



5 SELF-SERVICE DATA CHALLENGES AND HOW TO OVERCOME THEM



CHAMPION
GUIDES

EBOOK

TABLE OF CONTENTS

- 3** Why Self-Service Data is a Business Imperative
- 5** Challenge #1: Delays in Time to Value from Too Many Compute Engines
- 6** Challenge #2: Letting Domain Teams Work with Their Tools of Choice
- 7** Challenge #3: Discovering and Accessing Data Locked Up in Silos
- 8** Challenge #4: Governing Data Across Clouds
- 9** Challenge #5: Managing Users and Costs
- 10** Conclusion
- 11** About Snowflake



WHY SELF-SERVICE DATA IS A BUSINESS IMPERATIVE

Modern enterprises generate huge volumes of data in every corner of the business—from marketing and finance to supply chain management and shipping. That's great for improving operations or customer experiences, but it leaves IT and engineering teams buried in requests to ingest new data sources and deliver reports for a variety of business functions. Additionally, most domain teams only utilize a relatively small slice of this data because it is locked away in silos that limit discoverability and accessibility.

That's where the concept of self-service data comes in. Self-service data means giving employees across a variety of business functions the ability to easily access, integrate, and analyze previously siloed data sets in order to deliver business line reporting, access advanced analytics,

and build innovative data applications. By doing so, organizations can increase their efficiency, improve customer experience, access new markets, and create new revenue streams.

The State of California offers a **compelling case study** for the enormous improvements in user experience and agility that can be unlocked by self-service data. At the outset of the COVID-19 pandemic, the California Department of Technology (CDT) opted to make Snowflake its single source of truth for COVID-19 data, including the current number of cases across the state, the number of suspected cases, how many people were receiving treatment in a hospital ICU, the number of hospital beds and ventilators in use, and the demographics of COVID-19 patients.

CDT built dashboards that draw on live data to let California residents run queries based on a number of parameters.¹ For example, people could find out the current number of hospitalizations in their county and how that broke down by age and race—with the assurance that the data was always up to date.

Then, in early 2021, the agency launched an application that let Californians submit a request for a digital vaccine record; if the query finds a match inside Snowflake (where the California Department of Public Health's vaccine data resides), a QR code is returned along with the submitter's name and vaccination information.² This can be used to demonstrate proof of vaccination in lieu of carrying around a card that's too large to fit in a wallet easily. High-volume applications like this require speed and data freshness—CDT delivered both to millions of constituents across the state.

While self-service data is enormously valuable, there are challenges involved in managing access to data across different clouds and sources in a secure, governed manner and nailing performance across the workloads that need that data. By enabling large numbers of users to run concurrent workloads and offering the ability to scale computing resources up or down as needed, Snowflake addresses the challenges of processing, finding, and controlling data in order to make it self-service for each domain team's needs. This ebook will explore five specific issues that Snowflake solves for:

- 1 Delays in time to value from too many compute engines**
- 2 Letting domain teams work with their tools of choice**
- 3 Discovering and accessing data locked up in silos**
- 4 Governing data across clouds**
- 5 Managing users and costs**



Self-Service Data Challenge #1

DELAYS IN TIME TO VALUE FROM TOO MANY COMPUTE ENGINES

Domain teams are often bogged down by the limitations of the processing engines they need to accomplish their respective tasks. This leads to organizations deploying dozens of different platforms and solutions to better optimize around latency requirements and support diverse compute needs. Since some engines only support certain workload capabilities, data can end up siloed in the wrong place, making it difficult for domain teams to obtain timely insights or build data-powered products.

Snowflake helps solve this problem by offering a single elastic performance engine capable of accessing, processing, analyzing, and publishing data. This lets an organization set up independent virtual warehouses for different teams across a consistent platform. These virtual warehouses can be scaled up independently of one another, allowing for a much higher velocity of development and diverse sharing of data sets. Each domain team can have a dedicated set of resources that can scale with their needs. Within one Snowflake account, an administrator can grant individual domain teams their own virtual warehouses. This isolated compute resource can scale up, down, and across to meet their needs. Further, this architecture can be used to track the project costs to determine ROI on a project or use case basis.

