DEPARTMENT OF COMPUTER SCIENCE
RAJAGIRI COLLEGE OF SOCIAL SCIENCES
(Autonomous)
KALAMASSERY - KOCHI – 683104

MASTER OF COMPUTER APPLICATIONS

SEMINAR REPORT
MCA403

NAME : ASHIQ S

SEMESTER : 4

REGISTER NO : 2224020

DEPARTMENT OF COMPUTER SCIENCE
RAJAGIRI COLLEGE OF SOCIAL SCIENCES
(Autonomous)
KALAMASSERY - KOCHI – 683104

MASTER OF COMPUTER APPLICATIONS

SEMINAR REPORT
MCA403

NAME : ASHIQ S

SEMESTER : 4

REGISTER NO : 2224020

DEPARTMENT OF COMPUTER SCIENCE

RAJAGIRI COLLEGE OF SOCIAL SCIENCES (AUTONOMOUS)

KALAMASSERY- 683104

CERTIFICATE

This is to certify that the seminar titled ***Navigating Insights With Google Bigquery*** *is a bonafide work carried out by* ***Ashiq S*** *in partial fulfillment of the requirements for the award of the Master of Computer Application degree of Rajagiri College of Social Sciences (Autonomous), affiliated to Mahatma Gandhi University, during the year 2022- 2024. This project report has been approved as it satisfies the academic requirement of seminar work prescribed for the Master of Computer Application.*

Priyanka E Thambi
Seminar Co-coordinator

Dr. Bindiya M Varghese
Dean- Computer Science

Examiner –I **(Seal)** **Examiner -II**

Place : Kalamassery

Date :

NAVIGATING INSIGHTS WITH GOOGLE BIGQUERY

ABSTRACT

Google BigQuery is a powerful tool that may be utilised by companies who work with a significant amount of data. Because it makes organising and analysing data much simpler, virtually anyone can use it. BigQuery is an easy-to-use tool that enables you to store, explore, and analyse your data fast and efficiently. This makes it much simpler to discover insightful information. Even extremely huge datasets can be processed very quickly. The simplicity with which one can get started using BigQuery is certainly one of its many strengths. To be proficient in the use of technology, you need not need an extensive knowledge base. Using BigQuery, it will be much simpler for you to manage your data and do research based on it. Able to make informed decisions and enhance your business procedures without needing to have a strong understanding of computers. Google BigQuery is secure, effective. It assists you in transforming unprocessed data into actionable insights.

INTRODUCTION

In this modern age where data is important, organisations have to deal with a huge amount of information all the time. It is very important to get useful information from this dataset so that you can make smart choices, improve operational efficiency, and gain a competitive edge. Google BigQuery is a powerful and cutting-edge data warehouse and analytics tool that is very important in this situation.

Google Cloud's main data warehouse platform is BigQuery, which was designed to handle large datasets effectively while providing exceptional speed and efficiency. When businesses use this technology, they can run complex analytical queries, get instant insights, and find hidden trends in their data without having to keep an eye on the basic infrastructure. BigQuery is a big change in the way data is analysed because it combines the freedom of a serverless system with the limitless power of the cloud.

BigQuery gives people who work in data science, analysis, and business leadership a lot of possibilities. It lets SQL queries run quickly, makes it easy to connect to AI and machine learning services, and offers scalable and cost-effective options for data warehouse needs. In this age of digital technology, BigQuery is an important part of using data successfully. It helps people turn their knowledge into useful insights and drives their organisation to achieve unimaginable heights.

COMPONENTS OF THE STUDY

- Introduction to Google BigQuery
- Data modelling and Schema Design
- Google Dremel Technology
- Data Integration
    - Create Dataset
    - Creating Tables
- BigQuery Batch Loading
    - Google Cloud Platform Console
- BigQuery View Authorisation

INTERPRETATION OF THE CASE

Data Modeling and Schema Design

Data modelling

In BigQuery, data modelling means planning how your data is organised to meet certain analysis needs. Defining tables, relationships, and entities to improve query speed is part of this. BigQuery works with both logical and physical data models, which gives you a lot of options for how to show complicated datasets.

