume, query complexity, scalability requirements, storage efficiency, and ease of use and integration should all play into this decision.

In conclusion, this benchmarking study has provided an independent and detailed comparison of ClickHouse and BigQuery, two leading columnar databases. It is our hope that the insights and recommendations presented will serve as a valuable guide for enterprises on their journey to select the most suitable database solution for their OpenTelemetry data storage and analysis needs.

## 10. About the contributors

DifiNative is a globally operating IT services company based in Bengaluru. Incepted in 2021 by industry stalwarts known for their exceptional contributions in Global SIs, DifiNative has demonstrated its proficiency by guiding its clients through transformative, large-scale programs and designing bespoke practices and IPs tailored to various industries.

Specializing in automation and cloud-native solutions, DifiNative's broad spectrum of services include consulting, development, deployment, and support. Driven by their commitment to assisting businesses in their digital transition, DifiNative's team of experienced engineers and consultants relentlessly strive to ensure client success by delivering premier, customized solutions.

At DifiNative, the focal point is cloud-native solutions built around Kubernetes, with an array of accelerators dedicated to the data & AI lifecycle. To supplement this, the company offers a suite of tools geared towards deployment, management, monitoring, and security.

DifiNative also takes pride in its unique product, Squirrel Vision. This distinctive, computer vision-based SaaS platform harnesses the power of cutting-edge AI to meet all requirements of a planet-scale enterprise solution. Labeled as a "DevOps++" for computer vision, Squirrel Vision incorporates proprietary algorithms, processing pipelines, MLOps, and DataOps, positioning it as a comprehensive, versatile solution for enterprises.

Authors:

- **Saji Thoppil**

Saji Thoppil is an industry veteran with a successful career of three decades, which included a notable three-year tenure on the board of Linux Foundation Edge. Currently, he holds the positions of Founder & CTO at DifiNative Technologies. Previously, he had an illustrious career at Wipro Technologies, where he was recognized as a Wipro Fellow and held the role of CTO for its Cloud & Infrastructure business.

- **Prasanth Sekharan**

Prasanth Sekharan brings over 20+ years of IT industry experience with a strong focus on open-source technologies. He has worked closely with renowned Fortune 500 companies, providing comprehensive IT support across various technologies. Currently, Prasanth is the VP leading IT Operations and service at DifiNative Technologies.

## 11. Note of thanks

We would like to extend our heartfelt gratitude to all the stakeholders, participants, and the wider community whose invaluable input and support made this benchmarking study possible. Your contributions have been instrumental in making this work a comprehensive and insightful resource for the industry.

- **Shiv Iyer**

Shiv is an accomplished business leader in open-source databases. He has worked with Twitter, Pinterest, and PayPal. Currently, he is the founder of MinervaDB Inc. and ChistaDATA Inc., offering consultative support and managed services for various databases.

- **Alkin Tezuysal**

Alkin is an open-source database evangelist, global database operations expert, and cloud infrastructure architect. Alkin is known for being an inspiring technical and strategic leader, and is an accomplished author, speaker, mentor, and coach. Currently, he holds the position of EVP - Global Services at ChistaData.

- **Vijay Anand**

Vijay Anand is an accomplished professional with over 10 years of experience. He possesses extensive hands-on coding expertise in developing and deploying Python-based applications. Furthermore, he has authored a book called "Up and Running with ClickHouse" published by BPB publications.

## 12. Acknowledgements

We would like to acknowledge the following open-source programs from GitHub that were instrumental in conducting this benchmarking study:

- CNCF Open telemetry
  - <https://github.com/open-telemetry>
- Open telemetry operator
  - <https://github.com/open-telemetry/opentelemetry-operator>
- Grafana
  - <https://github.com/grafana/grafana>
- CNCF Jaeger
  - <https://github.com/jaegertracing/jaeger>

We extend our sincere gratitude to the authors and developers of these open-source programs for their valuable contributions. The availability of these programs greatly facilitated our research and enhanced the quality of our benchmarking study.