International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I6P107

Volume 71 Issue 6, 42-46, June 2023

Review Article

How Snowflake is Transforming Data Science

Priya Sengar

Data Scientist, Old Dominion University

Received: 01 May 2023 Revised: 29 May 2023 Accepted: 12 June 2023 Published: 30 June 2023

Abstract *- In a fast-growing and evolving business environment, companies globally are looking for ways to enhance their data governance capabilities. The advent of Snowflake as a rich and evolving data cloud service allows offering-controlled access to cutting-edge tools, apps, and services, as well as nearly endless amounts of data. With the help of the Data Cloud, companies work together locally and internationally to uncover novel insights, develop previously unanticipated business prospects, and recognize their customers at any time through smooth and pertinent interactions. Implementing Snowflake as a data cloud service enhances efficiency, scalability, security, and uniformity. This article emphasizes how Snowflake transforms data science by acting as a unified data hub for varied types of data blended with a rich ecosystem of partner products and native connectors where users can leverage their choice of machine learning tools to Snowflake data. The article also discusses how prominent companies across different domains are using Snowflake as their cloud data service to scale and improve their data and ml ecosystem. Although Snowflake provides rich benefits, this article also demonstrates its limitations, such as unavailability on-premises, limitations on high query performance, price elevations, not access to granular support, and difficult adaptation. By overcoming these challenges, companies do have a significant ability to adopt the technology with careful consideration.*

Keywords *- Data engineering, Machine learning, Snowflake, Snowpark.*

1. Introduction

Snowflake's Data Cloud is powered by a state-of-the-art data platform made available as a self-managed service. It is possible to create data storage and statistical solutions with Snowflake that are quicker, simpler, and more adaptable than those that are now accessible. No existing database or "big data" software framework, including Hadoop, is used to build the Snowflake data platform. Being a truly managed self-service platform, Snowflake does not require users to choose, install, configure, or manage any hardware. Almost no software needs to be installed, set up, or maintained. Snowflake is responsible for all ongoing management, maintenance, updates, and tuning. Snowflake uses a central data repository for persisting data accessible from all compute nodes in the platform, much like shared-disk systems. Snowflake uses virtual computing instances to meet its computing demands and store persistent data. It uses a storage service. Private cloud infrastructures (hosted on-premises or otherwise) cannot be used to run Snowflake. It is not possible for a user to install Snowflake because it is not packaged software. Snowflake handles the entire software installation and update process. As soon as data is loaded into Snowflake, the software rearranges it into a columnar, optimized, and compressed format. This streamlined data is kept in the cloud using Snowflake.

Companies use machine learning to educate a system on how to use machine learning algorithms to solve the current problem and improve over time.

The exponential boom of Data Science includes machine learning as its significant component. Using different statistical techniques, algorithms are trained to produce classifications and regression predictions to identify meaningful insights from the vast data pool. These insights produce quality decisions that drive the different growth indicators across industries.

Huge investment is being made in ML enterprises across a vast number of industries, including education, retail, banking, health care, government, and telecom. Machine learning is important because it helps create new products and gives organizations an overview of consumer behavior trends and operational business patterns. According to a recent study, spending on machine learning is anticipated to increase from roughly $1.58 billion in 2017 to $20.8 billion in 2023 [1]. The manner in which an algorithm learns the patterns of data to improve its accuracy is a common way to segregate different types of machine learning.

There are four different types of machine learning: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Depending on the data and the final outcome which has to be predicted, Data Scientists determine the respective Machine Learning algorithm.

In supervised learning, the output is anticipated by the algorithms using labeled training data that has been used to train the machines.

In unsupervised learning, a model must search for patterns in a dataset without labels and with little human supervision.

Between supervised and unsupervised learning, there is a type of machine learning known as semi-supervised learning. It is a technique that trains a model using both a sizable amount of unlabeled data and a modest amount of labeled data.