Schema Design

Schema design is important for keeping data in BigQuery organised. Picking the right data types, making sure that stacked and repeated fields are optimised, and dividing tables in the right way can all have a big effect on how quickly queries run. BigQuery lets you show complicated, hierarchical data structures by supporting nested and repeated fields.

Google Dremel Technology

Dremel is part of Google's BigQuery service, a fully-managed, serverless data warehouse that enables super-fast SQL queries using the processing power of Google's infrastructure.

Data Integration

Create Dataset

A collection in BigQuery is a group of tables and views. When you create a dataset, you need to give it an ID and set any optional parameters, like the place and time that the dataset will expire. Datasets help you keep your tables organised and manage who can see them.

Create Table

Your info is stored in BigQuery tables. You create a schema that shows how your data is organised when you make tables. This includes giving names to columns, types of data, and any restrictions that are desired. To speed up queries, tables can be partitioned and grouped.

BigQuery Batch Loading

Google Cloud Platform Console

Tools like the Google Cloud Platform Console are often used to load a lot of data at once into BigQuery. Users can share and manage datasets through this web-based interface. It works with many file types, such as CSV, JSON, and Avro. Visualising and controlling the loading process is easy with the console.

BigQuery View Authorization

View Authorization

Views in BigQuery are like virtual tables that are set up by SQL searches. Controlling who can see these virtual tables is part of authorising views. You can make sure that people have the right access to the data by giving or taking away permissions at the dataset or view level. This is very important for keeping info safe and correct.

All of these things work together to make BigQuery a good place to organise and use data. Correct data modelling and schema design are the building blocks for fast queries. Data integration, batch loading, and view permission make sure that data processing goes smoothly and that only the right people can see insights. Knowing how to use the Google Cloud Platform Console makes managing information even easier and makes it easier to load many files at once. Understanding and using these parts will make using BigQuery for data processing and analytics more efficient and effective as a whole.

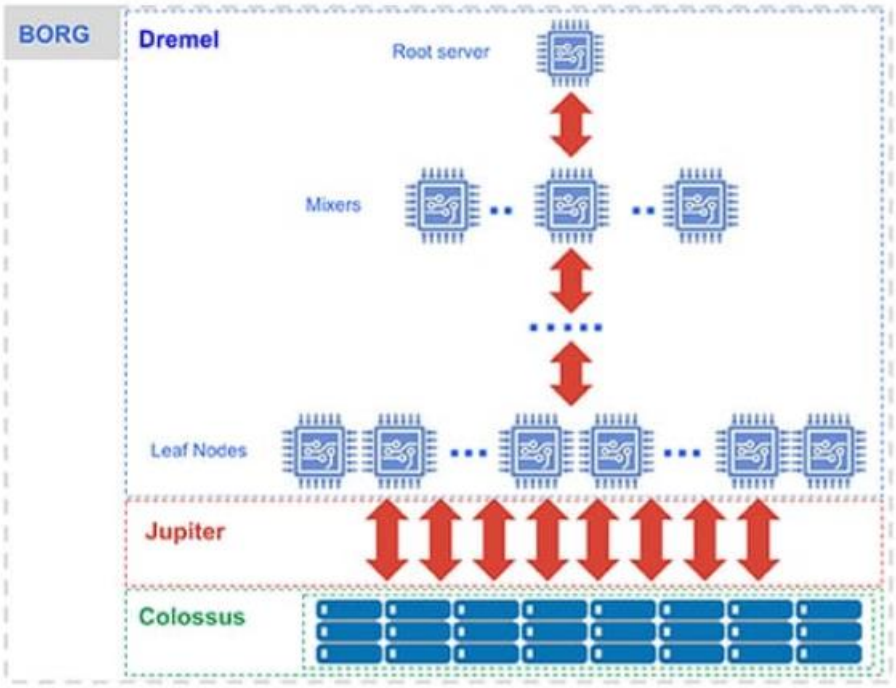
THE EXPERIMENT/TECHNOLOGY AND IMPLEMENTATION

The framework that Google BigQuery runs on is strong and designed to handle large numbers at high speeds and in parallel. There will be a thorough study that includes unique features like the use of a highly specialised SQL dialect and Google's own Dremel technology.

