Deloitte.

To Snowflake or not to Snowflake

Written by Kevin Kewal, Jeannine de Kuijper, and Jeroen Gelderblom

Snowflake has made a grand entrance in the market as a Software as a Service provider. Are all key elements they promote as disruptive as suggested? Through this article, we present the most well-known Snowflake components and discuss their capabilities, zoning in on separation of storage and computation along with the costs, automatic scaling, zero-copy cloning, semi-structured and schema-less data support, data protection, data owners, and third parties.

Who knows, this might just be the information you need to decide on whether or not you should transition to Snowflake.

It is the year 2012 when in the sunny San Mateo, California, Benoit Dageville and Thierry Cruanes combined their forces to found Snowflake, the first company to separate cloud data storage from computing resources. Their love for data and computer science inspired them to build a system that would adapt its resources to the demand - so that customers would only pay for their consumption - and include the possibility for the system to scale to infinity. Hereby Snowflake uses two technologies pioneered at the CWI (Centre for Mathematics and Computer Science) Database Architectures group in Amsterdam: vectorized query execution and lightweight compression methods in its columnar data storage. Vectorized query execution means that the data rows are batched together and represented as a set of column vectors, and columnar data storage reduces the amount of data that needs to be loaded from the disk. Marcin Żukowski, a PhD graduate who helped develop these technologies, brought these methods to Snowflake's attention. After sitting idle for two years, Snowflake launched publicly in 2014 and since then it has been a rapid growing enterprise leading to its nomination of being ranked first in the Forbes Cloud 100 in 2020.

Since the launch of the Snowflake Partner Network (SPN), Deloitte has been the fastest organization to achieve Snowflake's top level of recognition as an Elite Services Partner. With over 500.000 hours of hands-on Snowflake implementation experience, Deloitte has a track-record of delivering analytic modernization projects for its clients and has demonstrated capabilities, comprehensive approaches and accelerators. Now, in 2021, Snowflake has gradually climbed its way up the Gartner Magic Quadrant, into the Leader and Visionary quadrant. So, what makes Snowflake so special? Why is it becoming such a hot topic despite its cold name?

Deloitte's collaboration with Snowflake and its experience in migration projects for clients have given us an insight into the benefits and drawbacks of implementing such a system. In this article we lay out the unique Snowflake architecture and dive deeper into a few of the core topics: separation of storage and computation along with the costs, automatic scaling, zero-copy cloning, semi-structured and schema-less data support, data protection, data owners, and third parties. Our experience with these Snowflake areas enables us to provide the advantages and drawbacks of each and can guide the decision for why a company should or should not transition to Snowflake.

"Deloitte's collaboration with Snowflake and its experience in migration projects for clients have given us an insight into the benefits and drawbacks of implementing such a system."

Dageville and Cruanes began their journey of setting up Snowflake by steering away from the conventional data warehouses and repurposing a brand-new architecture for the cloud, thereby supporting fault isolation, performance isolation and elasticity of the cloud.

By setting up a new data warehouse from scratch and incorporating the demands of the customers, they have been able to provide unique advantages of the Snowflake system to the market. The Virtual Private Cloud (VPC) consists of three distinct layers: data storage, query processing (also known as the compute layer), and cloud services.

Database Storage
When data is loaded into the Snowflake Database Storage, Snowflake manages all aspects including the file size, compression, organization and structure.

Query Processing
These data objects are then only visible by querying the data in the Query Processing (Compute) layer using SQL query operations.
The compute layer uses virtual warehouses where each virtual warehouse is a Massively Parallel Processing (MPP) compute cluster that is independent from others.

Cloud Services
Finally, the cloud services layer combines a selection of services including authentication and control, infrastructure management, optimization, metadata management and security. This layer runs on compute instances from a third party cloud provider (Microsoft Azure, Amazon Web Services (AWS) or Google Cloud Platform (GCP)) and contains the optimized data once the loaded data has been compressed and put into columnar format.

