SnowFlake and its APIs

Project Report for

CS 446 Cloud Computing

14-Dec-2020

Team Members

Asma Ahmed

Mhealyssah Bustria

Justin McCarthy

Mane Telpian

Table of Contents

1. Deliverables

Presentation Slides (URL) and Presentation/Demo Video (URL) can be found in this GitHub repo:

https://github.com/bustr003/cs446_snowflake

2. Executive Summary

2.1 What did we do?

During this project, we explored Snowflake and its APIs. The most important capabilities that we investigated are: data warehouse provided as a cloud service, data lifecycle, continuous data protection, architecture, automation, pay as you need, user interface, connecting with Snowflake, alternative competitors, and advantages of Snowflake.

2.2 What were our most interesting experiences?

Here are the most interesting/impressive aspects of the experience for each team member.

Asma Ahmed: What I found impressive about Snowflake is how easy they make it for users to access and make use of their data warehouse. With no installations or management done by users, it is very convenient and easy for customers to start using Snowflake.

Mhealyssah Bustria: It was interesting to learn that if a team is using a traditional data system instead of Snowflake, team members often have to wait for each other to finish before they can start their own workloads. Even when team members can run their workloads at the same time, they have to allocate the resources amongst themselves. It was interesting to learn how these issues are eliminated by Snowflake, and how big of an impact this could make on productivity.

Justin McCarthy: I was surprised that this technology was in such demand for something that seems relatively simple. I was also interested to see that snowflake

doesn't have any integrated support for visualizing data and relies on 3rd party tools instead. It would be nice to have the tools directly in Snowflake.

Mane Telpian: I was interested in learning that despite the fact that there are many alternatives to Snowflake, Snowflake has one of the most convenient costs and pricing methods. Moreover, data sharing is done very easily and seamlessly within the Snowflake organization and its customers.

3. Project Overview

3.1 Problem Statement

The purpose of this project is to explore Snowflake. We focused on four main questions when exploring this service:

1) What is Snowflake?

2) What features does Snowflake offer?

3) How can people use Snowflake?

4) How does Snowflake compare to its competitors?

Through these four questions, we aimed to learn more about what Snowflake can be used for, how people can use Snowflake, and what makes Snowflake different from other data services.

3.2 Project Team

Name of the project: SnowFlake and its APIs

Name of the Team member	**Responsibility**	**Contribution %**	**Notes**
Asma Ahmed	Background of Snowflake	25%	Section 4

Mhealyssah Bustria	Features of Snowflake	25%	Section 5
Justin McCarthy	Using Snowflake	25%	Section 6
Mane Telpian	Snowflake and its competitors	25%	Section 7

4.What is Snowflake

4.1 Key Concepts

Snowflake is an analytic data warehouse architecture that uses the Software-as-a-Service (SaaS) distribution model. This means that a third-party hosts the application and makes it available to users/customers over the internet. Snowflake is a software that provides its customers with a data warehouse that is easy to maneuver, fast, and more flexible than its competitors. Unlike other data warehouse softwares that are built on existing databases or “big data”, Snowflake uses a new SQL database engine, with its own unique architecture that is developed for the cloud. Although Snowflake has similarities to other enterprise data warehouses, it also has additional functionalities and its own unique abilities that make it different from its competitors.

4.2 Data Warehouse Provided as a Cloud Service

Snowflake makes it easy for users to access with:

- No hardware to select, install, configure, or manage
- No software to install, configure, or manage
- Maintenance and management are handled by Snowflake

With these benefits, users are able to make use of Snowflakes data warehouse much easier and faster without the hassle of downloading and installing packages to make it run on their device. Snowflake is also run completely on the cloud. This means that it will not take up any storage on your personal device. All of its services are run on a public cloud infrastructure and cannot be accessed by any private cloud infrastructures. Like mentioned before, all the software installations and updates are handled by Snowflake and are not packaged software offerings that can be installed by the user.

4.3 Data Lifecycle

In Snowflake, user data is shown as tables that can be aquired and adjusted with the use of standard SQL interfaces. Each table is formatted to belong to a schema where in turn belongs to a database. Data can be organized into databases, schemas, and tables. With Snowflake there is not a limit to how many databases you can create, or the number of schemas you can create within a database, or even the number of tables you can create within a table. The data can be inserted directly into the tables and DML is provided by Snowflake for the use of loading data into Snowflake tables from externally formatted files. Once the data is stored into a table, all the data can be accessed and modified by using SQL commands such as SELECT to query the data or DELETE to remove data.

