ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

AN ANALYSIS OF SNOWFLAKE: A PIVOTAL TOOL IN THE FIELD OF DATA ANALYSIS

Shweta More and Priya Daniel
Assistant Professor, SIES (Nerul) College of Arts, Science and Commerce

ABSTRACT
In today's world data is the center of every strategic business decision taken in all the industries across all the domains. It is one of the most valuable assets of a business, as it allows all the stakeholders to take informed decisions after performing a thorough analysis of the data available. But as the data is generated in an exceptionally large volume, the businesses need to have tools that have features like high performance, scalability, and ease to use to get the most out of their voluminous data according to their business needs and in less computing cost and time. In this era of distributed computing, we have many cloud platforms and Software as Service providers available that are offering unlimited computation and storage resources on demand. This paper discusses Snowflake Elastic Data warehouse, one of the tools used in many organizations to solve their real-world data problems like handling large amounts of data with high performance. This paper discusses Snowflake architecture, data sharing and storage, data ingestion and transformation capabilities, analytics, and visualization along with its other features that makes this tool the most dominant in the organizations for solving real world data problems and handling substantial amounts of data with high-speed performance.

1. INTRODUCTION

The cloud platform service industries have gained momentum over the past few years, as its features like scalability and high performance with no infrastructure and maintenance overheads has made the process efficient for all the small, medium, and large-scale organizations. Snowflake is an emerging data cloud warehousing solution that provides on demand cloud resources and works on pay as you go model by saving the costs for idle time.

Traditional data warehousing solutions can't deal with the volatility of today's data as they were designed to run on small and static clusters with structured data sources like ERP applications [1]. There are many cloud infrastructure solutions like Amazon Web Services, Microsoft Azure and Google Cloud Platform and open-source big data warehouse's that overcome all the problems faced by traditional warehousing solutions.

Amazon Redshift is a cloud based Amazon data warehouse supported by Amazon Web Services, the company's current cloud infrastructure, and is extensively scalable. It can quickly scale up to meet shifting capacity requirements which can be expensive and complex for organizations with rapidly changing information needs. Google's Google Big Query is a cloud data warehouse that offers quick SQL queries and straightforward research of huge datasets. Big Query was first created to process read-only data and was based on Google's technology. The platform makes use of a columnar storage technique that enables considerably faster data scanning as well as a tree-like data structure [2].

Snowflake is a hybrid cloud native platform that works as a data warehouse and as a data lake. It is an analytical data warehouse provided as Software as a Service that is not built on existing databases or other big data platforms like Hadoop. Snowflake is faster and easier to use as compared to other traditional data warehouses. It has a unique architecture designed for the cloud and uses a new SQL database engine for query processing which makes it more competitive than others.

3. Architecture

Snowflake has a unique and innovative architectural design that is comprised of three layers: Database Storage Layer (Storage Layer), Query Processing Layer (Compute Layer) and Cloud Services Layer (Service Layer). The storage and compute layers are independent which gives the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

customers more flexibility in scaling up or scaling down according to their needs. The service layer allows data sharing in real time, and in secure and governed environment.

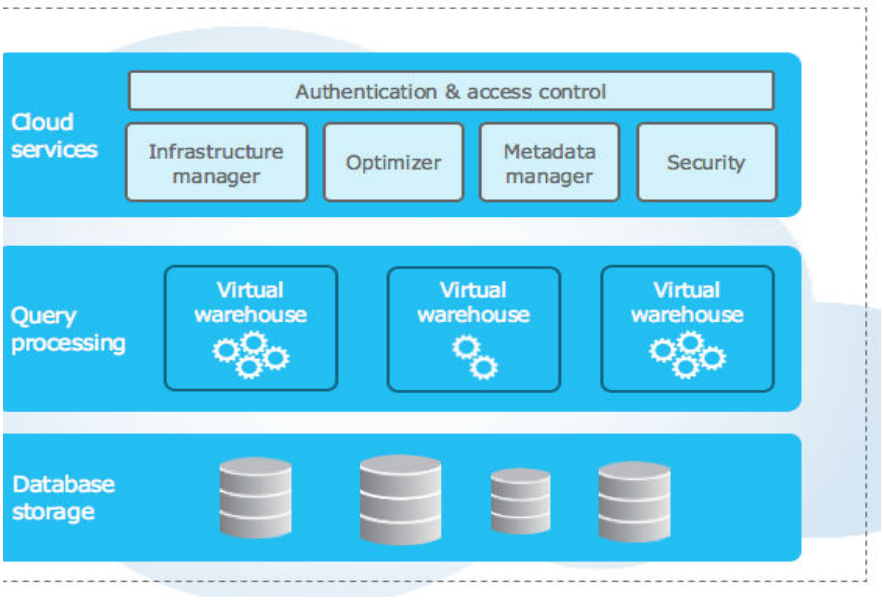

Figure 1: Snowflake Architecture

Database Storage Layer (Storage Layer)

This layer stores and manages all the data in Snowflake data warehouse. Whenever the data is loaded or imported into snowflake, it is micro partitioned into columnar format which is internally compressed, clustered, optimized and then stored in its cloud storage. This helps snowflake in effectively handling large tables by retrieving the data faster and improving the query performance. A micro-partition's data size ranges from fifty MB to five hundred MB.