Reinforcement Learning focuses on maximizing the reward for a specific business problem. The most appropriate course of action to take in every scenario is identified by using an array of programs and machines. In reinforcement learning, there is no right or wrong answer; instead, the reinforcement agent chooses how to carry out the job.

Snowflake for data science was created from the ground up to serve applications driven by machine learning and AI. In addition to being tightly integrated with Spark, R, Python, and Qubole, Snowflake is a crucial Data Science technology. Supporting reliable machine learning models depends on performance speed. Scaling up or down is possible with Snowflake. Apart from making its mark in Machine learning, Snowflake has transformed the feature engineering process and data preparation capabilities by using smooth feature engineering using ANSI SQL, external functions to orchestrate Data pipelines, leveraging streams and tasks for effective Data transformations, and CDC.

Snowflake's Zero-copy cloning features allow the creation of a full copy of data in production within seconds at no cost.

In order to carry out memory-exhaustive operations like mathematical analysis, data transformations, training, and prediction within Snowflake at scale, we have Snowpark, which is a developer framework, and Snowpark-optimized warehouses with nodes having 16x and 10x local cache hence boosting the streamlining of ML pipelines [2].

2. Snowflake For Data Science

2.1. Central Repository for Data

Snowflake acts as a single point of access to a worldwide network of reliable data, reducing the time spent searching for and requesting access to data. With native support for organized, semi-structured, and unstructured data, Snowflake practically sorts all the data into the machine learning model without building complicated processes. By aggregating data from varied environments, Snowflake consolidates all the data into a unified extensive-computing platform reducing the inactivity caused by ETL jobs. Snowflake offers the ability to profile and clean data directly, guaranteeing a high quality of data consistency. Additionally, Snowflake offers data discovery tools so customers can find and access their data more quickly and conveniently. With Snowflake Data Marketplace performance of machine learning models can be improved by using shared data sets from a company's ecosystem and outside data.

2.2. Snowflake Marketplace

The Snowflake Marketplace enables data consumers to access and query available data sets and services to drive informed data-driven decisions. Data providers can publish data sets or make data analytics services available to Snowflake customers. Data consumers securely access life and governed shared data sets directly from their Snowflake account and receive automatic updates in real-time [3].

Following are the different categories of data supported by Snowflake:

- Structured Data: Relational Database and Spreadsheets
- Semi-Structured Data: JSON, Avro, Orc and Paraquet
- Unstructured Data: Pdf documents, Audio files, medical images, videos, and emails

2.3. Faster Speed With Reduced Complexity

Snowpark provides data science professionals to transform data into meaningful ML insights by using their language of choice and using Anaconda integration equipped with pre-installed open-source libraries and effective dependency management to speed up the performance of Python–based workflows. Snowflake's flexible and multi-cluster architecture scales up processing for huge volumes of data and users.

2.3.1. Snowpark

A user-friendly library for searching and processing data at scale in Snowflake is offered via the Snowpark library. Snowpark library allows the creation of applications that process data in Snowflake using a library for any of the three languages without transporting the data to the system where your application code is executed. These applications can process data at scale as part of the elastic and serverless Snowflake engine. At the moment, Snowflake offers Snowpark libraries for Java, Python, and Scala.

Following are the three key components of Snowpark for streamlining the ML architecture:

- Dataframe Api: Snowpark leverages well-known Data Frames to create queries and data transformations. In order to scale out processing in Snowflake, operations are transformed to SQL.
- User-Defined Functions (UDF): It is a personalized program created by the user in native Java, Python, or Scala code, and it may be applied on top of any built-in function hence enabling engineers to work with massive databases more effectively by reusing scripts or programs that can carry out particular functions rather than having to write them from scratch each time. Data engineers need to pass in the parameters for their

particular use case when calling these functions as needed [4].

- Stored Procedures: A stored procedure is a piece of code that encapsulates logic from SQL statements to carry out database operations. A stored procedure is often used to carry out automation tasks that are run more frequently. Snowpark stored procedures provide the leverage to build and execute the data pipeline within Snowflake by harnessing the Snowflake warehouse as the source of computation. Snowpark API for Java, Python, Scala, and JavaScript is used to build the data pipeline code required for executing the Stored procedure.

