**Data Warehousing and Business Intelligence**

**INFO 7260**

**Fall 2020**

**HW06a - Research Cloud Databases**

**By**

Abhay Kurian 001491312

Nikhil Kukreja 001055376

Shivam Verma 001344145

Smit Patil 001378302

**Amazon Redshift**

Amazon launched Redshift in November 2012, which helped on-premises data storage to data storage on the Cloud. In the Cloud, Redshift is a completely controlled, petabyte-scale data warehouse operation. Redshift is compatible with the client's data lake, has up to 3 times higher performance than any other data center, and costs up to 75 percent less than any other data warehouse in the Cloud.

Steps for creating a Redshift data warehouse:

1. Launch a set of nodes called Amazon Redshift cluster
2. Provision cluster and upload data set
3. Perform data analysis queries

**Features:**

1. **Query data:** Using familiar ANSI SQL, the user can query free file formats like Parquet, ORC, JSON, Avro, CSV, and more directly in S3 and export data to the data lake by simply using the Redshift UNLOAD command in the SQL code.
2. **Efficient storage:** On databases varying in size from gigabytes to petabytes, Amazon Redshift provides quick query efficiency. The amount of I / O needed to conduct queries is reduced by columnar storage, data compression, and zone maps.
3. **Concurrency**: Even with thousands of concurrent requests, Amazon Redshift offers reliably fast performance, whether they query data in your Amazon Redshift knowledge warehouse or directly in your Amazon S3 data lake.

**Pricing:** Amazon provides Redshift for a two-month free trial. For just $0.25 per hour with no obligations, you can start cheap, then scale out for just $1000 per terabyte per year. Amazon Redshift is the only cloud data warehouse that provides on-demand pricing at no upfront prices, pricing of reserved instances that can save you up to 75% by subscribing to a 1- or 3-year period, and pricing per question depending on the volume of data scanned in your Amazon S3 data lake.

**Move data to and from:** When moving data to and from an Amazon Redshift cluster, AWS Glue jobs issue COPY and UNLOAD statements against Amazon Redshift to achieve maximum throughput.

**Performance:** To make substantial benefits, Redshift provides numerous ways of tuning both the queries and data structure. Redshift contains SORTKEY and DISTKEY clauses that can be used while the data is set up. These clauses may offer a substantial reduction in running times for queries containing JOINs and WHERE clauses if used effectively.

**Scaling:** Redshift provides a function called concurrency scaling, which, when there is an increase in concurrent read query load, will automatically increase the cluster power.

**Snowflake**

Snowflake is a cloud-based data-warehousing company that was founded in 2012. Snowflake offers cloud-based storage, analytics, and an MPP system (Massively Parallel Processing). It runs on AWS, Microsoft Azure, and Google Cloud Platform. It allows users to store and analyze data using cloud-based hardware and software technologies. Its Snowflake Data Exchange allows customers to exchange data securely between snowflake users. By separating storage, it enables multiple warehouses to access the data source simultaneously. You can interact with Snowflake DW through web browsers, command line, analytics platforms, or Snowflake ODBC.

**Features:**

* **UNDROP command:** Snowflake has the UNDROP command, which helps recover from mistakes like dropping the wrong table. Traditionally much time is wasted in recovering the backup and restoring data. With Snowflake, using the UNDROP command helps to bypass all those steps.
* **Support for Semi-Structured data:** Snowflake can ingest semi-structured data in the format of JSON, Avro, ORC, Parquet, and XML and provides a special column type, VARIANT, which allows storing semi-structured data. Flattening nested structures in semi-structured data is accessible using functions like parse, extract, cast, and manipulate.
* **Continuous Data protection – Time Travel:** Snowflake maintains historical data through Snowflake Time Travel. You can recover historical data using this feature.

**Pricing:** Both Snowflake and Amazon Web Services (Redshift) have On-Demand payment where you only pay for what you consume. At the end of the month, a bill is generated with the details of usage for that month. There is a $25 minimum for every month, and for data storage, the rates are typically set to $40 per TB. Snowflake provides 30-70% discounts to companies that choose prepay.

**Moving Data: Secure Data sharing using Snowflake cloud account**: Secure Data Sharing enables account-to-account sharing of data through Snowflake database tables, secure views, and secure UDFs. So, you can share data with other users without ever having to copy any of the data.

**Performance: Optimized query through Micro-partitions and Clustered tables**: With Snowflake, you do not have to worry about indexing to speed up your queries. Snowflake automatically divides tables by grouping rows into micro-partitions of 50-500 MB of data. During micro-partition creation, clustering metadata is collected. This metadata of clustered tables is used to speed up the querying process.

**Scaling:** Snowflake has an Auto-scaling feature to meet demands greater than maximum capability. Snowflakes allow you to define a minimum and maximum cluster size, and Snowflake automatically scales between this range to meet demand increase.

**Google BigQuery**

Google BigQuery was launched in 2011 for general availability and is the corporate data warehouse for business agility developed by Google Cloud. BigQuery is a serverless, petabyte-scale, and cost-effective analytics data warehouse from Google Cloud that operates super-fast SQL queries near real-time over massive data volumes. With BigQuery, there is no framework to set up or maintain, and using regular SQL helps users generate useful insights. With on-demand and flat-rate solutions, BigQuery has customizable pricing models.

**Features:** Some significant features for the Google BigQuery are:

1. There is no cluster implementation, no virtual machines, no indexes or registry keys, no applications. Import the data and run SQL.
2. It is deployed by default across several data centers, with multiple replication factors to maximize optimum data longevity and operation uptime.
3. There is no need for several clusters to be distributed, and data duplicated into each one. Access control lists can manage permissions for projects and datasets.
4. For real-time analytics, streams millions of rows per second.
5. Separate compute and storage, compute scales based on usage, no need for cluster resizing.

