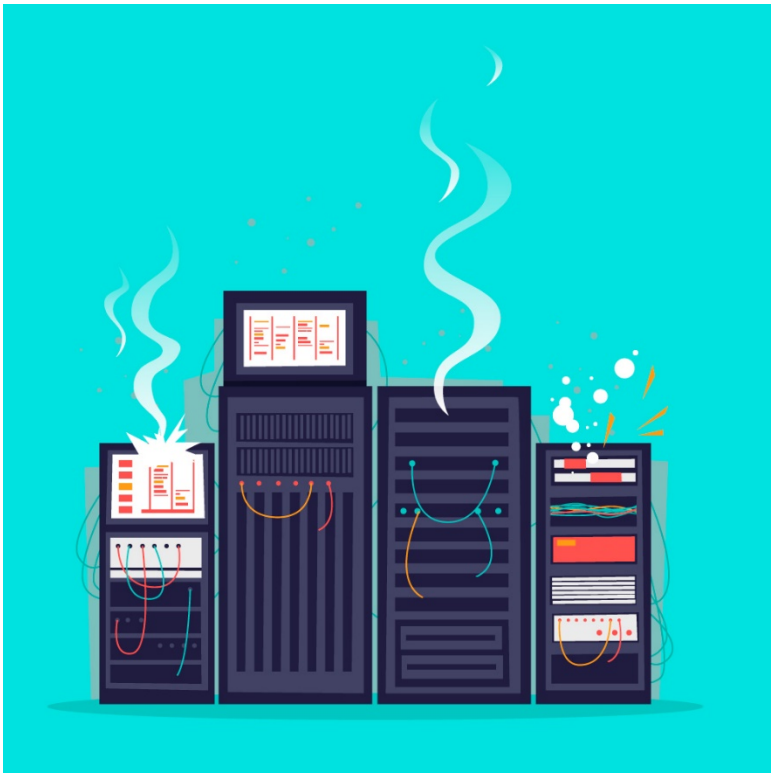
onebridge

Data Modernization with Snowflake

TABLE OF CONTENTS

What's the Buzz About Data Modernization?

We've all heard by now that businesses need to be increasingly data-driven to stay at the top of their game. Organizations that effectively use their data can better serve their customers, increase operational efficiency, and surpass their competitors.



But becoming a data-driven organization requires accessing, managing, distributing, and analyzing all your data while it's still valuable, as well as tapping into new data sources. This drive to use data effectively has created new challenges, especially that data source volume is growing at an exponential rate. For many organizations, quantities of data can reach almost unimaginable levels. **Harnessing the power of all that data can be daunting, and most run into issues with hardware and server performance/storage.**

Legacy servers may be unable to keep pace with user demand, expanding sources of data, and increased volume. Even once server upgrades take place, there are still more costs involved with setting up IT Operations capability to manage server infrastructure and implement software upgrades to keep systems running smoothly. Software licensing costs are often recurring, and with the constant fast pace of innovation, hardware upgrades will eventually become necessary.

However, it's clear that data is an asset to any organization. To start with, there are hidden stories in your data. Stories about your customers/clients that can help you serve them better. Stories about market opportunities your company can take advantage of, even stories that reveal the smartest way to take advantage of those opportunities.

Data as an asset isn't just about business opportunity – it's also about risk, compliance, and privacy/security. **Your data may not just be an asset to your company, it may be sought after by others who can benefit from it, too.** As the world of data keeps growing, so do regulatory requirements. Modernizing your data strategy is pivotal for keeping your organization ahead of the regulatory curve and your client data protected.

Data monetization is now possible, too. According to The Insights Professional's Guide to External Data Sourcing by Forrester Research, most businesses look at external data sources to supplement their own data. **This demand for external data creates an opportunity for every company to earn revenue from the**

information it generates, with raw data, packaged data, data analysis/insights, data enhancement, and data trade often the most attractive.

Relying on traditional business intelligence (BI) methodologies doesn't cut it anymore in today's fast-paced data economy. While successful, these methods often take six months or more to implement. **Modernized platforms and methods can reduce the time to data insight to a few months or even weeks through self-service analytics, agile BI, and data vault methodologies.**

While investment is required, if done correctly there will be a substantial return on that investment. And your investment in data modernization can be closely governed by choosing tools that offer flexibility and scalability to your teams, as well as cost savings and compelling ROI.

Along Comes the Cloud

Luckily, the cloud offers an incredible solution to many of the problems of data modernization.

In its purest form, cloud computing is a suite of computing services enabling companies to scale quickly, focus on solutioning, and reduce the need for operational overhead. Storage, computing, software, database, and networking services are typically all baked into a cloud service and provided as a pay-as-you-use subscription that operates over the Internet – as long as you have a connection, you have on-demand access to all of these services.