This unique structure is at the core of some of Snowflake's greatest advantages, however, some topics also include a few drawbacks which might not be so obvious at first glance.

"Based on our experience with Snowflake, we are able to provide the advantages and drawbacks of several Snowflake focus areas and can guide the decision for why a company should or should not transition to Snowflake."

1 Storage vs. Computation Costs

The most obvious topic for discussion is the separation of the storage database from the computation. This means that you have complete control over each individual component and that you can decide how to load data into the database storage without paying the extra costs of computing resources. This brings us to a direct correlation with the "pay for what you use" mentality that is provided by Snowflake, and which they use as one of their unique selling points. All costs are based on usage of data storage, compute resource needs (virtual warehouses) and usage of the cloud services. So, each layer in the Snowflake architecture has its own costs and is independent of each other.

Database storage costs are charged regardless of whether it is in Active, Time Travel, or Fail-safe state and is calculated on the daily average amount of data (in bytes) stored in the system.

The usage of compute resources is calculated based on the number of Snowflake credits consumed per minute by the virtual warehouses. (depending on warehouse size) for executing queries, loading/unloading data and other DML (Data Manipulating Language) operations.

Finally, the usage for cloud services is only charged if the daily consumption of cloud services for the account exceeds 10% of the daily usage of the compute resources.

As each layer has its own system usage and billing, companies are able to significantly reduce costs by paying solely for what they consume. Furthermore, Snowflake provides several user-friendly methods (web interface and SQL) so that you are able to monitor and control your storage, credit, and cloud services usage. No surprises there. A downside, however, is that the separation of the costs might prevent accurate predictions of possible future costs.

2 Automatic scaling

Another great advantage of Snowflake comes from its elasticity. Dageville and Cruanes made it a priority that the storage and compute resources could be automatically and seamlessly scaled without impacting the data availability or performance of concurrent queries. What does this mean?

Imagine you are running queries on a small virtual warehouse, but more data scientists are using the data warehouse, resulting in a significant increase in the number of queries that need to be performed. Snowflake provides an auto-scaling (economy and standard) which can not only increase the size of the warehouses (scale-up), but can also increase the number (scale-out) of virtual warehouses necessary for your data loads. This is extremely useful for unpredictable data sources and loads such as application logs, web applications, mobile devices, social media and sensor data (IoT).

However, automatic scaling requires knowledge of the scaling policies. If you set up the multi-warehouse environment with a standard policy, this means that Snowflake will automatically scale up or out the virtual warehouses to comply with your computing needs. This in turn can drastically increase your computation costs without your knowledge. Automatic scaling is a lovely addition to the environment as you do not have to worry about having enough compute power, but it can drill a hole in your pocket and let the money flow.

3   Zero-copy Cloning

Zero-copy cloning might be one of the greatest perks of Snowflake. Copying data is often difficult, time-consuming and expensive, but with Snowflake, if you want to copy objects such as the database, schema or table to test or experiment with the data, you can use this zero-copy cloning to avoid altering sensitive data. It is a point-in-time version of the data on which you can run queries and alter tables. The reason it is called zero-copy is that it does not physically copy the data, so there are no extra storage costs until you change a cloned table (add, delete or update new data), and then costs will be incurred when you modify the original table. Furthermore, the data protection of Snowflake ensures that cloned objects do not inherit the source's granted privileges, so you must grant users privilege to the cloned objects.

This is a great feature to have, but there is one point of consideration: zero-copy cloning can only be used within the same Snowflake account. If you have set-up multiple Snowflake accounts (e.g., separate accounts for production data, testing data, and acceptance data) zero-copy cloning will not work.

4   Data Ownership

A very important topic to consider, is that when loading your data onto Snowflake, you no longer own the data. I can hear you thinking: what? Well, let us look at how this came to be. One of the great advantages of Snowflake is its data protection, one of which is known as Key Rotation. Snowflake encrypts all data using a strong AES 256-bit encryption with a hierarchical key model where the keys are rotated regularly – a great data protection method, let us be clear about that. However, these access keys for the data are owned and stored separately by Snowflake. Therefore, if there is ever a problem at Snowflake which prevents you from accessing the key, it directly restricts you from accessing your own data. You are completely reliant on Snowflake.

