DATAOPS IN SNOWFLAKE

How Data Professionals Can Surmount the Challenges of Modern Data Management

CHAMPION
GUIDES

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In recent years, principles from the DevOps world, initially created to encourage Agile software development, have been applied to data modeling and design. A newer term, DataOps, short for data operations, is used to describe the practices that automate the data engineering cycle. Today, just as organizations look to DevOps practices to streamline software development, a growing number of organizations are embracing DataOps to improve and formalize their data management practices—leading to better, more accurate data and more complete analytics.

In this ebook, data professionals will learn how to apply Agile principles to the activities of data ingestion, modeling, and transformation in order to uphold rigorous governance, auditability, and maintainability, yet still push updates to production in a short amount of time. Read on to learn how your data team can create value faster by enforcing predictable delivery and change management of data, data models, metadata, and related artifacts, improving the quality and reducing the cycle time of data-driven projects.

You will learn how to:

- Apply the principles of continuous integration (CI) and continuous development (CD) to data
- Simplify the overall development/testing/production cycle for creating data models, managing metadata, and enforcing good governance
- Integrate different commercial and open source tools into the processes used to create data pipelines, and more

INTRODUCTION: SIZING UP THE PROBLEM

Modern businesses generate immense amounts of data, making it progressively more difficult for data professionals to keep up with demands. Data flows in from enterprise software applications, SaaS solutions, social channels, mobile devices, IoT sensors, and more. As a result, many data professionals face huge backlogs for creating new data sets and modifying existing ones.

According to a May 2021 survey by Forrester Research of more than 4,000 data decision-makers across a variety of industries, 88% of data strategists across the world admit they are neglecting either their technology and processes or culture and skills—or both. While 70% of data decision-makers are gathering data faster than they can analyze and use it, 67% say they constantly need more data than their current capabilities provide. “Firms today are generating, demanding, and collecting more data than ever before, but overwhelmed data teams are struggling to analyze and secure that data,” Forrester sums up.[1]

Many of these data professionals are stymied by manual procedures for everything from modeling data to engineering pipelines to verifying the quality of new data sets. Whether it’s managing dev/test environments or documenting changes, critical data management tasks are often slow, disconnected, and error-prone. Even simple data requests can take months to complete, hindered by the laborious process of capturing, staging, testing, and moving data into production settings.

Further complicating matters is the need for proper data governance and security. Business users want new data fast, but data teams must enforce corporate data-access policies and observe proper data privacy controls. To reduce project backlogs, data professionals need automated procedures for designing, testing, and orchestrating database models, schema, hierarchies, libraries, and reference data as well as auditing team activities.

In recent years, principles from the DevOps world, initially created to encourage Agile software development, have been applied to data modeling and design. A newer term, DataOps, short for data operations, is used to describe the practices that automate the data engineering cycle. Today, just as organizations look to DevOps practices to streamline software development, a growing number of organizations are embracing DataOps to improve and formalize their data management practices—leading to better, more accurate data and more complete analytics.

How do you apply the principles of continuous integration (CI) and continuous development (CD) to data? Is there a way to simplify the overall development/testing/production cycle for creating data models, managing metadata, and enforcing good governance? What is the best way to integrate many different commercial and open source tools into the processes you use to create data pipelines?

This ebook explains how you can apply Agile principles to the activities of data ingestion, modeling, and transformation. It describes how data teams can uphold rigorous governance, auditability, and maintainability, yet still push updates to production in a short amount of time. Read on to learn how your data team can create value faster by enforcing predictable delivery and change management of data, data models, metadata, and related artifacts, improving the quality and reducing the cycle time of data-driven projects.

DEFINING TERMS AND CONCEPTS

DataOps helps data teams to reduce development times, increase data quality, and maximize the business value of data by bringing more rigor to the development and management of data pipelines. DataOps engineers don't only work with the data itself. They also engineer processes that allow other data professionals to build data products. These activities not only increase the scale of data analytics development, but also improve quality and boost staff productivity.