Understanding The Bigquery Architecture

BigQuery is Google's data warehouse system that runs in the cloud. Its serverless design kept storage and computing separate. It means that both storage space and computing speed can be increased on their own. This design gives users choices about how much storage they want and how much it costs. The Google cloud takes care of everything, so users don't have to worry about the structure underneath. It lets people focus on business issues even if they don't know much about systems. Google's distributed file system stores the data. It is very stable and can grow or shrink automatically based on the load. BigQuery interfaces, such as Web UI, REST API, CLI, and client-side languages, let users ask questions. First, the data is pulled from the storage. It then moves to the compute part, where it is grouped together and calculated. The two parts—storage and compute—are linked by Google's very stable network.

Dremel, Colossus, Jupiter, and Borg are just a few of the low-level tools that are working behind the scenes.



1.Dremel

The processing engine is what turns SQL queries into trees. Slots (the leaves of trees) let you read data from the store (Colossus). Mixers (branches) are used for aggregation. The leaf nodes link to the colossus through Jupiter, which is a durable network system. It also gives users slots on the fly, so the same person can get more than one slot for a query.

2.Colossus

Google's shared file structure is what it is. It gives users enough disc space. It also handles backups and restore in case the disc crashes. The data is stored in a columnar and compressed format that saves room and keeps the cost of running low.

3.Jupiter

It helps people talk to each other and is a stable network between computers and storage. It gives you a lot of bandwidth, which helps you spread out heavy tasks quickly and efficiently.

4.Borg

It's a method for managing a lot of clusters. It keeps things from going wrong, like when machines crash or power sources stop working, etc.

Dremel Technology

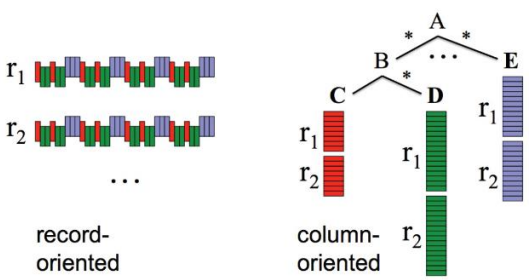
A key Google technology called "Dremel" is used outside of Google as BigQuery. With or without an index, Dremel can quickly go through a billion rows. Dremel is a massively parallel query service that runs in the cloud. It uses Google's infrastructure, which means that each query can be run on tens of thousands of computers at the same time.

Why does Dremel work so much faster?

- **Columnar Storage:** Data is kept in columns, which lets you get a very high compression ratio and scan throughput.
- **Tree Architecture:** This type of architecture sends questions to thousands of machines and collects their answers in a matter of seconds. Yes, it's a lot like Map-Reduce.

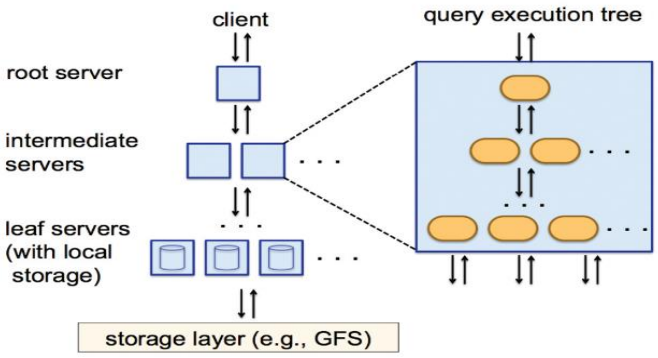
Columnar Storage

Dremel stores information in columnar storage, which means that it divides a record into column values and stores each value on a separate storage volume. This is different from how most databases store records, which store the whole record on a single volume.