5 Data Protection

Authentication, access and security are all part of the cloud services layer in the Snowflake architecture. For companies that rely heavily on data protection due to their data sensitivity, Snowflake offers several possibilities. The first is setting up your Snowflake edition, which can be Standard, Enterprise, Business Critical or Virtual Private Snowflake (VPS). Each of these editions builds upon the other and increases its data protection possibilities. The VPS, for example, is perfect for financial institutions and other organizations that collect, analyze and share highly sensitive data as it is a completely separate Snowflake environment, isolated from all other accounts. This means that you do not share any resources with other Snowflake accounts outside the VPS.

Another method of data protection is Continuous Data Protection (CDP) which encompasses a set of features that help against human error, malicious acts, and software or hardware failure. These features include:

- Network policies for granting or restricting access based on IP address

- Verification/authentication required for all users

- Security roles for controlling user access to all objects in the system – users have to be granted permission and access to objects from the SYSADMIN role

- All ingested data stored in Snowflake tables are encrypted

- Time Travel for querying and restoring data that has been changed or deleted

- Fail-safe for disaster recovery

Even in the case of a system failure or other catastrophic event, Snowflake is able to recover the data.
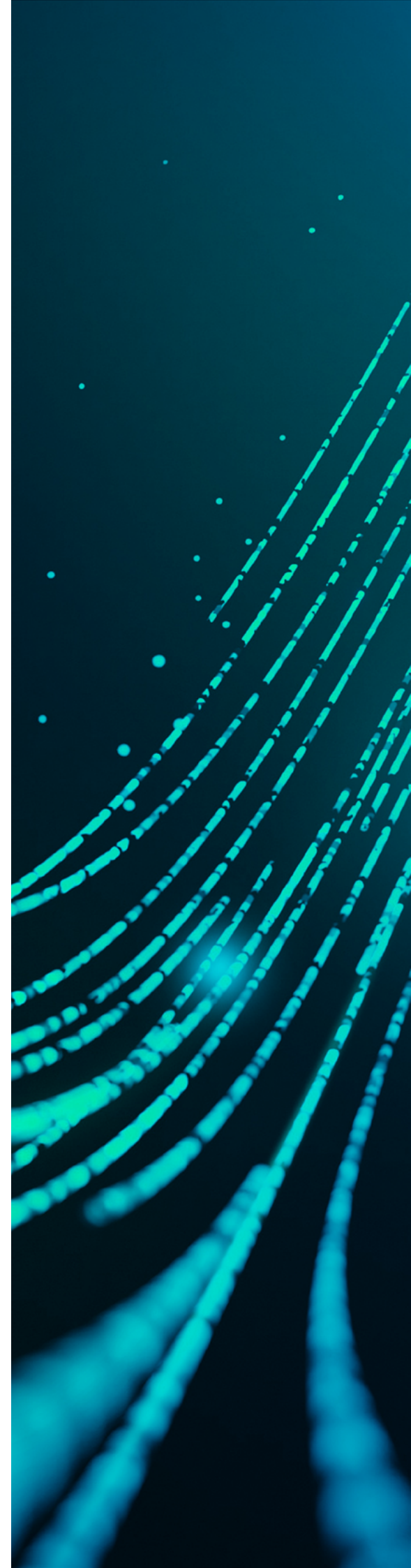
6 Support Semi-structured and schema-less data

APIs (Application Program Interfaces) provided by cloud applications or third party data providers often provide the data in semi-structured formats such as JSON and XML. Snowflake offers SQL extensions (VARIANT, ARRAY, and OBJECT) for traversing, flattening and nestling these types of semi-structured data which warrants an ELT (Extract-Load-Transform) manner instead of the traditional ETL manner. Furthermore, this enables transformations to be performed on the data load without having to specify the document schema, which makes the operations on semi-structured data nearly as fast as operations on relational data.

