Abstract for Data 603 Technical Research Paper

Abstract

We live in the age of distributed computing, and public cloud platforms have offered unlimited and flexible compute and storage resources for daily demand. With the Software-as-a-Service, such as Google App and Salesforce, the number of users is adapted to the newly cheaper, convenient, and fewer maintenance options. As unstructured data evolves, the demand for data analytics was increased. A traditional data warehouse is not able to handle well. For example, resizing of resources cannot be performed without compromising availability and performance. This means service users typically end up with either over-provisioned or under-utilized expensive resources to accommodate possible peak demand.

The cloud-based data warehouse was introduced as a modern solution for addressing the needs of data processing and reporting which requires flexibility and adaptability in both computing and storage resources. There are a few competitors in the cloud data warehouse market who can make the transition from the traditional star schema-based data warehousing to a modern elastic data warehouse platform possible: Snowflake, Amazon Redshift, and Google BigQuery.

Snowflake is an analytic data warehouse provided as Software-as-a-Service. Snowflake was built by a SQL database engine with its unique architecture for the cloud. Snowflake provides a cloud-based data storage and analytics service, and it allows the corporate users to store and analyze data using cloud-based hardware and software. Recently in 2020, Snowflake went public with its IPO listing on the NYSE which was the

highest value software IPO. Amazon Redshift is a fully managed, cloud-based, petabyte-scale data warehouse service by Amazon Web Services (AWS) which was initially released in 2012. It is an efficient solution to collect and store enterprise data and enables users to perform analyzation by using various business intelligence tools to acquire new insights for business and customers. On the other hand, Google BigQuery is a highly scalable, serverless multi-cloud data warehouse that is designed for business agility. It is a Software as a Service that supports querying using ANSI SQL with a built-in machine learning capability.

Although the three data warehouses have their similarities, they have the differences that make them unique. A business may have difficulties to choose the best cloud-based data warehouse that can fulfill their needs. This paper will describe and compare the architecture, cloud/on-premises options, storage and compute, scalability, and security of the three cloud-based data warehouses. The paper will also focus on comparing the databases' performance, maintenance, pricing as well as application recommendation for business.

Potential References:

1. https://www.abhishek-tiwari.com/rise-of-elastic-data-warehouse-and-database-services/

2. https://panoply.io/data-warehouse-guide/data-warehouse-concepts-traditional-vs-cloud/

3. https://medium.com/@richiebachala/snowflake-redshift-bigquery-b84d2cb60168

4. http://pages.cs.wisc.edu/~remzi/Classes/739/Fall2018/Papers/p215-dageville-snowflake.pdf

5. https://statsbot.co/blog/modern-data-warehouse/

6. https://medium.com/2359media/redshift-vs-bigquery-vs-snowflake-a-comparison-of-the-most-popular-data-warehouse-for-data-driven-cb1c10ac8555

7. https://aws.amazon.com/redshift/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

8. https://www.snowflake.com/

9. https://cloud.google.com/bigquery