If you compare today's data workflows to the flow of raw materials in a factory, DataOps can be likened to supply chain planning. It covers everything from laying out material flows to anticipating production capacity to minimizing quality defects. Just-in-time manufacturing ensures that production capacity can be quickly scaled to satisfy the flow of orders, so that the factory can efficiently meet customer demands. Similarly, DataOps ensures that business users have accurate, governed data moving through data pipelines to meet the needs of analytics, data science, and other downstream applications. Instead of managing material workflows, DataOps engineers manage the flow of data. That includes automating, monitoring, and auditing the processes that are required to keep data current, accurate, and available. Without careful planning, orchestration, and logistics, there is increased risk that data pipelines falter, defects escalate, and customer satisfaction suffers.

Well-coordinated data management procedures enable businesses to ensure data quality, track versions, enforce data privacy regulations, and keep data applications moving through a continuous cycle of development, integration, testing, and production (see Figure 1).

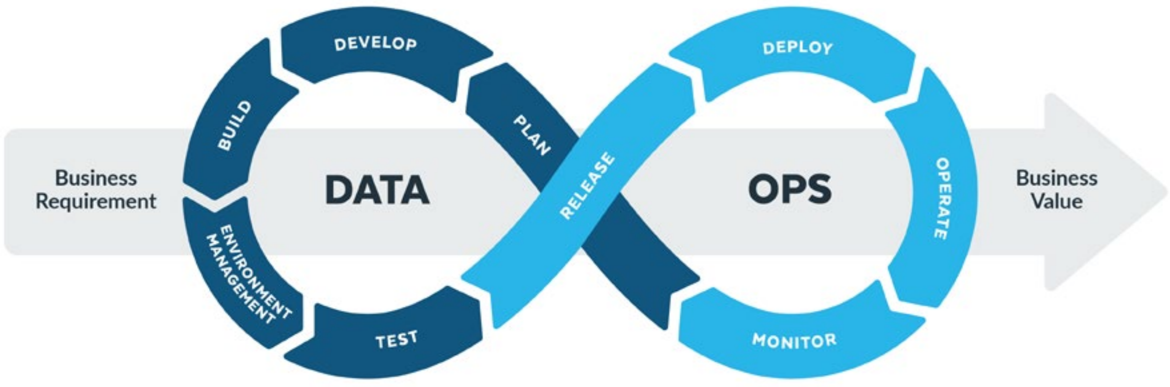


Figure 1: DataOps automates critical data engineering activities and orchestrates handoffs throughout the data management cycle, from plan, develop, build, manage, and test to release, deploy, operate, and monitor.

WHO IS ON THE DATAOPS TEAM?