This means that the cloud can deliver whatever amount of IT resources are required at that particular time for your business. Want to scale down during times of low activity and save a little cash flow? You can do that. Need to scale up for extra support during a big project? That's easy, too. Self-service, on-demand access allows resources to be administered in minutes, usually with just a few mouse clicks. The cloud is elastic in its ability to quickly adapt to your needs.

Contrast this with traditional on-premise server environments, where hardware is sized to the highest expected workload. In these cases, the server is rarely used at capacity. Being able to scale on demand to increase compute or storage, or even dynamically based on the compute demand at any given moment, offers a lot more flexibility.

It also offers a lot more control over spend. **Cloud services typically offer a pay-for-use model and can scale based on usage requirements.** For example, an analyst running data discovery queries can use a lower rate service model, while a mission-critical / customer-facing service may use a high-compute model at a higher rate.

The Right Tool for the Job

Here at Onebridge, we've discovered the best tool for the job: Snowflake, the data warehouse built for the cloud from the ground up. Why?

With Snowflake, there's **no new language to learn or convert your existing code to** – most standard SQL commands are supported, and very minimal changes are required to existing code.

One of Snowflake's key benefits is its "dynamic elasticity" – which means that it's extremely flexible and scalable for future growth and allows businesses to rapidly address immediate business challenges. **Variable compute allows you to increase or decrease resources in minutes to align your compute power with database needs.** Limited system maintenance is required, so you can spend time maintaining your data instead of optimizing database servers.

With Snowflake's multi-cloud support, you can move your data to any cloud in any region, without having to re-code your applications or learn new skills. Snowflake is more secure by default with high levels of encryption and security enabled from the start, and **operates as a service on Amazon Web Services, Microsoft Azure, and Google Cloud Platform**. It plugs into many strong single sign-on solutions and features a host of integrations with industry-standard formats, with new software only needed for outside connections. Built-in location mirroring allows you to quickly and easily share between multiple supported geographies, a feature that is frequently used for disaster recovery efforts.

Here are more of Snowflake's most compelling benefits:

Key Benefit: Separate Storage from Compute

Snowflake has built an environment that allows the customer to scale their data warehouse as they want. The innovative, patent-pending, Multi-Cluster Shared Data Architecture in Snowflake takes full advantage of the elasticity of the cloud. This unique environment gives users the ability to add more compute power on demand, even in the middle of a process.

How it works:

- Data is stored, once, on elastic disk storage in the cloud (for example it is currently stored in AWS S3)
- Compute clusters (for example using AWS EC2 nodes) are allocated dynamically as required to do work (queries or loads)
- The compute clusters are connected to the data storage. The access to the data is based on the approved security credentials of the user that is running the query

The important point is that **storage and warehouse are not tightly coupled together. This means you can grow and reduce both independently of each other.**

Say you are running a series of complex SQL queries on a Small (2 node) warehouse in Snowflake. Normally you have until the close of business to run the reports and prepare the data. But today, at 8 AM, your management requests the results of that analysis for a board meeting at noon. So now you have half the time to get the reports to run. In a traditional data warehouse, you would pretty much have only one option: kick everyone off the system and hope things run faster.

If, however, you are running the reports on the Snowflake Elastic Data Warehouse, all you have to do is increase the size of the warehouse to a Medium which doubles the horsepower of the data warehouse (i.e. it doubles the number of nodes). And you can perform this change with a few clicks in Snowflake's web-based interface, even if the process has already started.

Once the size is increased, the next statement will use the additional resources immediately, in about half the time as the prior statement despite including about double the data. You don't need to interrupt the process, start over, or restart the service. It just takes effect automatically, at the push of a button. And once the process is done, you can reduce the warehouse back to a Small or even just turn it off.

And in this way, the unique cloud-based architecture of Snowflake allows you to truly take advantage of the elasticity of the cloud for your data warehouse.

Key Benefit: Multi-Cloud Support

Snowflake's cloud data platform supports a multi-cloud strategy, including a cross-cloud approach to mix and match clouds as you see fit. With a common and interchangeable code base, Snowflake delivers advantages such as global data replication, which means you can move your data to any cloud in any region, without having to re-code your applications or learn new skills.