4.3 Continuous Data Protection

Continuous Data Protection (CDP) is a feature in Snowflake that helps protect data that is stored in Snowflake. The CDP protects that data from human error, malicious acts, and any software or hardware failures that may occur. In any stage within the data lifecycle, Snowflake allows users to access and recover any data that might be accidentally or intentionally modified, removed or corrupted. Some of the features in the CDP include:

- Network policies that grant or restrict users access on the site based on their IP address
- Verification/ authentication is required for any users to access their account
- Security roles that control user access to any and all objects in the system
- Encryption of data and files
- Maintenance of historical data and failsafe

Most of these features are included in all Snowflake editions making it more accessible and safe users. These safety measures allow for users to feel protected while using Snowflake.

5. Key Features of Snowflake

5.1 Architecture: The Three Layers of Snowflake

Snowflake's architecture consists of three layers:

1) the centralized storage layer

2) the multi-cluster computing layer

3) the cloud services layer

Using Snowflake, team members are able to separately work on the same copy of data, while each workload will have its own compute cluster. This is possible because the computing layer is separate from the storage layer. The centralized storage layer allows all relevant compute nodes to access the data, and each node in the computing layer can process its own queries.

The cloud services layer assists with important aspects such as security, authentication, access control, query optimization, and management of infrastructure and metadata.

Users of the Snowflake service can deploy and manage all three of Snowflake's layers on a cloud platform. Users can host their Snowflake accounts on Amazon Web Services, Google Cloud Platform, and Microsoft Azure. However, there are some limitations when hosting a Snowflake account on GCP or Azure.

5.2 Automation of Administrative Tasks

Snowflake automates many tasks that otherwise would have to be done manually. The automation of tasks reduces "up to 80%" of the manual administrative tasks that humans have to do, which gives human users more time to focus on other meaningful tasks, such as analytics.

The automation of tasks is made possible by three key features of Snowflake: multi-cluster virtual warehouses, self-organizing cloud storage, and materialized views.

5.2.1 Automatic Scaling

The use of multi-cluster virtual warehouses allows automatic scaling.

Each workload has its own dedicated computer cluster, even if multiple workloads are accessing one data source. Since each computer cluster is independent, each cluster the size and scale of each cluster can be modified at any time, without affecting the other clusters. Because of the isolation of workloads, users are able to work on the same data without interfering with each other, and without having to wait for each other.

Having the resources isolated in each cluster prevents data administrators from having to spend time worrying about how much of various resources to allocate to each workload.

5.2.2 Self-organizing cloud storage

Automatic clustering allows self-organizing cloud storage.

Snowflake offers a feature called "cluster keys" so that users can take advantage of automatic clustering. Since clustering data can be automated with Snowflake, users do not need to spend time manually clustering the data. The automatic clustering feature is especially-useful when queries only need to be run on a smaller portion of a big table.

Snowflake continuously reclusters the data in the background, so the user does not need to do so manually. While taking advantage of this automation, users can also manually modify the cluster keys whenever needed.

5.2.3 Materialized Views

Snowflake's materialized views allow improved accuracy with less manual work.

Snowflake can continuously maintain the accuracy of materialized views, so that users do not need to manually update the materialized views as the data is modified. The resources required for this automatic maintenance is managed and scaled by Snowflake, but the user can resume and suspend the automatic maintenance as needed.

Furthermore, users can take advantage of the automatic clustering and materialized views features together to generate various representations of the same data. The

combination of these two features allows users to accurately view the current state of the database, while still only looking at the needed clusters.

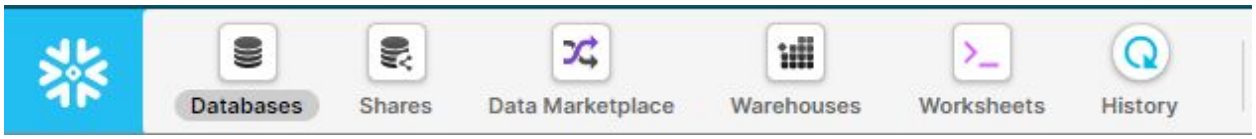
5.3 Pay as you need

With Snowflake, companies do not need to keep a data warehouse running at full capacity at all times. Doing so would cause the company to lose money because the warehouse is running even when it is not being used. Snowflake's use of virtual warehouses allows users to resume a data warehouse when it is needed, and suspend that warehouse when it is not needed. Many other features offered by Snowflake can also be resumed and suspended as needed, so companies only have to pay for what they need and how much they need.

6. Using Snowflake

6.1 User Interface Overview

Snowflake runs completely on cloud infrastructure, providing its services directly through a web-based interface. The following are the main components accessible through Snowflake's UI.



Databases

- Easily upload and store data. Snowflakes can handle data in many forms such as compressed or delimited data files and formats such as JSON and XML, automatically reorganizing for optimization. Snowflake is also well integrated with other cloud platforms making it easy to transfer data already stored in the cloud on AWS S3, Google Cloud Storage or Microsoft Azure.
- Snowpipe is a REST API that facilitates continuous bulk loading of data from internal or external files. For example, a pipeline can be set up that automatically transfers data in micro-batches from an Amazon S3 button to a table in Snowflake without having to manually upload each time something new is added.