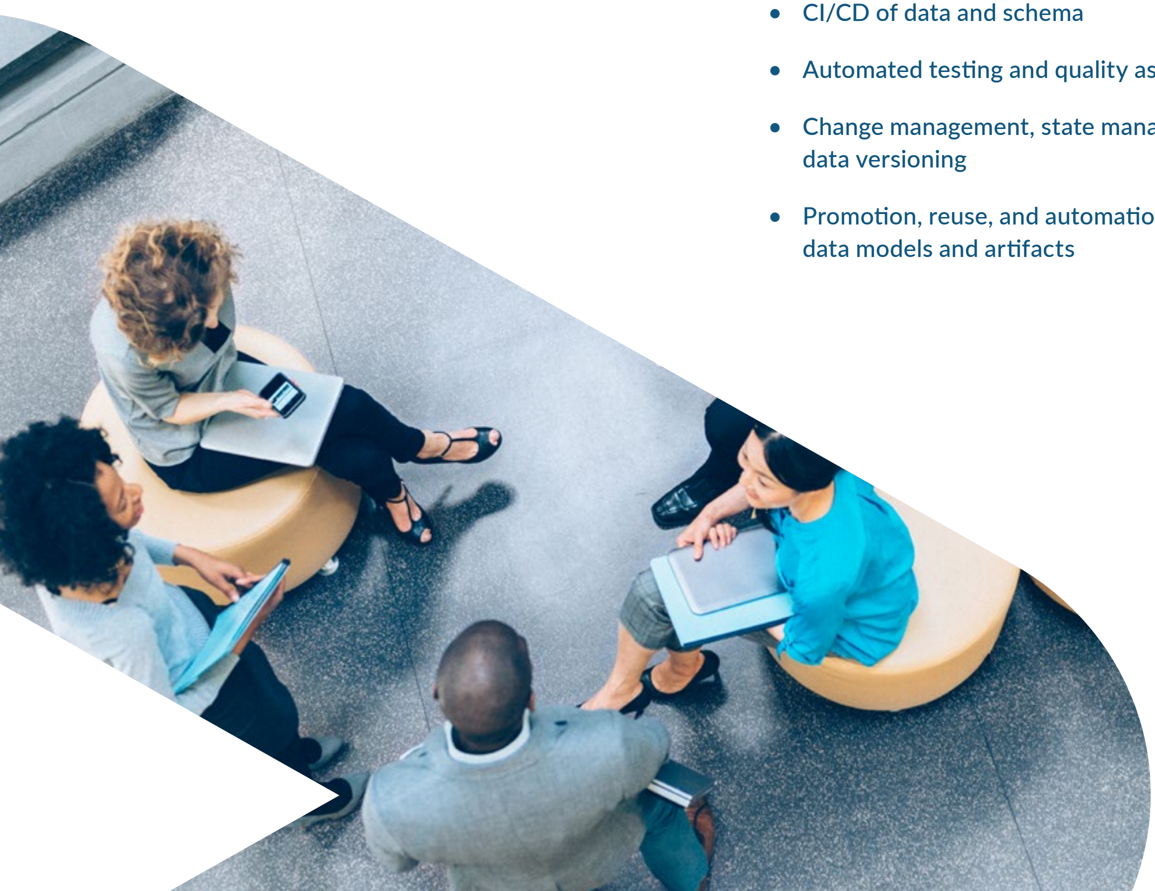
DataOps brings together data engineers, data scientists, business analysts, and other data professionals to apply Agile best practices to the data lifecycle, from data preparation to data science and advanced analytics. Their expertise is especially valuable in data-intensive industries such as technology, healthcare, finance, and retail.

LEARNING LESSONS FROM THE DEVOPS PLAYBOOK

DevOps has become the standard for how progressive organizations develop and manage applications. It thrives via peer-review processes for developing, testing, and deploying new code. Successful organizations that have adopted DevOps frameworks are pushing hundreds or thousands of code changes a day into production.

DataOps applies these same concepts to the creation and maintenance of data—not just the data itself, but also the schemas that govern that data. As DataOps methods grow in popularity, data professionals can mimic the achievements of today's DevOps teams in the following important areas:

- Agile development throughout the DataOps "loop": testing, debugging, deployment, production
- CI/CD of data and schema
- Automated testing and quality assurance (QA)
- Change management, state management, data versioning
- Promotion, reuse, and automation of data models and artifacts



COMPARING DEVOPS TO DATAOPS

DataOps is the application of DevOps principles to data. While DevOps is all about creating better synergies between development and IT operations teams, DataOps helps data professionals model, maintain, and share data by shortening the feedback loop among developers, testers, and other stakeholders. These principles are especially helpful to data scientists and data engineers as they create machine learning (ML) models that support AI/ML workflows.