Three major challenges exist today that make a multi-cloud strategy incomplete:

- **Cloud silos** are created as soon as data exists in a public cloud. Because each major cloud provider created a unique offering with proprietary APIs for data management, there's no easy way to copy or share data from cloud to cloud. Making matters worse, it's hard to find DevOps employees who have the skillset to work in multiple clouds, which often leads to separate cloud teams within an organization (yet another silo).
- Cloud services work best when users are in close proximity. As a result, geography plays a role in creating **data silos by region**, especially for organizations that operate in multiple locations (regions, countries, and continents).
- **Data portability** is a problem for all organizations, including those that use open source technologies and open data formats. Today, there's no easy way to lift multiple petabytes of data to change clouds, open source or otherwise.

The true benefits of a multi-cloud strategy will not materialize until data can be shared and replicated across clouds and regions. Fortunately, **cross-cloud capability** is the answer. Two basic requirements exist for cross-cloud capability:

1. A cloud-agnostic layer must provide a **unified data management platform**, which sits on top of each cloud region and all cloud infrastructure regardless of which cloud platforms are used. By providing identical functionality across all cloud platforms, the data management platform enables a cost-effective and seamless method to securely share data.
2. Data must move anywhere easily, which requires a high-throughput communication "mesh" that enables **complete data portability**.

With cross-cloud capability, organizations will be able to securely share data across regions and cloud accounts while adhering to the same rules of data sharing (data exists locally in a single source where it's accessed rather than moved). Plus, the platform will make cross-region asynchronous data replication possible without impacting the performance of accessing primary data.

In short, **a cross-cloud capability removes all barriers to data** so organizations can:

- Analyze all data for decision-making, no matter where the data is located
- Ensure business continuity and disaster recovery through cross-cloud replication
- Perform account migration without data portability concerns

Key Benefit: Data Sharing

Traditional sharing methods require laborious and costly procedures to share even just a slice of your data warehouse. Today's volume, variety, and velocity of data only intensify sharing issues.

With Snowflake's Secure Data Sharing technology, organizations around the world are revolutionizing how they share data and use data shared with them. With Secure Data Sharing, data doesn't move, eliminating the cost, headache, and delays associated with legacy data sharing methods that deliver only slices of stale data, and limited insights associated with that data. **Snowflake's modern approach to data sharing provides any organization easy access to shared data, so they can combine it with their existing data to get the deepest insights possible.**

Snowflake Secure Data Sharing enables organizations to instantly and securely share their data. Companies can share governed data internally between departments, subsidiaries, and operating groups, or externally with partners, suppliers, vendors, and even customers. **Organizations can also monetize parts of their data through external data sharing to create new revenue streams.**

Snowflake's data sharing capabilities include:

- **Share live data with ease**: Within minutes, share any part of your data warehouse while maintaining a single source of truth.
- **Eliminate data movement**: share read-only data and database objects without moving data, rebuilding schemas, or creating pipelines.
- **Create more assets from data**: share more data more effectively and efficiently, internally or externally, compared to slow, costly, and labor-intensive approaches such as FTP, EDI, APIs or email.
- **Provide controlled, customized views**: make available a single table or your entire data warehouse to partners, vendors, and data customers.
- **Cut out the middleman**: create your data sharing business agreements direct with other enterprises and without additional costs to share data.
- **Pay only for what you use**: as a data provider, read-only sharing means no extra storage cost to share data. As a data consumer, pay only for the compute charges you use on shared data.

Snowflake is currently offering a data sharing rebate program to further reduce cost. Customers automatically receive an equivalent of 10% of the credits their data consumers use when performing operations on shared data.

Key Benefit: Supports Multi Data Types

Today, data arrives in diverse forms from diverse sources. The rapid decrease in the cost of storing data and the growth in distributed systems has led to an explosion of machine-generated data. This includes data from applications, sensors, mobile devices, and more.

Semi-structured data formats such as JSON, Avro, and others have become the de facto form in which this data is sent and stored. Semi-structured data is easy for these applications to create and capable of representing a wide array of information.

With Snowflake, there's no need to implement separate systems to process structured, semi-structured, and non-structured data. You can eliminate complex Hadoop and data warehouse pipelines. Snowflake can perform both roles much more efficiently and with better business results at a lower cost.

Snowflake was designed from the start to handle semi-structured data without the trade-offs of current approaches. Their patent-pending approach makes it possible to have both the schema flexibility of semi-structured data and the performance optimizations of relational data. To do that, they designed the database engine to handle both semi-structured data natively without transformation.

You can query both structured and machine-generated, semi-structured data (i.e., JSON, Avro, XML, Parquet) using relational SQL operators with similar performance characteristics as if querying structured data. Loading semi-structured data is painless. Schemas are dynamic and are automatically discovered during load. This support for dynamic schemas enables efficient query execution using natural extensions to SQL.

