# An Inside Look at Google BigQuery

Google BigQuery is a cloud-based interactive query service designed for processing massive datasets. It is built on Google's core technology, Dremel, which enables fast and ad hoc querying of large datasets. BigQuery offers the same performance and core features of Dremel to third-party developers, allowing them to leverage its capabilities for their data processing needs. Compared to technologies like MapReduce, Hadoop, and traditional data warehouse solutions, BigQuery stands out in terms of its query performance, scalability, and user-friendly interface. It provides a cost-effective and efficient solution for businesses and developers to analyze and process big data without the need for complex programming or managing large computing clusters.

Dremel's columnar storage offers benefits such as reduced data transfer, improved query performance, and higher compression ratios. By storing data in columns, Dremel can efficiently process queries by accessing only the necessary columns, minimizing data transfer and speeding up query execution. The ability to achieve higher compression ratios is particularly advantageous for datasets with columns that have low variation in values. Overall, Dremel's columnar storage contributes to its exceptional performance and efficient utilization of storage resources.

Dremel's tree architecture plays a crucial role in achieving its massive parallelism and fast query processing. By using a distributed tree structure, queries can be efficiently dispatched and results can be aggregated across thousands of machines. This enables Dremel to process queries in seconds, regardless of the size of the dataset. The parallel execution of queries across multiple machines ensures high scalability and enables Dremel to handle large-scale data processing tasks with ease. The combination of columnar storage and tree architecture forms the foundation of Dremel's exceptional performance, allowing it to provide fast and efficient query processing for large datasets.

Dremel's significance within Google has led to the externalization of its core features in the form of Google BigQuery. Released as a publicly available service, BigQuery allows businesses and developers outside of Google to leverage the power of Dremel for their own big data processing requirements. BigQuery provides the same underlying architecture and performance characteristics as Dremel, offering scalability, high-speed query processing, and seamless integration with Google Cloud Storage. With BigQuery, users can benefit from the massive computational infrastructure of Google, including features like multiple replication across regions and high data center scalability, all without the need for managing the infrastructure themselves. BigQuery represents a powerful tool for unlocking the potential of big data analytics and enables organizations to process and analyze massive datasets at "Google speed" using Dremel's capabilities.

BigQuery, the cloud-based query service built on the foundation of Dremel, allows businesses and developers to harness the power of Dremel's massively parallel query engine for their big data processing needs. With its comprehensive set of features, including a REST API, command line interface, and web UI, BigQuery provides a user-friendly and efficient platform for querying and managing large datasets. It distinguishes itself from technologies like MapReduce by offering near-instantaneous query response times and enabling ad hoc data analysis, making it accessible to both programmers and non-programmers. By externalizing Dremel's capabilities, BigQuery brings the benefits of fast query performance and cost-effectiveness to a wider

audience, empowering them to leverage Google's infrastructure for efficient data processing.

In contrast to MapReduce, Dremel and BigQuery are designed as interactive data analysis tools for large datasets, prioritizing fast query performance and providing a user-friendly experience. Unlike MapReduce, which requires programming and customization, Dremel and BigQuery offer a more intuitive approach to data exploration, making them accessible to non-programmers. They enable ad hoc queries and interactive analysis, completing most queries within seconds or tens of seconds. The data flow in MapReduce involves mappers and reducers for batch-oriented processing, while BigQuery offers a direct SQL-like query interface for interactive analysis. Overall, BigQuery provides a more user-friendly and performant solution for large-scale data analysis compared to traditional batch processing frameworks like MapReduce and Hadoop.

MapReduce has been widely adopted for parallel data processing and has proven effective for applications such as log analysis, user activity analysis, and data mining. It offers a cost-effective and scalable solution for processing Big Data. However, its batch processing nature and slower turnaround time make it less suitable for ad hoc analysis and interactive data exploration. The need to restart jobs in case of errors further adds to the time and effort required for analysis tasks. In contrast, Dremel and BigQuery provide a more user-friendly and interactive approach to data analysis. By leveraging Dremel's fast query performance and BigQuery's intuitive interface, users can explore and analyze large datasets in a more efficient and timely manner, without the need for complex programming or lengthy job execution.