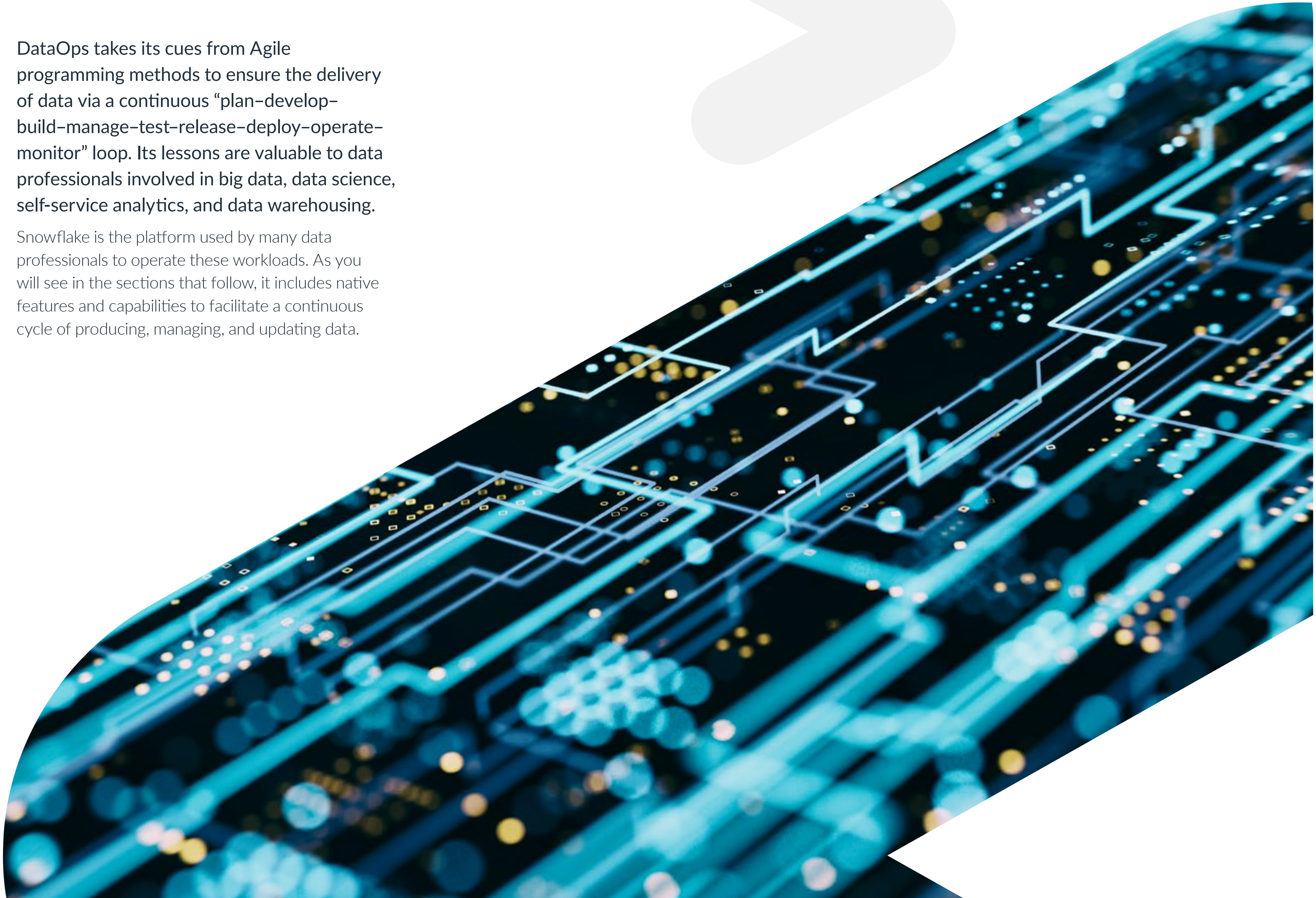
- DevOps automates version control and application configurations to streamline software development. It allows for continuous, component-based development of data and analytical pipelines, with automation to orchestrate data cleansing, data transformation, data matching, and data integration.
- DataOps automates data acquisition, modeling, integration, and management to deliver fast, trustworthy insights for data-driven applications. It applies automation to the entire data lifecycle, from data preparation to reporting. Most important, it facilitates collaboration by recognizing the interconnected nature of data analytics teams and IT operations teams.

As with DevOps, DataOps employs Agile methods to shorten delivery lifecycles and encourage cross-functional collaboration among multiple teams. DataOps also uses many different types of automation to accelerate new development.

HOW SNOWFLAKE HELPS WITH DATAOPS INITIATIVES

DataOps takes its cues from Agile programming methods to ensure the delivery of data via a continuous “plan–develop–build–manage–test–release–deploy–operate–monitor” loop. Its lessons are valuable to data professionals involved in big data, data science, self-service analytics, and data warehousing.

Snowflake is the platform used by many data professionals to operate these workloads. As you will see in the sections that follow, it includes native features and capabilities to facilitate a continuous cycle of producing, managing, and updating data.



COMBINING DATA, ELIMINATING SILOS

With Snowflake you can make your data centralized or decentralized—whichever your organization prefers—and this can include structured, semi-structured, and unstructured data. Centralizing your data reduces the number of stages the data needs to move through before it becomes actionable, which simplifies data pipelines and accelerates productivity.