Tree Architecture

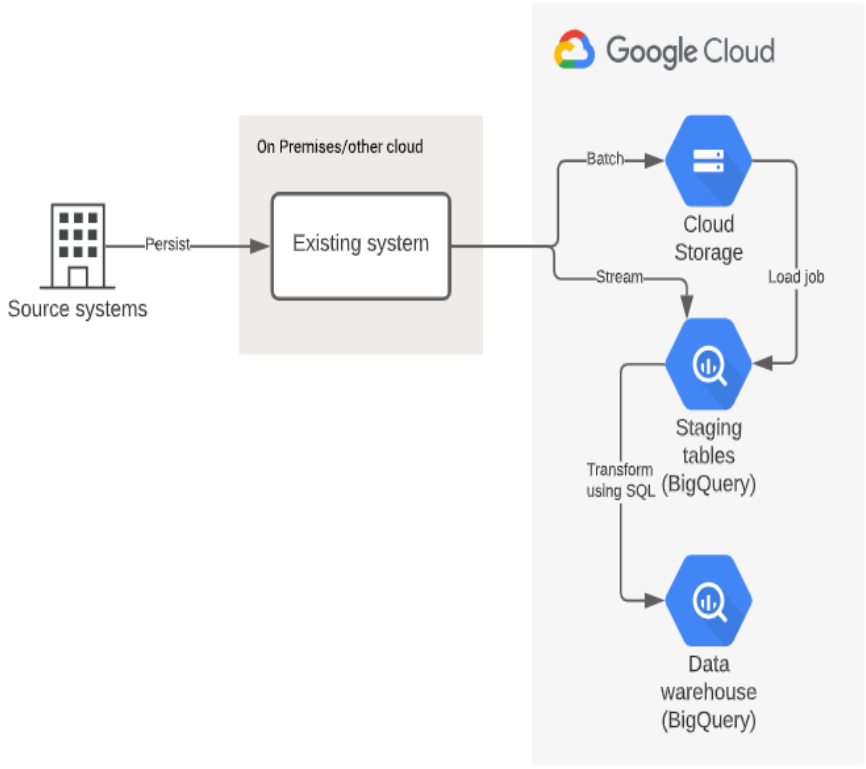
One hard part for Google in making Dremel was figuring out how to quickly send queries and gather data from tens of thousands of machines. Using Tree design helped solve the problem. The architecture sets up a massively parallel distributed tree so that a query can be sent down the tree and the answers from the leaves can be gathered very quickly.



DataFlow in BigQuery

Sources of data from outside the company: These are the different places where the info comes from. They can come from databases, the cloud, streaming info, or other places.

- **Data Ingestion:** means that data is brought into Google BigQuery from outside sources. Depending on the data source and needs, this could be either a batch process or a live process.
- **Google BigQuery:** This part of the computer does all the processing and stores and queries info. You can do the following in BigQuery:
- **Data Storage:** Data is saved in BigQuery tables. For better keeping and retrieval, these tables can be split up and grouped together.
- **Query Processing:** SQL queries are sent to BigQuery to be analysed and worked on.
- **Data Export:** You can send data from BigQuery to other places, like Google Cloud Storage, tools for visualising data, or other tools for analysing data.
- **Data Outputs:** This is where the results of searches and data exports from BigQuery are shown. The info that comes out can be used to make reports, make graphs, or do more research.
- **Users/Consumers:** These are the people or programmes that use Google BigQuery. They are made up of partners such as data scientists, data analysts, business intelligence tools, and more.



It's important to remember that BigQuery is mainly a service for storing and analysing data, so the main focus is on handling and analysing data. The DFD shows how data comes into and goes out of BigQuery. It does this by showing the external data sources, data storage, query processing, and data export parts in a simpler way. How complicated your DFD is will depend on how it is used and how it connects to other services and processes.

Methods Used To Build

As a complete and fully managed data warehouse system, Google BigQuery makes it easy to run fast and scalable SQL-like queries on large datasets. Even though the exact methods used to build it are kept secret, the system uses a variety of approaches and optimisation techniques to make queries run quickly. The basic ideas and methods that BigQuery uses are

- BigQuery uses a columnar storage format, which means that data is organised and saved between columns instead of rows. This format is better for analytical queries because it can compress data better and only access the columns that are needed for a given query, which cuts down on the number of input/output processes that need to be done.
- Data compression is used to reduce the amount of space needed for storing and the costs of sending data. A number of compression techniques have been added to BigQuery to make its storage more efficient.
- BigQuery uses techniques to cut down on the amount of data that needs to be sent between nodes during processes like joins and aggregations. This process involves moving and sorting data in a different way so that there is less network traffic.
- Cost-based query optimisation is a way for BigQuery to figure out which processing plan is best for a certain query. During the evaluation process, many methods are looked at, and the one that uses the fewest resources is chosen.