The limitations of MapReduce in terms of speed and agility become apparent when considering the experience of an AdWords API traffic analyst. By adopting Dremel (the technology underlying BigQuery) instead of MapReduce, the analyst was able to complete analytic tasks more quickly, often finishing them within a short timeframe. This showcases the advantages of tools like Dremel and BigQuery in facilitating faster and more interactive analysis on Big Data. Users can perform iterative and ad hoc queries with ease, enabling efficient and agile data exploration and analysis.

BigQuery and MapReduce serve different purposes in data processing. BigQuery is designed for structured data analysis using SQL queries, making it suitable for OLAP and BI use cases. It excels in quickly finding specific records, aggregating statistics with dynamic conditions, and enabling trial-and-error data analysis. On the other hand, MapReduce is a better choice for programmatically processing unstructured data, executing complex data mining algorithms, performing large join operations, and exporting substantial amounts of data.

Combining both technologies can result in a comprehensive solution. For example, MapReduce can be used for large join operations and data conversions, while BigQuery can handle quick aggregations and ad hoc data analysis on the resulting dataset. Alternatively, BigQuery can be used for preflight checks and quick data analysis, while MapReduce is employed for production data processing or data mining.

Traditional data warehouse solutions and appliances have been widely used for OLAP and BI use cases. These solutions include Relational OLAP (ROLAP) and Multidimensional OLAP (MOLAP). ROLAP relies on relational databases and requires building indices for improved query performance. However, indexing can become complex, expensive, and less efficient for large datasets. MOLAP involves designing and building data cubes or data marts based on predefined dimensions, providing fast results for specific queries.

However, it requires significant time and resources for cube design and can be inflexible in accommodating changes.

In contrast, BigQuery offers an alternative approach. Rather than relying on indices or pre-built data cubes, BigQuery leverages full-scan speed to achieve fast query performance. By accessing all records on disk drives without indexing or pre-aggregated values, BigQuery eliminates the need for upfront design and accommodates ad hoc queries and trial-and-error data analysis. The key to its performance is disk I/O throughput. While traditional solutions may use in-memory databases or flash storage to improve disk I/O throughput, they can be costly. In comparison, BigQuery leverages Google's infrastructure to provide high disk I/O throughput without the need for expensive hardware.

BigQuery stands out as a cloud-based interactive query service that offers high-performance analytics on massive datasets. It leverages two key technologies: columnar storage and parallel disk I/O. Columnar storage provides better compression ratios and disk I/O efficiency compared to row-based storage, making it ideal for data warehouse solutions. The parallel disk I/O capability allows BigQuery to achieve full-scan performance by leveraging Google's infrastructure, resulting in fast query execution times even without indices.

The cost-effectiveness of BigQuery is another significant advantage. Previously, achieving similar query performance required expensive data warehouse appliances or specialized database clusters. With BigQuery, businesses can execute full scans of billions of rows within seconds without the need for costly hardware investments. The pricing model is transparent, with a cost per query based on the amount of data scanned, making it a more cost-effective solution compared to traditional data warehouse solutions.

Importing big data into BigQuery is a two-step process: uploading the data to Google Cloud Storage and then importing it into BigQuery. This process can be executed via command-line tools, a web UI, or APIs, and BigQuery provides efficient data import capabilities, enabling the ingestion of large datasets within a reasonable timeframe.

Utilizing the Google Cloud Platform offers additional advantages. BigQuery is a fully managed service, eliminating the need for capacity planning, provisioning, and 24x7 monitoring. Google's experts handle operations and security patch updates, reducing the total cost of ownership for data handling solutions. The REST API provided by BigQuery allows developers to build interactive dashboards and mobile front-ends, empowering users to access meaningful data anytime, anywhere.