Why is this important? Data silos make it difficult to know what data is available and how to wrangle it from various sources. Pulling all your data into a centralized repository allows you to establish a consistent metadata layer, which makes tracking and working with data easier.

In addition to this primary data store, Snowflake allows you to access and use data in external tables—read-only tables that reside in external repositories and can be used for query and join operations. DataOps teams can leave data in an existing database or object store, yet apply universal controls, as if it were all in one cohesive system. Snowflake supports the object stores of all major cloud providers and will soon support open table formats as well, such as Apache Iceberg (in preview).

BALANCING GOVERNANCE AND AGILITY

Having all your data in one place dramatically simplifies management, administration, and governance—knowing precisely what data you have, where it resides, who is authorized to access the data, and how those authorized are permitted to use it. However, you don't want to curtail productivity by imposing data governance procedures that are too onerous or obtrusive.

According to a white paper from TrueDataOps.org, ungoverned self-service data preparation creates a "Wild West" scenario of inconsistency and reinvention of data sets throughout the enterprise. Businesses may end up with multiple copies of the same fundamental data, each prepared by different teams, potentially yielding different results. "The more complicated and governance-heavy these data environments are, the less agile they become under the weight of all that governance, security, and privacy," the paper states. "This is because of the largely manual methods we use today to build and manage these environments."[2]

Ultimately, data teams need to provide a managed environment that allows the user community to be creative, while also ensuring that data is secure, consistent, and compliant with corporate policies and data privacy regulations. This type of governance is much easier to achieve when all database objects are centrally maintained and updated by a centralized data platform.

Ideally, the data platform should apply fine-grained governance across all the different objects, not just the database, and those governance policies should always be replicated with the data. Managed self-service procedures, backed by continuous DataOps delivery methods, help ensure that data is clean, accurate, and up to date.

FUNDAMENTALS OF DATA GOVERNANCE

As described in *Cloud Data Engineering for Dummies*,[3] a complete data governance practice should pay close attention to thefollowing functions:

- **Data access**: Data access rules determine who can inspect and manipulate the data, with particular attention to sensitive data and personally identifiable information (PII), which may need to be masked or tokenized to uphold data privacy regulations.
- Data quality: Business users need to know whether the data in reports, dashboards, and predictive models is correct. In the case of consumer data, the business may face penalties and fines if data is mishandled. Thus your data governance framework must incorporate solid data security and change control procedures.
- Change management: Data stewards commonly monitor who makes changes to databases and data pipelines, and how these changes impact the applications that depend on those assets. Change management tools allow them to audit usage so they can minimize the chance of errors and deletions.
- Data lineage: Being able to track where data originated, where it is used, who has access to it, and what changes are made to it over time is important when you need to trace errors back to the root cause in a database or analytics process. For example, if you delete a field in a table, how will it impact the reports and dashboards that display that data?
- Data catalog: One way to track the lineage of your data and monitor who can access that data is to create an information schema, also known as a data catalog. This gives you visibility into the metadata to see how data objects are being accessed, changed, and moved. It helps data professionals find pertinent data as well as determine which data is relevant for particular use cases.

SIMPLIFYING THE DATAOPS CYCLE BY CLONING DATA

In traditional data workflows, data professionals create individual environments to house development, test, and production databases, and then physically copy tables and data among them—a costly and time-consuming process. As these databases grow, it becomes progressively more difficult to obtain a copy of the production data for development and testing purposes because it requires too many hardware and software resources to replicate and copy data.

Instead, today's DataOps teams use cloning technology to create multiple copies of a database without incurring extra storage costs. For example, Snowflake Zero-Copy Cloning allows you to develop, test, and deploy a database into production without physically copying the data. With a single command, you can create logical copies of a database of any size. Using CI/CD processes, you can more easily apply changes to a development environment, test new features, promote those features to QA, conduct user-acceptance testing, and deploy a new version into production. Instead of physically copying data, you can clone as many environments as you need to enable this cohesive DataOps cycle.

CASE IN POINT: LARGE CPG COMPANY

Consider the situation at a large consumer packaged goods (CPG) firm, which streams hundreds of thousands of data flows each day from 85 different markets and dozens of regional systems into one global system. Its DataOps team includes 300 full-time employees and nearly 100 developers.