2.4. Unification of Teams and Tools

Snowflake utilizes sophisticated integrations from a large ecosystem of partner products and native connectors to link users preferred machine-learning tools to Snowflake data. Snowflake provides the capability to implement models running as UDFs inside of Snowflake to make scalable and safe machine learning inferences or use External Functions to connect to a secure model endpoint. Snowflake makes it simpler for teams and applications to access model results in Snowflake so they can quickly consume and act on ML-driven insights. Through Snowflake's broad partner ecosystem, users can gain direct access to all current and future data science tools, platforms, and languages, such as Python, R, Java, and Scala; open-source libraries like PyTorch, XGBoost, TensorFlow, and sci-kit-learn; and notebooks like Jupyter and Zeppelin. Furthermore, Snowflake supports connections to the latest ML libraries and tools, such as Dask and Saturn Cloud. By offering a single consistent source for data, Snowflake eliminates the need to update the underlying data if tools, languages, or libraries are changed. Alteryx, AmazonSageMaker, Bigsquid.ai, Boostkpi, Databricks, and Dataiku are a few of the major data science platforms which have partnered with Snowflake to provide flawless native connectivity.

2.5. Companies In Different Sectors Using Snowflake

2.5.1. Media and Entertainment

Utilizing data from every point of contact with the consumer to build holistic views and provide more individualized service. With Snowflake's Global Data Clean Room capabilities, companies move beyond third-party cookies to work on identification resolution, campaigns, and attribute determination. Warner Music Group's CIO claims that the company actively uses Snowflake Data Cloud for musicians to help them better understand their fans and consumer patterns at the same time, which also helps them decide where to invest in new artists and new content kinds [5]. According to Audacy's VP of Data and Analytics, the company actively utilizes Snowflake to connect all of its data (from transactions and organized information to data collected from connected cars), resolution of identity, and creation of rich, 360-degree views, which has allowed it to surpass clients' expectations, implement distinguished attribution metrics, and foster increased revenue [5].

2.5.2. Healthcare

Novartis uses Snowflake's Healthcare & Life Sciences Data Cloud to transform its data into amazing avenues for knowledge by speeding up its ability to acquire, collect, and distribute data in an efficient way that would not be possible with other solutions, according to the company's head of Digital Platform & Product Delivery. Snowflake is assisting them in providing their patients with the medications they need more quickly [6].

2.5.3. Manufacturing

Snowflake streamlines process within and between production facilities while utilizing shop floor data in almost real-time to forecast maintenance requirements, assess cycle times, enhance product output and quality, and achieve sustainability targets. With Snowflake’s flexible architecture and computation, companies now focus on swift data collaboration with suppliers and customers to build massive supply chain resilience in a secure and scalable direction. Honeywell harnesses the power of Snowflake to bring data from ERP, CRM, supply chain systems, and more together in Snowflake to drive business impact [7].

2.5.4. Government and Education

Data sharing, teamwork, and well-informed decision-making are all made possible in the public sector by Snowflake's Government & Education Data Cloud. In addition to collaborating and sharing data within and between agencies and departments, organizations can modernize and speed up their migration to the cloud, along with helping to identify and safeguard against fraud, waste, and abuse and giving public sector officials comprehensive perspectives of people and students [8].

2.5.5. Retail

Major retail companies now leverage Snowflake to optimize their marketing strategies to cater to customer needs, optimize relevance and produce flawless experiences. Snowflake builds holistic customer profiles by building personalized, closed-loop campaigns and breaking down data silos. According to Kraft Heinz's Global Head of Machine Learning Operations, the company uses Snowflake's Retail Data Cloud to gather information from multiple sources on orders for goods, supplies, and production and to work together on data virtually in real-time with partners like Albertsons, all in one location, for end-to-end supply chain visibility that removes all doubt from their business [9].