- BigQuery uses a cache to store and retrieve data that is used a lot. This cuts down on the time it takes to process routine searches. The implementation of caching has the ability to make query response times much more efficient.
- BigQuery's streaming updates make it possible for new data to be available for queries right away, which makes it ideal for real-time data processing. When it comes to real-time data research, this feature works out well.
- Materialised views are a feature of BigQuery that involve computing query answers ahead of time. These answers that have already been calculated can improve the performance of queries in certain situations. The views that are shown are always being updated based on the data that they are based on.
- Integrating machine learning methods is made easier by BigQuery, which works perfectly with Google's machine learning services, such as BigQuery ML. The features allow users to create and run machine learning models inside the BigQuery platform, which makes predictive analytics easier to use.
- Partitioning and Clustering: You can speed up the processing of queries by partitioning and clustering tables in BigQuery. Partitioning is the process of breaking up very large tables into smaller pieces that can be managed more efficiently. Clustering, on the other hand, puts the data in a certain order to make searches faster.

BigQuery turns out to be a strong tool for business information and data analytics. It is possible for users to focus on writing SQL searches while BigQuery handles the optimisation tasks in the background.

CHALLENGES

No doubt that Google BigQuery is a great tool for analysing data, but it does have some problems that need to be fixed. Some of the most usual problems that Google BigQuery users run into are

- **Safety of Data:** Protecting the privacy and safety of private data is very important. To keep their info safe, users must set up access controls, encryption, and authentication.
- **Query Performance:** BigQuery is known for being fast, but searches that aren't optimised well can slow it down. To make sure that data retrieval works quickly, query optimisation is always a problem.
- **Data Modelling:** Making the right data model is important for searching quickly and cheaply. Queries that are too slow or cost too much can be caused by bad data modelling.
- **Managing costs:** BigQuery has a flexible pricing plan, but costs can go up if they are not managed well. Costs need to be kept an eye on and managed, especially for big numbers and a lot of queries.
- **Data Ingestion and Integration:** It can be hard to move data into BigQuery, especially when the data comes from different sources and forms. Having good methods for importing and integrating data is important.
- **Complex Queries:** It can be hard to write and improve complex queries, especially ones that use a lot of joins and aggregations. To solve this problem, you need to know how to use the SQL syntax that BigQuery uses.
- **Data Import and Export:** Moving data into and out of BigQuery can be hard at times. You need tools and formats that are suitable to make the process go smoothly.
- **Data Governance:** It's important to set up policies and procedures for data governance to make sure that data is correct, consistent, and follows the rules.
- **Monitoring and Debugging:** Queries and data flows need to be effectively monitored and debugged so that problems can be found and fixed quickly.
- **Scalability and Concurrency:** Companies that need to analyse a lot of data should make sure that BigQuery can handle big workloads and multiple queries at the same time.
- **Data Retention and Storage:** Managing how long data stays in BigQuery tables can change how much it costs to store. When choosing how long to keep data, you should give it some thought.
- **Learning Curve:** People who are new to BigQuery need to go through a learning curve to get to know its benefits, how to use it most effectively, and what it can do.

Even with these problems, Google BigQuery has a lot of answers and best practises that can help users deal with them well. A lot of these problems can be fixed with good planning, optimisation, and ongoing management.