**Data Storage:** BigQuery uses columnar storage, where each column contains a different block of data. Data can be conveniently appended to BigQuery tables, and current values changed or removed. Without restrictions, BigQuery supports mutations (INSERT, DELETE, UPDATE, MERGE). BigQuery uses a proprietary format. The storage engine will develop in parallel with the query engine to maximize query execution, which takes advantage of extensive data structure knowledge.

**Moving and Querying Data:** BigQuery can directly ingest datasets into its native storage from various formats. Google handles the native storage of BigQuery entirely, including backups, replication, scaling, and much more. One way for querying external data sources directly without being loaded into BigQuery storage is to use a federated query.

**Security:** The security paradigm of BigQuery firmly integrates with the rest of Google's Cloud Infrastructure. BigQuery uses Identity and Access Management (IAM) access control framework to grant special permissions to individual users or groups of users. BigQuery is also strongly linked to the Virtual Private Cloud (VPC) policy controls of Google, which defends against users seeking to access data from outside the company or distribute it to third parties.

**Pricing:** To satisfy both the technical requirements and budget constraints, BigQuery offers flexible, competitive pricing options. The cost of storage is dependent on the amount of data contained in BigQuery. Storage charges can also vary based on the frequency of use of the data. There are two pricing models for the query costs, one is on the amount of data processed by each query, and the other is a flat rate option.

**Microsoft Azure Synapse Analytics**

Microsoft's scalable, cloud-based data warehousing solution is Azure Synapse Analytics. It was launched to bring enterprise data warehousing and Big Data analytics together in 2019. Using either serverless or provisioned tools, users have the flexibility to query data on their terms. To ingest, plan, handle, and serve data for immediate BI and machine learning needs, Azure Synapse offers a seamless experience. Azure Synapse allows for both vertical and horizontal scaling of the data warehouse. Vertically by changing the service tier or placing the database in an elastic pool. Horizontally by adding more data warehouse units. Azure Synapse offers cloud data warehousing, machine learning analytics, and dashboarding in a single workspace.

**Features:**

* **Massively Parallel Processing –** Massively parallel processing (MPP) database technology is used by Azure Synapse, allowing it to effectively handle analytical workloads and aggregate and process large volumes of data. Unlike transactional databases that store rows as objects in a table, MPP databases store each column as objects. Data is often distributed through MPP databases through several nodes that function parallel to process various queries. This design of the database enables complex, long-running analysis.
* **Cloud-Native Hybrid Transaction/Analytical Processing (HTAP) Implementation –** To achieve real-time data integration with the Azure databases that make up operational database infrastructure, Azure Synapse Analytics uses "Synapse Connect" and HTAP implementation technology. The outcome is insights from live, operational data from real-time machine learning and business intelligence, without impacting existing operational systems.
* **Programming Language Compatibility** **–** A wide variety of scripting languages are compatible with Azure Synapse, including Scala, Python, .Net, Java, R, SQL, T-SQL, and Spark SQL. Synapse's compatibility with so many different languages makes it ideal for a broad range of analytics tasks.

**Pricing:** Synapse computes power in the form of data warehouse units (DWU). This allows your users to quickly and seamlessly scale compute resources at will. For serverless querying, data from data lake price is $5 per TB of data processed. Pricing is for provisioning resources starting at 100 DWU for $1.20/hour. Also, Data storage is charged at the rate of $122.88 per TB of data processed. SQL Database Threat Detection billed at $0.02/node/month.

**Performance:** Synapse facilitates query performance tuning and optimization via limitless concurrency, workload isolation, and workload management. It allows orchestration of data pipelines to separate historical data (into a data warehouse optimized for high-speed read operations) from real-time operational databases. Synapse can easily integrate with Microsoft and Azure ecosystems to provide seamless analytics. Like a data warehouse running on traditional SQL Server, it requires monitoring and tuning the distribution keys, indexes, caching, and partitions to ensure the best performance.

**Security and Privacy:** Synapse provides new security and privacy innovations, such as real-time data masking, dynamic data masking, always-on encryption, authentication via Azure Active Directory, authentication with a single sign-on, and automatic detection of threats. The platform also allows users to monitor column-level and row-level security access to sensitive data.

**Scaling:** Synapse SQL monitors resource usage by assigning concurrency slots to each query to ensure that each query has enough resources to execute effectively. Based on value and concurrency slots, the system puts queries into a queue. Queries wait in the queue until there are sufficient concurrency slots open. Based on importance, queries are queued and executed. If equal significance occurs, queries are performed on a first-in, first-out basis.

**Data Migration**: Azure Data Share can be used to share data, in any format, and any size, from multiple sources with other organizations. It also supports automated data sharing processes for productivity and predictability.

**Conclusion:**

A comparison is made between Amazon Redshift, Snowflake Cloud Data Platform, Microsoft Azure Synapse Analytics, and Google Big Query to find a leading cloud data warehouse for the Enterprise. Cloud data warehousing provides businesses of all sizes with benefits and flexibility in Data storage and management. Factors like Pricing, Performance, Data Security, Ease of moving data to the warehouse are considered for the comparison. Snowflake Cloud Data Platform has few advantages over other cloud data warehouses like it supports Multi-cloud, Massively Parallel Processing, Columnar, and Concurrency. However, different cloud data warehouses have their pros and cons, as discussed in the document. It depends on the Enterprise to incorporate a cloud data warehouse according to its need. To conclude, Snowflake Cloud Data Platform is an excellent Cloud data warehousing service in terms of power, the flexibility of Big Data platforms, and Cloud elasticity at a low cost.