# GOOGLE DREMEL – BIG DATA ANALYTICS TOOL

## 1. Introduction

Google's Dremel system is a cloud-based big data analytics platform. It is designed to handle large data sets and provide interactive analysis of them - highly scalable and can process billions of records in seconds. Dremel was introduced while trying to deal with an encompassing problem - big data. Big data is a mammoth that must be dealt with in today's time when working on data processing and analysis. The key takeaway from Dremel is the nested columnar storage strategy (data is stored in columns rather than rows) which includes splitting, encoding, and assembly of big data. This nested strategy is very useful when dealing with data in a distributed file system, like Google File System (GFS), as it helps reduce the processing time from hours to minutes when compared to using a record-based storage strategy. Dremel is based on the MapReduce framework, has a distributed architecture, and can scale to petabytes of data, while being highly efficient and providing real-time results, due to its low latency. Google has used Dremel internally for several years and has made it available to select customers through its Google Cloud Platform [1]. It is currently used by many Google products, including Google Search, YouTube, and Google Earth.

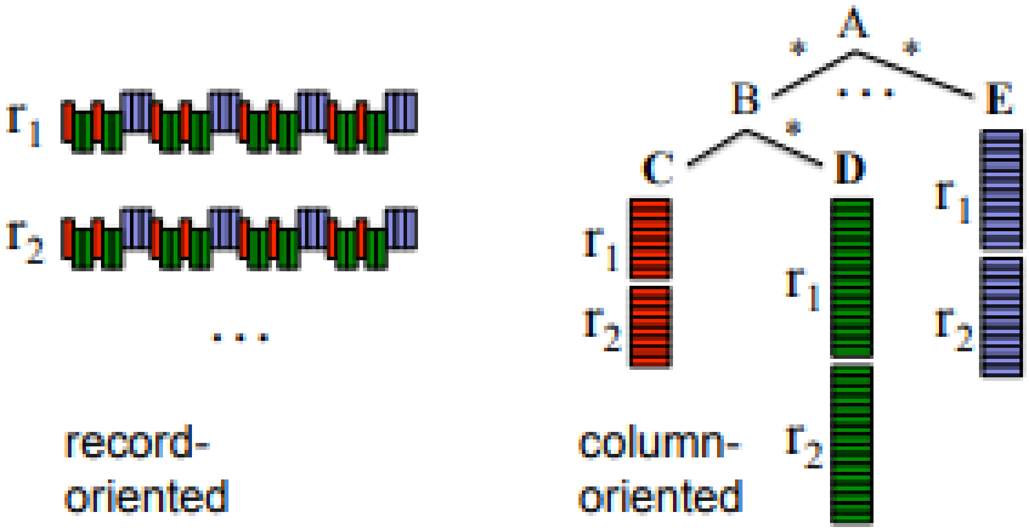

**Fig 1: Record-wise vs. columnar representation of nested data**

Dremel also supports SQL-like queries which makes it well-suited for analytics workloads, where queries often involve aggregating data across many columns and joining multiple relations.

Google's Dremel SQL Query Service is a cloud-based data analysis tool that enables users to run SQL queries on data stored in Google's BigQuery data warehouse. It is also designed to be fast, easy to use, and interactive with a tree architecture. Overall, Google's Dremel SQL Query Service is an excellent data analysis tool. It is also very scalable, making it an ideal choice for organizations with large data sets along with the ability to handle a variety of data types.