2.5.6. Financial Services

By implementing a granular governance strategy, systematic code, and secured networks for all three clouds, all portfolio, reference, market, and risk data can be safeguarded, stored, and made accessible using Snowflake.

According to the Senior Director Of Engineering of CapitalOne, while other financial companies are gradually making their mark in the public cloud space, CapitalOne has made its mark by displacing all its data into the public cloud. According to the Cloud Data Engineering & Conversational AI Group Leader of Western Union, By lowering the amount of technology, servicing, licenses, and assistance required, Snowflake is fostering annual savings of millions of dollars [10].

3. Limitations of Snowflake

3.1. Not Available On-Premises

Since Snowflake is a SAAS (Software as a Service) offering, users cannot host it on-site. As a result, there is comparatively less control over the execution. Even though Snowflake offers numerous cloud options, there are only three available: AWS, Azure, and Google Cloud. Users should consider carefully if using a different cloud platform, such as one from Alibaba, IBM, or Oracle. Due to the dependence on the current cloud providers, the selection of geographies is likewise constrained. Snowflake itself manages both the infrastructure and the software. This is effective in many respects, yet it has several limitations.

3.2. Restriction on High Query Performance

For straightforward queries, query performance is good, but as complexity rises, performance suffers. Due to Snowflake's lack of support for indexes, a query must search through every row of data stored. Compression techniques and the related information, which aid in avoiding unneeded storage blocks, are the foundation of optimization. Even these solutions may encounter difficulties if the dataset is huge. Performance degradation can occur with searches that include joins or subqueries in addition to huge datasets.

3.3. Price Elevations

Snowflake offers computation and storage costs in terms of seconds. Moreover, when compared to other possibilities, this seems reasonably priced. However, in some instances, this can be a myth. Snowflake provides a variety of price and support options. Standard is the least expensive. Users may have to pay more if they want something specifically tailored, such as dedicated hardware, dedicated computing nodes, quicker results, more data, or higher availability. As was previously mentioned, there are not many choices available if users want to boost query performance. The most practical alternative is to increase the computing resources, but doing so can dramatically raise the cost. A different cluster is also needed to maintain performance for concurrent queries that are more than eight. Clusters of virtual data warehouses may quickly increase expenditures. For workloads made up of often occurring, repetitive query patterns, materialized views can also be leveraged to enhance query performance. This also entails additional expenses.

Additionally, storage fees may rise dramatically. It necessitates the storage of metadata and logs since it permits time travel. According to one study, a 90-day retention period can enhance base table capacities by 200 times or more. When there are numerous inserts, updates, data copies made across partitions, and clustering, storage capacity utilization may skyrocket.

3.4. Granular Access Not Supported

One of the most important components of a cloud-based solution is security. Although cloud-based security has improved significantly, and Snowflake makes use of it, it still lacks a crucial data aspect: fine-grained access. The data access cannot be defined at the row, column, or label levels. An organization needs to plan differently if it has several tenants (customers). Only virtual data warehouses can provide tenant isolation.

3.5. Difficult Adaptability

Dedicated hardware or software is not required, thanks to Snowflake. The operational worries are significantly diminished as a result. Many users, however, find it difficult to use the Snowflake platform. They still have to build the clusters even though we do not require any hardware. Based on various configurations, users still need to select the nodes. Untrained individuals are unable to complete this. Users still need someone to periodically evaluate how well data ingestion and access are doing, even if a cluster has been formed. Finding the best fit for an organization’s needs may need some experimentation.

The data cannot be copied simply if there is a need to migrate it from current systems. Indexes, restrictions, and other features that Snowflake does not support may be used by other databases. In order to avoid conflicts, users must clean up the data files. Furthermore, Snowflake does not impose restrictions like those requiring a unique key or primary key. If not managed appropriately, this could lead to some significant data quality difficulties. AWS S3, Azure Blob, and Google Object Storage's cloud-based storage solutions can load data smoothly into Snowflake. However, it does not do very well at loading data from transactional or outside sources. To do so, we must create unique pipelines, which may call for outside resources like Talend and Hevo.