Compute engines are typically designed for specific data types or use cases, such as normalized, denormalized, or unstructured data, which also leads to maintenance problems. Drift between different versions of the same data becomes likely, making it hard to join and analyze data across engines. Multiple ingestion, exfiltration, and transform pipelines are required to move data between engines.

Snowflake simplifies data architectures by removing specialized components for large-scale data exploration, interactive data applications, or needle-in-the-haystack searches across large tables. This architectural consolidation enables organizations to leverage **Snowsight**, a query editor with data visualization. Instead of connecting to a dedicated editor or visualization tool, Snowsight helps you quickly explore data and create reporting prototypes, as well as find and leverage your colleagues' previous queries, resulting in faster delivery of new insights.



Self-Service Data Challenge #2

LETTING DOMAIN TEAMS WORK WITH THEIR TOOLS OF CHOICE

According to Gartner®, “the proliferation of augmented capabilities, cloud and other accelerants within data access, data management, analytics, and data science makes once-distinct markets collide. Fueled by cloud approaches, the collision has ushered in a transformation in not only how data and analytics is used, but also who can effectively leverage it.”³

There are two major forms of interaction in data processing: SQL-based and DataFrame API via a programming language such as Scala or Python. While a flexible system like Spark can support running both types of interaction, it's difficult to manage the system so that it can handle low-latency queries while also performing cost-effective batch workloads. On the other hand, pure SQL systems like Presto are much easier to scale and manage but are limited in the types of queries they can run.

Most organizations today run both types of systems, segmenting low-latency, ad hoc workloads from batch workloads that can tolerate more variable latency. But running both types of clusters introduces far more complexity to manage, scale, and govern data access across the diverging engines.

Snowflake addresses this problem through its Snowpark developer experience.⁴ Snowpark supports development in Scala and, more recently, Python,⁵ using DataFrame-style APIs. This means that users don't have to leverage both a real-time, multiuser SQL system and Spark; they can stay in the Snowflake ecosystem for their data processing needs, leveraging the same scaleout, governance, performance, and isolation in one system, instead of using two different platforms. They can also continue to program in the language they're most comfortable with, increasing productivity.

This is a win for IT teams since it's more complex to offer Spark to a wide range of users as a query backend than an SQL-only system, which offers convenient interfaces (Spark requires the use of additional technologies, such as Apache Livy). The upside for users is getting to use their preferred programming language within a familiar DataFrame-style API instead of being forced to learn a new language or rely on SQL for more complicated tasks.

Snowflake also recently started supporting UDFs, which help make self-service data a reality by letting engineers add functionality into SQL queries. IT only needs to create the function once and then maintain it, not write every query that uses it themselves.

Once created, UDFs can be shared between users or across the organization to enable specific tasks—all without engineering intervention. For example, users

can leverage a currency conversion function to specify that all monetary amounts be converted into a single currency (such as U.S. dollars) before being graphed for analytics purposes.

With the foundation of Snowpark's DataFrame API and UDFs in place, domain teams can build and deploy powerful data applications, such as embedded analytics, that integrate rich data insights into the context of an application so customers can see visualizations in their workflow. In practice, this could be as simple as an HR platform that lets sales managers instantly query the data platform for associates' sales volume, customer service comments, or other data. Managers wouldn't have to use separate tools or export custom reports into a different engine to access these insights.

Domain teams can also lean on Snowflake's high-concurrency, low-latency performance updates⁶ to deliver otherwise impossible results, such as allowing millions of California residents to query a data warehouse and obtain instant, up-to-date COVID-19 insights as described in the **California Department of Technology example**. Without Snowflake's unique capabilities, enabling residents to verify their vaccination status would have been considerably more complicated, requiring a process of exporting data sets to other engines. It also would have involved multiple technologies, resulting in a much longer development lead time.

Self-Service Data Challenge #3

DISCOVERING AND ACCESSING DATA LOCKED UP IN SILOS

Today, most organizations store their marketing data, finance data, supply chain data, operations data, and other data across siloed data marts and unorganized data lakes, which can stymie cross-functional collaboration. If data is shared between domain teams, there are often multiple copies of the same data and a lack of governance on private or sensitive data. This practice leaves domain teams in the dark about information and insights that are theoretically available to them or wading through a swamp of incomplete and inaccurate information.