Key Benefit: Data Streaming for Near Real-Time Reporting & Insights

Rapid and immediate access to live data and the ability to analyze it in the moment are key requirements so the data-driven enterprise can streamline operations, better serve customers, and uncover new market opportunities. Data pipelines in Snowflake can be batch or continuous, and processing can happen directly within Snowflake itself. Thanks to Snowflake's multi-cluster compute approach, these pipelines can handle complex transformations, without impacting the performance of other workloads.

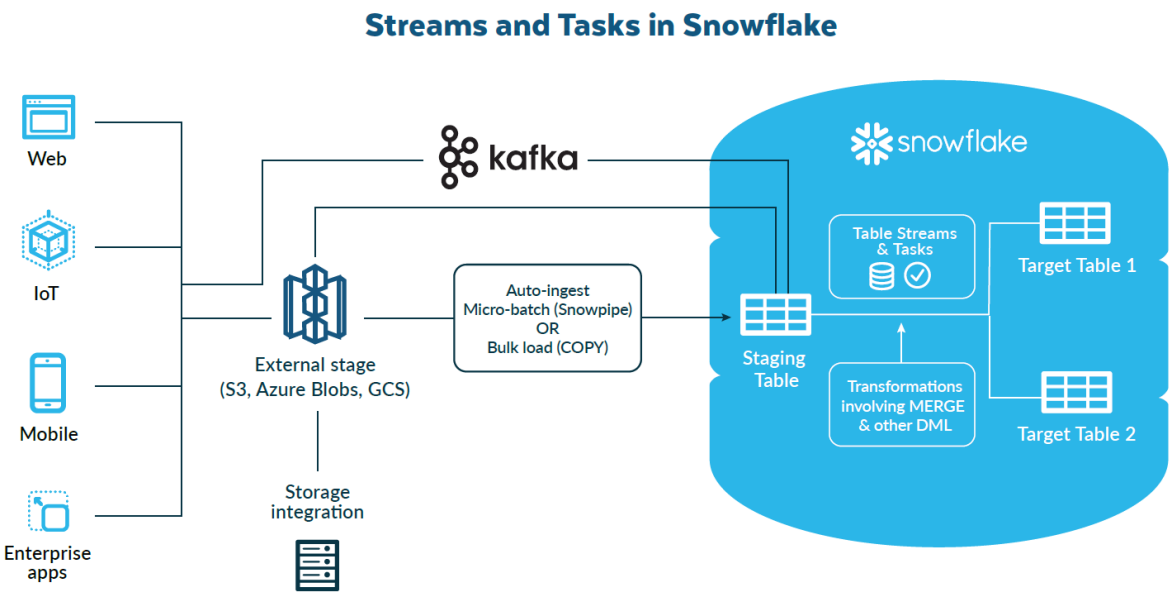
Snowpipe is an automated service that utilizes a REST API to asynchronously listen for new data as it arrives to your cloud storage environment. **As data arrives, whenever it arrives, Snowpipe then loads that data into Snowflake.**

Snowpipe doesn't require any manual effort to configure or run, and uses servers separate from the customer environment to ensure workload isolation. **Customers only pay for the per-second server time they use, keeping data ingestion costs predictable and affordable.**

With Snowpipe, you get:

- **Continuous Data Pipelines**: Simplify ingesting and processing of data across a broad range of use cases. Move data within a Snowflake environment continuously, without the need for scripts or custom code.
- **Stream Data from Apache Kafka**: Support event-driven architectures by ingesting JSON or Avro messages from Kafka using a Snowflake native connector.
- **Auto-Ingest Using Snowpipe**: Automatically stream files from Amazon Web Services, Microsoft Azure, or Google Cloud Platform object stores as new data arrives. Parse and insert files using Snowpipe without separate scheduling or orchestration.
- **Implement Change Data Capture**: Capture changes made to the underlying table via Snowflake table streams. Enable change data capture (CDC) use cases for continuously loading or changing tables, regardless of data scale. Table streams enable CDC use cases for continuously loading or changing tables, regardless of data scale.
- **Build Robust Data Transformations**: Schedule workloads to execute data transformations when needed. For use in more complex event-centric pipelines, trigger tasks that are based on outcomes of other tasks.
- **Support All Your Users**: Easily scale and allocate resources to different workgroups without data or resource contention. Take advantage of Snowflake's instant and infinite scaling of independent workloads for any level of concurrency without impacting performance.

Diagram on next page...



Ingesting structured and unstructured data from Kafka and a data lake into Snowflake.