The storage layer is independent of the compute layer, so it is effective as the organizations have to pay the price only for the data that they store irrespective of its computation. The data stored in snowflake cannot directly be accessed or viewed, it can only be done by using the SQL Queries and running them on the compute layer.

Query Processing Layer (Compute Layer)

The query processing layer uses the data from the storage layer in order to run the run the queries. Queries in snowflake are processed on a virtual warehouse as it a Massive Parallel Processing (MPP) platform. Each virtual warehouse is comprised of compute clusters made up of many nodes owning memory and CPUs. The client can select the cluster that they need as per their data and performance requirement. Snowflake warehouse sizes ranges X-Small having one cluster to 4X-Large having 128 clusters. So, if one needs the processed data in few minutes and data is in large volumes like Petabytes and more, the client can use cluster above size Large i.e. X –Large or 2X / 3X / 4X–Large varying on his needs and has to pay according to the computation pricing and time span required to complete the query, independent of the storage. The more the size, the more are the number of clusters, providing fast computation in less time and hence more is the cost. Snowflake supports the features of auto – suspending and auto – resuming along with auto – scaling.

Cloud Services Layer (Service Layer)

Service layer is the layer that manages all of the Snowflake's activities, including authentication, security, data management, and query optimization. A cloud service is a stateless computing resource

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

that utilizes readily available and useful data while operating across many availability zones. A SQL client interface is provided by the cloud services layer for data operations like DDL and DML. This layer maintains the metadata required to optimize a query or filter data. Query submissions to Snowflake must travel through this layer's optimizer before being directed to the Compute Layer for processing. User login requests must go through this layer in order to be processed by the cloud services layer.

3. Analysis:

As a Data Warehouse:

As a data warehouse, Snowflake is no different from other cloud data warehouses. It is a SaaS-based Cloud Data Warehouse constructed on top of either the Microsoft Azure or Amazon Redshift cloud architecture. The users are free from the menace of installing, configuring or managing hardware or software. However, snowflake architecture is designed differently than others that provides its users freedom to use storage, compute, and services, all of which are independent of others and fully elastic. It also allows users to write queries in Snowpark which is Snowflake's procedural programming language.

As a Data Lake:

A data lake is a repository of data that can either be structured or unstructured without any pre-defined schema. It is mostly used to store the data coming from the various incoming sources and it is stored as it is in raw and unprocessed format. Snowflake can also be used as data lake to store all the raw data of the organizations that can be used for data scientists or developers for all the required analysis especially when you need raw, transformed data. This data is accessible to all the users and is flexible as one can query the data that they need and matches to their analysis requirement or criteria anytime.

DATA LAKE BENEFITS

Data Access:
A distinguishing characteristic of Snowflake is democratic data access and straightforward, manageable data governance. The security measures and flexibility are intended to foster innovation. Snowflake's governance is built into the platform and includes object tagging for sensitive data for compliance, discovery, protection, and resource use as well as access control to accounts and users, column-level security, row access regulations, audit logging for access history.

Snowflake offers granular control over object access, allowing users to specify which objects may be accessed, what operations can be carried out on them, and who can create or modify access control policies[3].

It uses a combined access control mechanism from the following:

Discretionary Access Control (DAC): Each object has an owner, who can in turn grant access to that object [3].

Role-based Access Control (RBAC): Access privileges are assigned to roles, which are in turn assigned to users [3].

Data Ingestion and Pipeline Integrations:
A data pipeline is a method for transferring data from one location (the source) to another (such as a data warehouse). Data is optimized and modified along the journey, eventually reaching a stage where it can be examined and used to generate business insights.

A data pipeline is basically the process of gathering, arranging, and transporting data. Many of the manual processes required in processing and improving continuous data loads are now automated by

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

modern data pipelines. This typically entails loading raw data into a staging table for temporary storage, editing it, and then inserting it finally into the target reporting tables.

The "raw" data is frequently converted using a sequence of SQL statements before being inserted into the target reporting tables, usually after being temporarily loaded into an interim staging table. The workflow for this process that is most effective only transforms new or updated data.

Data can be loaded in bulk or continuous data loading can be done through Snowpipe, Snowflake Connector for Kafka and Third-party data integration tools

Building the pipeline into the platform improves speed, efficiency, and usability, just like most of Snowflake's capabilities. Engineers can manage workloads with little to no effort, making the system scalable to meet concurrency and computing needs.

For event-based real time ingestion into the table, Snowflake's Snowpipe enables importing data from files as soon as they are accessible in a stage.

Microsoft's Azure storage is directly accessible to Snowflake, it can ingest data from all of Amazon's AWS. Snowflake can directly ingest data from GCS.