Warehouses

- Warehouses are virtual compute instances that provide the resources required for manipulating and analyzing data stored on the Snowflake database. Snowflake provides a useful interface for managing these instances with usage/billing data, automatic scheduling of suspensions/resumptions and on the fly scaling of the warehouse cluster size.

Worksheets

- Worksheets provides an interface for creating and running SQL queries on your databases and viewing results. Worksheets can be saved for later use and easily shared between users.

Data Marketplace

- Provides instant access to a variety of datasets that you can join with your personal data. Users can apply as “Approved Data Providers” and upload their own datasets.

6.2 Connecting with Snowflake

Snowflake has provided developers with many options for building applications that can connect to and interact with Snowflake. They have built native programming interfaces for Go, .NET, Node.js, JDBC and ODBC which allow you to perform all standard operations done through Snowflakes web interface. An example of an application built with these drivers is SnowSQL. SnowSQL is a command line interface client built using the Snowflake connector for Python. It allows you to load, unload and manipulate the Snowflake databases and analyze data using SQL queries directly through a command line interface.

A wide variety of companies have utilized these tools to integrate Snowflake into their own applications that handle things such as data integration, business intelligence(BI), machine learning, data science, security/governance and SQL development/management.

We’ll take a closer look at the BI tools that integrate with Snowflake. These are mostly applications that help with data visualization. Companies can leverage the power of Snowflake to store and analyze their data and seamlessly receive the results in these BI tools for presentation. Snowflake has partnered with many of the leading BI applications to provide native connectivity such as Adobe, Looker, Sigma, Tableau and dozens of others. Several of these companies have also partnered with Snowflake to provide free trial accounts to test out the integration using Snowflake partner connect.

7. Snowflake and its Alternative Competitors

7.1 Overview of Alternative Competitors

As a data warehousing company, Snowflake has some main competitors including Amazon Redshift, Google BigQuery, IBM Db2 Warehouse, Cloudera, Databricks, Microsoft Azure Synapse, Oracle Autonomous Warehouse, Panoply, Yellowbrick Data, and more. The top alternatives as well as competitors used by most developers are Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, and IBM Db2 Warehouse.

7.2 Amazon Redshift

Amazon redshift by Amazon is one of the most used data warehousing solutions. It is a fully-managed petabyte-scale cloud based solution, developed to deal with large data set storage and analysis, as well as large scale database migrations. Redshift allows users to scale to petabytes, yet allows them to start with gigabytes of data. The database is column-oriented, allowing it to connect to SQL based clients and business intelligence tools in order to allow data to be available to users at all times. The reason many companies and developers use Redshift is because of its speed. It can provide fast query speed on large data sets, which is due to its architectural components, columnar data storage and massively parallel processing design (MPP). Because Redshift uses columnar storage for database tables, it lessens disk I/O requirements, which optimizes analytic query performance. The amount of disk I/O requests and data that needs to be loaded from disk is decreased. The MPP design automatically distributes workload across the nodes of each data cluster, allowing even large amounts of data to process fast. The multiple nodes will process all SQL operations simultaneously. Many top companies including Intuit, Coursera, Coinbase, Pinterest, Lyft, Yelp, and McDonalds use Amazon Redshift.

7.3 Google BigQuery

Google BigQuery is part of the google cloud platform environment. It is a highly scalable and serverless cloud data warehouse. One of the key highlights of BigQuery is that it is cost efficient. It allows you to run an analytics environment with a three-year TCO which is around 34% cheaper than most cloud data warehouses. It can also be integrated with google machine learning tools. BigQuery runs on the google cloud Storage infrastructure and can be used with a REST oriented API. It uses columnar storage

which allows fast data processing, and a tree architecture for dispatching queries and aggregating results. Companies that use Google bigQuery include Spotify, The NY Times, Monzo, Target, PayPal, HSBC, and much more.

7.4 Microsoft Azure Synapse

Another competitor is Microsoft Azure Synapse. The product is in fact an updated version of the Microsoft Azure SQL data warehousing product. It is a massively parallel processing (MPP), scale-out database that processes large amounts of data in the Microsoft Azure cloud platform. It combines cloud data warehousing and big data analytics into a single service platform. It allows the user to query data using serverless on-demand resources or provisioned resources. Some other offerings include workload optimization, business intelligence and machine learning tool integration, Cloud-native HTAP integration, and security features such as automated threat detection, always-on encryption, access control of column and row levels, and dynamic data masking which can protect real-time data.