The interactive platform traditional data warehouses offer consumers has evolved over time. Such maturity is missing from Snowflake. Without autofill or debug support, its SQL editor is extremely simple. Additionally, the interface is only accessible via the Web; a desktop version is unavailable.

The SQL itself has a lot of limitations. Complicated SQL constructs are not supported by Snowflake. In addition, learning JavaScript is a need if you need to build stored procedures.

4. Conclusion

In conclusion, Snowflake undoubtedly has been a game changer for businesses looking for a cloud-based data warehouse that seamlessly integrates scalability, flexibility, and ease of use. Snowflake reduces the complexity and latency caused by standard ETL processes by taking data from several settings and putting it all on a single high-performance platform. With Snowflake, users can immediately profile and clean data, ensuring high data integrity. In order to make it easier for customers to locate and access their data, Snowflake also provides data discovery tools. Snowflake additionally provides immediate access to many different third-party data sets via its Snowflake Data Marketplace. There, hundreds of providers offer exclusive third-party data, which is readily available upon request.

Since each workload and team has a dedicated compute cluster, thanks to Snowflake's innovative architecture, there is no resource conflict between data engineering, business intelligence, and data science workloads. Machine learning (ML) partners of Snowflake push down a large portion of their automatic feature transformation into Snowflake's scalable and dynamic cloud data pool, significantly transforming the nature of AutoML.

Since Snowflake has a wide partner ecosystem for different programming languages, customers can leverage rich benefits. Open-source libraries, data science notebooks, and numerous data science platforms

A wide spectrum of industries like retail, government & education, finance, manufacturing, entertainment, healthcare, and media are harnessing Snowflake to improve their data management and data science capabilities.

Snowflake began as a cloud-based data warehouse and has subsequently transformed into a provider of complete data solutions. Its feature set is expanding quickly. It has demonstrated ongoing innovation along the way, from the data warehouse to the data marketplace. There are multiple challenges when implementing Snowflake, including unavailability on-premises, limitations on high query performance, price elevations, lack of access to granular support, and difficult adaptation. Snowflake has the potential capability to overcome the existing challenges and limitations and drastically improve the ways enterprises manage their data and ML infrastructure.

References

[1] Michael Desmond, Machine Learning Market to Grow to Nearly $21 Billion by 2024, 2019. [Online]. Available:
[2] https://pureai.com/articles/2019/07/23/nwsdes-machine-learning-market-growth.aspx
[3] The Snowflake Website. [Online]. Available: https://www.snowflake.com/guides/feature-engineering
[4] Snowflake Marketplace, 2023. [Online]. Available: https://docs.snowflake.com/en/user-guide/ui-snowsight-marketplace
[5] Kavya Nagarajan, The StreamSets Website, 2022. [Online]. Available: https://streamsets.com/blog/snowpark-udfs-python/
[6] The Snowflake Website, Media Data Cloud. [Online]. Available: https://www.snowflake.com/en/solutions/industries/advertising-media-entertainment/
[7] The Snowflake Website, Healthcare and Life Sciences Data Cloud. [Online]. Available: https://www.snowflake.com/en/solutions/industries/healthcare-and-life-sciences/
[8] The Snowflake Website, 2023. [Online]. Available: https://www.snowflake.com/en/solutions/industries/manufacturing/
[9] The Snowflake Website, Government and Education Data Cloud. [Online]. Available: https://www.snowflake.com/en/solutions/industries/public-sector/
[10] The Snowflake Website, Retail Data Cloud. [Online]. Available: https://www.snowflake.com/en/solutions/industries/retail-cpg/
[11] The Snowflake Website, Financial Services Data Cloud. [Online]. Available: https://www.snowflake.com/en/solutions/industries/financial-services/