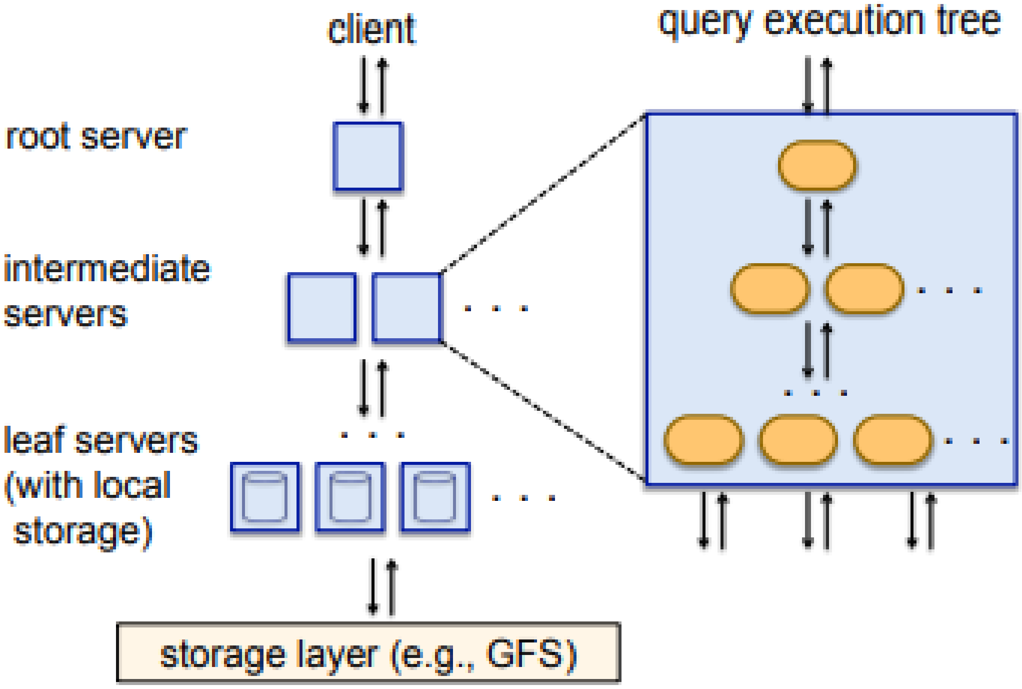

**Fig 2: System architecture and query execution**

## 2. Research

Research conducted by Sergey Melnik *et. al.* [2, 3] talks about the novel Dremel system, an interactive analysis tool for nested data, how it is used, its data model, querying language, and some experiments on the nested data. The authors perform multiple experiments using Dremel querying on different datasets containing up to TBs of data and trillions of records, the outcome of which helps establish Dremel as a great tool for processing big data. It has also been proven through these experiments that a MapReduce (MR) job when complemented with Dremel, helps reduce the execution time from minutes to seconds which is a very big advantage in the field. It is also discovered that leveraging some reference code from a Sawzall program and integrating it with Dremel and nested storage has further enhanced the processing power. The research also builds upon existing technologies, like nested storage, to provide novel algorithms for column stripping, record assembly, finite state machine creation, and query evaluation, which have proven to have assuaged the problem of dealing with big data.

The various experiments conducted, as part of the research, were on performance comparison of using a columnar vs. record-oriented storage strategy, fault tolerance by measuring Dremel's replication capability (as it's a multi-user system and hence needs to serve multiple queries at a given time), finding the performance enhancements after using Dremel in conjunction with MR jobs, comparison of query execution times as a measure of serving tree depths. Other experiments included measuring the performance of aggregation within records, the scalability of the system when dealing with trillions of records, and the impact of stragglers on the system. Each of these experiments used a different dataset that suited the needs of that experiment and helped build more confidence in the system. These well-defined experiments helped the readers understand the benefits of Dremel and nested columnar strategy better, along with some proof of its working.

## 3. Shortcomings

Aside from these well-developed experiments, some concepts which can be improved in Dremel are: During query execution, if a tablet takes a disproportionately long time to process, the query dispatcher reschedules it on another server, which can lead to some tablets being re-dispatched multiple times. This could potentially lead to increased processing times, which we are trying to optimize, and hence turn into a bottleneck. A suggestion would be to use a round-robin kind of approach for scheduling servers to reduce the time spent waiting for the original assigned server. For example, if server A is originally assigned to process tablet 1 and after X mins, it is unable to process Y% of the data, then the dispatcher should move the tablet processing to another available server B without waiting for a predefined time to realize server A would take disproportionately longer times.

There are some other limitations as well to the system. First, the system is not designed to handle real-time queries i.e., it may not be able to provide results in real-time for some types of queries. Second, the system is not designed to handle data that is constantly changing and may not be able to handle all data types equally well. Third, Dremel requires a lot of storage space and is not as fast as traditional relational databases. Finally, the system can be difficult to learn.

## 4. Conclusion

Despite of these shortcomings, I strongly believe that with the increase of big data in this fast-paced technological world, a tool like Dremel is needed to help reduce the processing time from hours to minutes to seconds, and to best utilize this tool, we need to perform valid and thorough experiments to test the system in a real-world setting. There is no limit to the potential scope of Google's Dremel. As the technology develops, it could be used for a wide variety of applications including:

- **Big data analysis:** Analyze large data sets to identify patterns and trends.
- **Business intelligence:** Generate insights about businesses and industries.
- **Scientific research:** Analyze data from scientific experiments and research studies.
- **Weather forecasting:** Analyze weather data to improve forecasting accuracy.
- **Disease detection and epidemiology:** Analyze health data to identify disease trends and track the spread of epidemics.

The possibilities are endless. As Dremel continues to evolve, it will become an increasingly powerful tool that can be used for a wide range of applications.

# REFERENCES

[1] Metz, Cade. “Google's Dremel Makes Big Data Look Small.” Wired, Conde Nast, 16 Aug. 2012, https://www.wired.com/2012/08/googles-dremel-makes-big-data-look-small/.

[2] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. 2010. Dremel: interactive analysis of web-scale datasets. Proc. VLDB Endow. 3, 1–2 (September 2010), 330–339. https://doi.org/10.14778/1920841.1920886

[3] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis, Hossein Ahmadi, Dan Delorey, Slava Min, Mosha Pasumansky, and Jeff Shute. 2020. Dremel: a decade of interactive SQL analysis at web scale. Proc. VLDB Endow. 13, 12 (August 2020), 3461–3472. https://doi.org/10.14778/3415478.3415568