Previously, this CPG firm had separate environments for development, testing, quality, and production. Today, they use Snowflake Zero-Copy Cloning to create virtual data environments. Each environment has its own security and encryption, yet all are managed uniformly as one master data set. Moving a 10 terabyte database from testing to production takes only a few seconds. No data is copied and no extra storage is required—Snowflake uses pointers to handle these operations, and all changes are tracked within the clone. This allows DataOps professionals to build up and tear down new environments in seconds.

This flexibility enabled the company to combine six regional and one global data warehouse into a single Snowflake account. Thanks to Snowflake's near-limitless compute and storage resources, which can be scaled independently, various members of the DataOps team can work in parallel without impacting the performance of the other teams.

The elastic nature of the Snowflake cloud service also boosts efficiency. Previously, a simple stress test on a database was prohibitively expensive. Today, being able to conduct these tests before going live with a production database—without copying data—reduces costs while minimizing risk. All access controls, tags, data aggregations, data masks, and other artifacts are cloned along with the data set.

With near-limitless scalability of compute and storage resources, large ingestion and reporting jobs can be executed simultaneously. Automated data flows and serverless tasks allow data professionals to incorporate testing and quality checks inside these processes, while row-level security restricts access by country.

INHERENT AUTOMATION IN THE SNOWFLAKE PLATFORM

By separating compute from storage and delivering features like Zero-Copy Cloning, Snowflake provides the underlying infrastructure to uphold the basic DataOps principles of agility, maintainability, security, and governance. Here are some other ways in which Snowflake enables and extends DataOps practices:

- Snowflake provides near-limitless compute and storage resources for development, iteration, testing, and QA activities. Data professionals can easily create accounts, scale capacity, and spin up new clusters to support these activities. In the background, Snowflake automatically handles provisioning, availability, tuning, data protection, and other critical operations.

- Snowflake's inherent support for all types of data simplifies data management. For example, it can natively process structured, semi-structured, and unstructured data from enterprise applications, machine-generated data from IoT systems, streaming data from social media feeds, weblog data from mobile apps, and image and audio files. Snowflake can efficiently ingest all these data types using standard SQL, while also supporting popular open source data-ingestion methods, including batch integration and streaming integration with Apache Kafka.

- Snowflake's automatic schema detection speeds up modeling and design work for data formats that contain metadata. For example, a Parquet format might include nested fields that represent countries, counties, cities, and towns. A relational format may contain thousands of table definitions. Snowflake can extract this information as the data is loaded. Snowflake can also read the data in external tables such as Iceberg. Some customers choose to leave data in the external table, and bring the metadata into Snowflake.

- The Snowflake architecture is sometimes described as infrastructure as code. That means the entire configuration can be defined and stored in a version control system. This allows DataOps professionals to create and modify the environment from the definitions in the code, including data structures, roles, and definitions governing how compute clusters will be deployed in production. Tracking these configurations can be difficult with other large data platforms, and may require manual deployment and administration.

- Snowflake Time Travel logs all changes into immutable micro-partitions that make it easy to restore a database to any point in time. For example, a DataOps professional can look back 10 seconds or 10 days to see how a data set changes and evolves. Snowflake continuously tracks changes to the data itself as well as changes to the structure of the data.

DATAOPS AUTOMATION WITH SNOWFLAKE

Snowflake automates critical DataOps functions as part of the basic cloud services that all customers receive. For example, Snowflake automatically:

- Encrypts all data, at rest and in motion, with industry-standard encryption
- Re-keys the encrypted data on a regular basis
- Scales data workloads up and down elastically to accommodate fluctuating demands
- Replicates data in dev, test, and production databases to ensure business continuity
- Sets up change data capture (CDC) procedures to keep multiple databases in sync
- Partitions data, tunes SQL queries, and optimizes performance to maximize the efficiency of DataOps teams
- Applies software updates, such as security patches, as soon as those updates are available, allowing data professionals to focus on being productive with their data
- Detects schema as raw data is ingested

FACILITIES FOR CREATING AND MONITORING DATA PIPELINES

A data pipeline represents the encoded flow of data from source to consumption. Some data pipelines are simple: they might merely export financial data from a general ledger system into a CSV file structured for analytics. A more complex pipeline might move a group of tables from multiple sources into a target database, merge common fields, and parameterize the data for rapid access through an executive dashboard.