INFERENCES

BigQuery is a fully managed, serverless data warehouse service that lets you run SQL-like queries on big datasets quickly and on a large scale. A few important conclusions about BigQuery can be drawn from this:

- **Query Optimisation:** BigQuery uses methods for query optimisation to make your queries run faster. Things like cost-based query optimisation, simultaneous processing, and caching are all part of this. The service automatically tweaks how your SQL searches are run so that they return results as quickly and cheaply as possible.
- **Partitioning and Clustering:** How your BigQuery data is organised. You can divide and group your tables based on certain categories. When filtering by a partition key, partitioning makes it easy to get rid of unnecessary data, and clustering speeds up queries by putting data in a certain order.
- **Automatic Indexing:** BigQuery may use automatic indexing to speed up searches by making and keeping up to date indexes on your tables. These indexes can make queries much faster, especially when filtering through big datasets.
- **Cost Estimates:** BigQuery gives you cost estimates for your queries, so you can know how much they will cost before you run them. This feature helps you keep track of your query prices well.
- **Where and how to store data:** BigQuery lets you store data in different places, and it instantly finds the best place to store and process the data based on where it is stored. Inference about where data is stored and where it is located is necessary to make sure data is available and to reduce delay.

You can get the most out of your data analytics jobs while keeping costs low. It is a strong tool for looking at big sets of data, and it has built-in optimisations that help it work as quickly as possible.

FUTURE SCOPE

Google BigQuery has become a more stable cloud-based data warehouse and analytics tool within Google Cloud Platform. Although particular new developments aren't available, conversations about BigQuery's possible future in the context of data analytics and cloud computing as a whole show a number of important trends. People are using the platform more and more, and that's because Google Cloud is always working to connect BigQuery to new services and products, especially those that deal with AI, machine learning, and live data in real time. Improvements are expected, such as better analytical tools that go beyond basic data warehousing and include advanced analytics, machine learning, and prediction modelling. The serverless design and scalability of BigQuery are likely to stay popular, helping businesses deal with growing datasets. Because data governance and security are becoming more and more important, future updates may include tools for encryption, access control, tracking, and compliance. Google BigQuery's future is being shaped by a lot of changing factors, such as the ability to use multiple clouds and hybrid setups, tools for lowering costs, industry-specific solutions, the growth of data centres around the world, and the ability to share data freely.

CONCLUSION

Google BigQuery is a strong and flexible tool for storing and analysing data. Its scalable infrastructure, cloud-native design, and real-time processing make it the best choice for businesses that want to get useful insights from huge datasets. It is easier for both new and experienced data workers to use because it works well with other Google Cloud services and has a querying language that is similar to SQL.

BigQuery is a cost-effective option for businesses of all sizes because it can handle huge datasets very quickly and let you choose your own pricing conditions. It has strong security features, like fine-grained access controls, that keep personal data safe and private.

As the field of data analytics changes, Google BigQuery stays at the top by providing a flexible tool that can meet the ever-evolving needs of modern businesses that rely on data. As companies try to figure out how to work in this data-driven world, Google BigQuery is a great tool that helps them with things like real-time analytics, machine learning, and complicated data modelling. It is Google BigQuery's constant creativity and ability to grow that will shape the future of data analytics as technology improves and data becomes more complicated.

APPENDIX

1. Understanding BigQuery: Architecture and Use Case - (analyticsvidhya.com)
2. BigQuery overview | Google Cloud
3. Query a public dataset with the Google Cloud console | BigQuery
4. Create and use tables | BigQuery | Google Cloud
5. ETL with Dataflow & BigQuery. Extract, Transform and Load using… | by Suraj Mishra | Analytics Vidhya | Medium

NAVIGATING INSIGHTS WITH GOOGLE BIGQUERY

ORIGINALITY REPORT

5%	5%	4%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ann Baby, A. Kannammal. "Network Path Analysis for developing an enhanced TAM model: A user-centric e-learning perspective", Computers in Human Behavior, 2019 Publication	2%
2	Dan Sullivan. "Official Google Cloud Certified Professional Data Engineer Study Guide", Wiley, 2020 Publication	1%
3	E. Ravindran Vimina, K. Poulose Jacob. "Feature fusion method using BoVW framework for enhancing image retrieval", IET Image Processing, 2019 Publication	<1%
4	medium.com Internet Source	<1%
5	www.mongodb.com Internet Source	<1%
6	bmcprimcare.biomedcentral.com Internet Source	<1%