Key Benefit: Operations Cost Savings

Snowflake allows for some IT operations to be absorbed into cloud services. For example, according to a Forrester Total Economic Impact study on the potential ROI of Snowflake versus traditional on-premise alternatives, "when ETL teams are no longer needed to move data from transactional systems to operations systems to data warehouses, then labor savings ensue, amounting to nearly $995,000 for the composite company." The study also identified legacy storage costs saved, legacy compute costs saved, reduced cost of database management, faster time-to-production, and capacity planning labor cost avoided.

With Snowflake's pay-as-you-use scalable compute structure, moving to a data warehouse in the cloud is win-win in terms of cost savings.

Snowflake Case Study: Major Automotive Services Company

A major automotive services company was faced with the following business and technical challenges:

- **Business Challenges**
  - Inability to rapidly share data internally and externally
  - Demand for faster data integration and access to company data sources/assets
  - Real-time analytics and insights
- **Technical Challenges**
  - On-prem servers and database platform unable to scale for data expansion and volume
  - Traditional BI methodologies inability to keep up with demand
  - Growing user demand of infrastructure and access to data
  - Data security, access, and delivery

After reviewing several big data platforms and cloud service offerings **Snowflake was chosen for its cloud data warehouse solution and its ability to address the business and technical objectives**. The solution was implemented across two projects.

The first project was to procure the service, onboard and train the development team, and pilot loading an existing fact table for benchmark testing. This project was estimated to take four months. However, the team was able to deliver in just under 45 days. **Benchmark tests showed over a 5000% improvement in query performance over the same on-prem query.**

The second project intended to capture a streaming source of live activity data and build a Tableau Dashboard to reflect the live data capture. **In just under 30 days, the team was able to capture and load data using Snowpipe from an S3 bucket. A dashboard was then developed and could demonstrate real-time activity by refreshing the Tableau Dashboard.**

With these two projects demonstrating key capabilities and the speed in which they could be implemented, a strategy was set in place for legacy on-prem and new project development with the overall goal to have all user interaction and consumption from Snowflake.

Case Study Key Objectives:

- ✓ New source integrations and new development would occur in Snowflake.
- ✓ The legacy on-prem data warehouse would take on a shift and lift replication plan with a roadmap to convert each source/existing solution over time.
- ✓ Front-end reporting would follow the same strategy.

How to Put Data Modernization to Work for You

Here are a few tips around taking advantage of the benefits data modernization can offer your business:

Identify business use cases and challenges. Defining your business use cases and overall key objectives up front will go a long way to realize the value of a data modernization plan. You'll have a higher percentage of business adoption and funding when you are able to demonstrate value in terms of time and money.

Plan for the future. Think big but start small. Figure out where you are going so you can figure out how to get there. Once you can see the big picture and take inventory of all the key deliverables, create a Business Opportunity heat map to identify high-value low-effort projects for the kind of quick wins and fast turnarounds that increase business engagement.

Choose a solution partner that can serve your unique needs. Look for a solution partner that has experience in the platform and technologies you are interested in. A reputable partner will have and share a successful track record of engagements with other organizations, have resources certified in the various technologies, and typically a partnership with the various technology companies.

Be prepared to think and develop differently. Data modernization will challenge the more traditional development methodologies. It helps to be open minded and willing to try methods that break the rules of traditional development life cycles. It's also a good idea to bring business partners closer to the development process. Your teams will become nimbler. They will deliver smaller, higher quality feature functionality more quickly, and your business partners will see value in weeks rather than months.

Consider a new methodology. Many new methodologies are available when considering data modernization:

- Agile / SAFe – to align business and IT to deliver feature sets quickly in an iterative model.
- Self-Service BI/Analytics – enable the business to engage early in the data development life cycle and interact with certified and third party data. This often leads the business to the ability to articulate requirements in the form of reports and dashboards.
- DevOps – development teams can deploy code more frequently offering high value and low impact to users and the business. Continuous Integration / Continuous Delivery, or CI/CD, is a set of guidelines in which development teams can automate code deployment and testing, further reducing the time for delivery.

Making the Leap to Data Modernization with Snowflake

Data modernization is critical to enabling the use of data as an asset, but many companies have outgrown their legacy data warehouses. Snowflake makes it easy and cost effective to upgrade to a data warehouse that allows you to scale to support the current volume of data and handle the complexity of accessing, managing, distributing, analyzing, and monetizing your data. With a multi-cluster, shared data architecture and ability to store, transform, and analyze all types of data, you can focus on using data to drive insights instead of maintaining your data warehouse.

Onebridge regularly hosts free training to help the data analytics community grow their capabilities, including frequent workshops focused on Snowflake. Sign up here to get notifications of our upcoming events: https://www.onebridge.tech/onebridge-training-and-workshops