As of yet, our experience with semi-structured and schema-less data support with Snowflake has seen no faults!

7 Third party cloud providers

Snowflake builds upon third party cloud providers (Azure, AWS and GCP). This means that if you already use one of these platforms, you can choose to host your Snowflake account on the same platform or host the Snowflake account on different platforms. Furthermore, you can use Azure, AWS or GCP as an external stage for loading tables. This is ideal for users who already have one of these accounts.

Nevertheless, the benefit we just proposed of Snowflake building upon third party cloud providers, automatically highlights an inherent flaw: it relies on third party cloud providers. This can be a great drawback when a service disruption occurs at the third party because Snowflake is then automatically affected by the same disruption. Another point of concern is that when using third party cloud providers to transfer data (loading data from external stages into tables), transfer billing charges may apply.

Conclusion

We have seen some great reasons why Snowflake has grown to become so large in the last ten years. They provide great advantages with their separate storage database and compute clusters, relatively lower costs as you only pay for what you use, automatic scaling, zero-copy cloning, support with semi-structured and schema-less data and data protection. However, most of these topics also have their drawbacks and some, such as not owning your own data, see no advantages at all.

Is Snowflake right for you?

Whether or not you should choose to transition to Snowflake is greatly dependent on your needs and industry. We have seen that the advantages Snowflake provides can also lead to some drawbacks and it is the weighing of these concepts that should ultimately lead to an informed decision. As an Elite Service Partner of Snowflake, Deloitte can help you with this decision. If you would like to have an in-depth comparison of cloud providers, for example if you are contemplating between Snowflake and Redshift, we can provide the options based on your needs and requirements, providing expertise on whether Snowflake is the right fit for you.

If you consider going for Snowflake, we have vast experience that can aid you in your decision making and can help you set up a migration strategy. With Deloitte's Migration Factory in collaboration with Snowflake, we focus on automation and business continuity, meaning that we find a way to migrate the data without disrupting the business. Alongside this, we are able to set up a pilot to validate the approach.

Want to know more about what we can offer you regarding Snowflake?
Reach out to one of our team members for any questions and/or comments!

Summary

Snowflake is a Software-as-a-Service (SaaS) and builds upon the resources from existing cloud providers

When data is loaded into the Snowflake Database Storage, Snowflake manages all aspects including the file size, compression, organization, structure, etc.

Data objects are then only visible by querying the data in the Query Processing (Compute) layer using SQL query operations

Snowflake separates the storage and compute resources and thus also the costs

Storage and compute resources can automatically and seamlessly scale

Zero copy cloning allows you to copy databases, schemas or tables within an account to test or experiment with the data, without incurring additional costs, as data not physically reproduced unless you alter it

Snowflake offers support for semi-structured and schema-less data, such as JSON and XML

As Snowflake builds upon third party providers, a service disruption at a third party will also result in a disruption in your Snowflake operation

Who owns the data? Snowflake controls the key to your data. In case Snowflake is not accessible, you will not be able to access your data

Emile Kramers
Partner

+31 (0)882882576
ekramers@deloitte.nl

Emile is a partner within Deloitte Consulting and the Artificial Intelligence & Data offering. He has over 25 years advisory and delivery experience in the Analytics, AI and Information Management domain.

His primary focus is on clients in the TMT industry, but has work in FMCG and Energy & resources industries as well. Before joining Deloitte Emile has led Analytics service lines across EMEA with Accenture, DXC and Atos.

Dennis Scheeren
Director

+ 31 (0)6 8333 0270
dscheeren@deloitte.nl

Dennis is a Director at Deloitte Consulting and Business Lead of the Data Modernization and Analytics team. He focusses on the Financial Services Industry (FSI) and specific the Banking industry.

Dennis has 15 years of experience in management consultancy within financial services on data transformation, data and regulatory remediation and operational analytics.

Deloitte.