Snowflake simplifies pipeline development by allowing data engineers to orchestrate multiple streams and tasks. A stream is a special object type that uses CDC technology to track the ongoing changes in a table, including inserts, updates, and deletes as well as data manipulation language (DML) changes. Streams allow analytic apps to query just the incremental changes rather than the entire data set. A task defines a recurring schedule for executing SQL statements, including statements that call stored procedures.

You can chain tasks together to support complex processing scenarios. In a continuous data pipeline, tasks can use streams to process new or changed data.

Snowflake's Snowpark processing engine allows data engineers to develop data pipelines using popular programming languages, including Java, Scala, and Python (in preview), and then process the data directly inside of the Snowflake platform. Because data does not leave Snowflake, governance and security are easy to maintain.

More pipeline observability and UX features are also coming to Snowflake. For example, as data is loaded into Snowflake, operations personnel can use the Snowflake Ingestion Dashboard (in preview) to inspect the data and monitor ingestion events, errors, completion status, and other variables through a graphical user interface.

Snowflake Materialized Views aggregate data to speed up performance, such as grouping data at the day level, week level, and month level. Once established, Snowflake updates these views and aggregations in the background when the underlying data changes.

SIMPLIFYING ADMINISTRATION ACROSS MULTIPLE CLOUDS

Tracking changes in CI/CD environments can be difficult when your data set incorporates external tables in multiple public clouds, such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform. For example, how do you ensure that the same security configurations and administrative techniques apply to all of your cloud providers? Will you have to resolve differences in audit trails and event logs? Will you need to deal with different rule sets, or work with multiple key management systems to encrypt data?

As a unified code base spanning all the major cloud platforms, Snowflake simplifies these operations. You don't need to hire people with unique skill sets or maintain familiarity with multiple clouds. You can move data among major public cloud platforms without having to redefine and reconfigure the data environment for each cloud, and manage everything through one common interface.

BROADENING YOUR DATAOPS PRACTICE

While the basic architecture of Snowflake facilitates DataOps practices, and many of its features uphold fundamental DataOps principles, many Snowflake customers also use third-party DataOps tools that assist with orchestration, testing, monitoring, and other tasks. These DataOps teams incorporate proven software services such as GitHub, Terraform, Dagster, and other tools to create a culture of continuous improvement.

It's easy to plug third-party tools into Snowflake, as well as to leverage popular development languages and frameworks such as Python, Node.js, Go, .NET, Java, and SQL. Snowflake can also leverage native cloud capabilities from Amazon, Microsoft, and Google to streamline version control, operations management, change management, and CI/CD.

Whether you use Apache Airflow to create and manage your data pipelines, Jenkins for change management, Alation to assist with data governance, or dozens of other options, Snowflake can "meet you where you are" on the DataOps journey. Snowflake also works with strategic partners, such as Datalytyx, to complement and extend the innate capabilities of the Snowflake platform.

You can also work with a broad ecosystem of partners who have built seamless integration with Snowflake. Learn more.

CONCLUSION: MAKE HEADWAY ON THE DATAOPS JOURNEY

Good DataOps procedures enable companies to ensure data quality, properly manage versions of data, enforce data privacy regulations, and keep data applications moving through a continuous cycle of development, integration, testing, and production. As Forrester notes, firms can become data champions by partnering with vendors that can help them balance technology and cultural changes: "Vendors can do this by helping firms adopt the right end-to-end technology to break down silos and align data environments and assisting them in fostering the right data culture."[4]

To learn more about how Snowflake can help your firm become a data champion, visit snowflake.com.

WHY PURSUE DATAOPS?

- Accelerate the delivery of data-centric applications
- Reduce the cost of managing data throughout its lifecycle
- Improve data quality, security, and governance

ABOUT SNOWFLAKE

Snowflake delivers the Data Cloud—a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **Snowflake.com**.

CITATIONS

[1] bit.ly/3hxI5z4

[2] truedataops.org

[3] bit.ly/35vcLOO

[4] Forrester Research, op. cit.