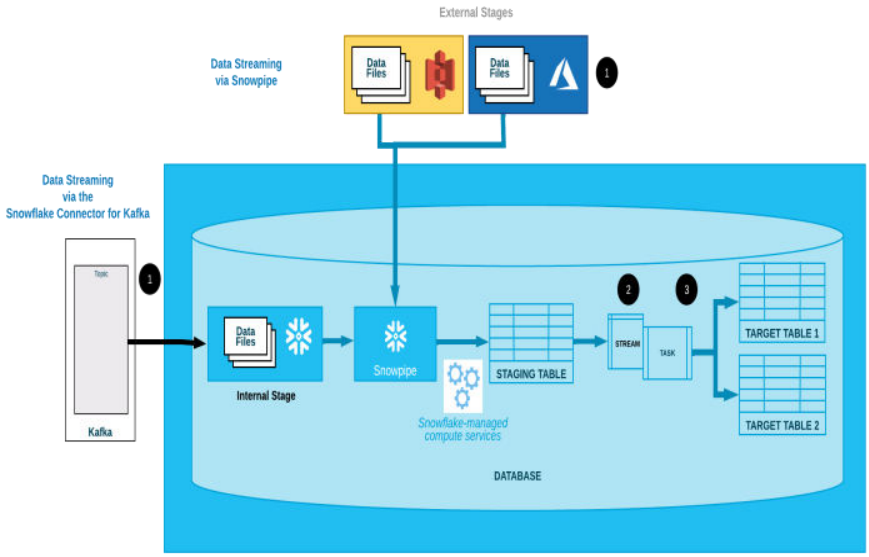

Figure 2: Snowflake Pipeline Workflow

Performance

With minimum setups or customization, Snowflake outperformed competing cloud data warehouses in head-to-head tests done by independent businesses in terms of query speed and associated costs. Because of this, Snowflake is practically a server-less solution.

Data Sharin

Data sharing is another foundational principle considered in the development of Snowflake. Data sharing is done without copying or relocating and fast cloning. Data sharing happens securely across Snowflake objects: tables, external tables, secure materialized views, UDFs, etc. With its cloning and extensive data sharing features, Snowflake is a fantastic option for businesses that need to exchange data with a wide range of partners, providers, or clients.

Data Analytics

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

Snowflake shows its Artificial Intelligence and machine learning capabilities by having built in integrations for Python, Spark, Java, Node.js, etc. It also supports AWS Sagemaker, Datalku and Data Robot. Snowflake has its own Snowpark that allows querying and processing data in a data pipeline securely. The Jupyter Notebooks or jar files can be directly accessed in Snowflake. It is compatible with many visualization tools like Power BI, Tableau, etc.

Pricing and Interoperability and Other Features

Snowflake follows pay as you go model, but what distinguishes it from others is its cost effectiveness due to its independent storage and compute layers. It is highly interoperable and supports data cloning and restoration, and cost-effective auto-scaling and VM sizing during SQL processing. Additional security levels are offered, including support for PHI data for HIPAA clients and encryption for all network connections. It is SOC 2 Type II certified.

4. CONCLUSION

A lot of the difficulties with traditional hardware-based data warehouses, such as limited scalability, challenges with data transformation, and delays or failures brought on by high query rates, are addressed with Snowflake, which is created expressly for the cloud. Snowflake shows high speed query performance as it is elastic in nature. It is a hybrid cloud solution that can store structured, semi structured or unstructured data as it can be utilized as data lake or data warehouse. Its architectural differences make Snowflake standout from all the other competitors. It has seamless data sharing, concurrency and availability. Snowpipe and Snowpark allow easy integrations and data loading. Snowflake also has its built in visualization capabilities, making it more competitive. Snowgrid, a recent addition to Snowflake features has enhanced cross cloud collaboration capabilities. Based on the capabilities and features listed, Snowflake stands as an emerging and pivotal leader in the field of data analysis by providing all solutions in one place to its clients, along with high performance, elastic nature and cost effectiveness.

5. REFERENCES

[1] B. Dageville, T. Cruanes, M. Zukowski, V. Antonov, A. Avanes, J. Bock, J. Claybaugh, D. Engovatov, M. Hentschel, J. Huang, et al. The snowflake elastic data warehouse. In SIGMOD, 2016.

[2] Raghavendra S., Hrithik Gautham T G, R Sindhu Rajendran, A comparative Study Between Hadoop and Snowflake. In International Research Journal of Engineering and Technology,Volume 7 Issue 5, 2020.

[3] Snowflake Documentation https://docs.snowflake.com/en/

[4] Dmitry Anoshin, Dmitry Shirokov and Donna Strok, Jumpstart Snowflake A Step-by-Step Guide To Modern Cloud Analytics. New York,Apress, 2020. https://www.projectpro.io/article/snowflake-architecture-what-does-snowflake-do/556

[5] A. Gupta, D. Agarwal, D. Tan, J. Kulesza, R. Pathak, S. Stefani, and V. Srinivasan. Amazon redshift and the case for simpler data warehouses. In SIGMOD, 2015.