7.5 IBM Db2 Warehouse

Lastly, there is IBM Db2 Warehouse, a relational database that offers advanced data management and analytics for transactional workloads. Some advanced features included by the db2 database are in-memory technology, advanced management and development tools, storage optimization, workload management, actional compression, and endless data availability. Some top companies currently using Db2 warehouse include JP Morgan Chase, Morgan Stanely, Wells Fargo, Unitedhealth Group, and more.

7.6 Snowflake’s Advantages

Although these products also have their cons, Snowflake has many advantages. A convenient aspect of Snowflake is its pricing structure, which is much different from most other data warehouses. Snowflake allows per-second pricing, meaning that users only pay for what they use. Snowflake’s architecture allows the storing and computing to scale separately from each other, which allows users to use and pay for the storage and the computation separately. This is good for organizations with large amounts of data, as it decouples the storage and compute functions, meaning organizations that have high storage needs but not as much for CPU cycles don’t have to pay for a bundle with both, since Snowflake provides all in one.

In regards to performance and speed, Snowflake allows users to scale up the virtual warehouse and take advantage of extra compute resources. Then, the warehouse can be scaled down and the user would only need to pay for the time used.

Snowflake also deals with concurrency issues using its multicluster architecture. Queries between virtual warehouses don't affect each other, and each warehouse is able to scale up or down as needed. This allows users to get what they need whenever without having to wait for other loading and processing tasks to finish. Snowflake also allows data sharing easily between other Snowflake users, meaning organizations can share data seamlessly within each other and with data customers. Availability wise Snowflake can tolerate component and network failures with little to no impact to customers.

8. References

Title: An architecture built for the cloud
URL: https://www.snowflake.com/product/architecture/
Author: Snowflake
Access date: 12/05/2020

Title: Key Concepts and Architecture
URL: https://docs.snowflake.com/en/user-guide/intro-key-concepts.html
Author: Snowflake
Access date: 12/05/2020

Title: Supported Cloud Platforms
URL: https://docs.snowflake.com/en/user-guide/intro-cloud-platforms.html
Author: Snowflake
Access date: 12/07/2020

Title: How Snowflake Automates Performance in a Modern Cloud Data Warehouse
URL:
https://resources.snowflake.com/ebooks/how-snowflake-automates-performance-in-a-modern-cloud-data-warehouse
Author: Snowflake
Access date: 12/08/2020

Title: Snowflake documentation
URL: https://docs.snowflake.com/en/
Author: Snowflake
Access date: 12/10/2020

Title: Top 10 Cloud Data Warehouse Solution Providers

URL:
https://em360tech.com/data_management/tech-features-featuredtech-news/top-10-cloud-data-warehouse-solution-providers
Author: Enterprise Management 360
Access date: 12/09/2020

Title: Overview of Data Lifecycle
URL: https://docs.snowflake.com/en/user-guide/data-lifecycle.html
Author: Snowflake
Access date: 12/12/2020

Title: What is Amazon Redshift?
URL: https://www.sumologic.com/blog/what-is-amazon-redshift/
Author: Kevin Goldberg
Access date: 12/12/2020

Title: What is Azure Synapse Analytics
URL: https://www.matillion.com/resources/blog/what-is-azure-synapse-analytics
Author: Matillion
Access date: 12/11/2020

Title: IBM Db2 Database
URL:
https://www.ibm.com/uk-en/products/db2-database?p1=Search&p4=43700052261275366&p5=b&cm_mmc=Search_Google-_-1S_1S-_-EP_GB-_-%2Bdb2_b&cm_mmca7=71700000064548121&cm_mmca8=aud-311016886972:kwd-302204620467&cm_mmca9=Cj0KCQjw4dr0BRCxARIsAKUNjWQCb-eJQMVfrcsNq9CpfFEqWWrwug6RDBQgo6AF5alDaEH9_gl4fuEaAtUDEALw_wcB&cm_mmca10=423579257697&cm_mmca11=b&gclid=Cj0KCQjw4dr0BRCxARIsAKUNjWQCb-eJQMVfrcsNq9CpfFEqWWrwug6RDBQgo6AF5alDaEH9_gl4fuEaAtUDEALw_wcB&gclsrc=aw.ds
Author: IBM
Access date: 12/12/2020

9. Glossary

API - (Application Programming Interface) Interface that defines the interaction between software systems.

BI - (business intelligence) Techniques used by companies to analyze data and to make business decisions.

CDP - (Continuous Data Protection) Feature on Snowflake

DML - (Data Manipulation Language) Language used to manipulate Oracle Database table data

JDBC - (Java Database Connectivity) An API for accessing database management systems.

MPP - (Massively Parallel Processing) Large number of computer processors simultaneously performing a set of computations at the same time.

ODBC - (Open Database Connectivity) An API for Java that determines how an client can access a database

SQL - (Structured Query Language) Language used to communicate with a database.