Snowflake helps eliminate silos and helps domain teams access and organize company data, including raw data stored in data lakes. Specifically, Snowflake natively supports structured, semi-structured, and unstructured data⁷ in the same system, which means that domain teams can interact with and produce analytics from inputs like JSON for the first time.

With Snowflake, domain teams can also identify popular data sets and their access history, providing insight into who typically uses them and for what purpose. In this way, organizations can learn which social media, image, or metadata is most useful to each functional area, informing longer-term data strategies and business decisions.

As data sets come together, understanding data lineage and object dependencies can become a challenge. By housing data in a single location, Snowflake ensures that data relationships can continue unimpeded and that there is a single source of truth along with clear provenance for every data set.

The power of Snowflake extends beyond a company's own data sets, too. Through **Snowflake Data Marketplace**, domain teams can connect with new data inside and outside their Snowflake account and collaborate with thousands of partners, suppliers, and applications in the Snowflake Data Cloud. Snowflake customers are also able to spin up their own internal data exchanges to streamline collaboration and ensure privacy where appropriate. For example, a holding company with multiple business lines can create a private exchange in which its portfolio brands can share data to the holding company for entity-wide analysis while keeping each data set separate from other portfolio brands. The holding company can apply governance policies to protect the data and apply masking features to ensure only those with rights to sensitive data can access it.



Self-Service Data Challenge #4

GOVERNING DATA ACROSS CLOUDS

In a multi-cloud architecture, simply identifying what data an organization has and how it's being used can be challenging—let alone monitoring and securing sensitive data within that organization and across cloud environments. This is especially true for multinational conglomerates in which cloud vendor selection may be market-specific, with one market using AWS, another running on Microsoft Azure, and yet another on Google Cloud.

Fueled by regulatory and compliance concerns, as well as by the goal to enable better decision-making, organizations are increasing their focus on how data is used and monitored.

According to a Gartner prediction, “by 2023, 95% of Fortune 500 companies will have converged analytics governance into broader data and analytics governance initiatives.”⁸

But defining unified data governance policies can be challenging. Security and governance policies tend to be custom-built for each workload, analytical system, and storage location, making them inflexible—and contributing to policy proliferation.

Snowflake addresses these problems by letting organizations define policies once and deploy them universally across multiple clouds, ensuring consistency in self-service data access. In this way, policies can be consistently applied not only across different clouds but also across different workloads and geographic regions. Standard authentication procedures deployed across applications ensure that no matter what tool or service an employee uses, their access to data is consistent. This approach works to protect data in motion and at rest.

With Snowflake, organizations can establish multiple account and role types, each with their own level of access and permissions. Features like dynamic masking, in which sensitive data is partially or fully hidden using constant value, hash, or custom functions, ensure security without changing the underlying stored data. Further, by keeping only one live copy of data as a single source of truth, companies can limit the potential for data leakage.

In short, Snowflake enables cross-cloud data governance so companies can choose their preferred tools and integrate different business units without disruption.



Self-Service Data Challenge #5

MANAGING USERS AND COSTS

Remember the old adage: Just because you can doesn't mean that you should. Concurrent workloads can expend a massive amount of computing power, and spikes in usage can push organizations over their contract limits and incur costly overages. When promoting collaboration and access to data, IT needs to work with the business to ensure a positive return on their data strategies.

Snowflake addresses this challenge with a consumption-based pricing model in which warehouses can be spun up and down as needed, and payment is calculated based on seconds of compute. Features like auto-suspend and auto-resume allow Snowflake to automatically turn off resources when query processing completes. Recent innovations have improved various aspects of Snowflake's performance, including storage compression rate efficiency, which further lowers costs for data stored in Snowflake.

Contract models are only one consideration when it comes to the cost of self-service data; usage across a distributed workforce is another. Organizations employing self-service data access have struggled to gain visibility into usage across the organization and guard against runaway user costs. Here, too, the Snowflake platform provides solutions that offer transparency. The first mechanism is role-based access, which can be set to limit access and activity by various teams or by individual employees. The next is resource monitoring, including real-time monitoring dashboards with details on compute usage and queries and automated start/stop functionality to prevent overconsumption.

Demonstrating ROI for self-service data can be difficult and will be key to growth and adoption over time. This type of monitoring data can be used to create a chargeback model in which each functional group or business unit is billed for its own usage, making ROI calculations much simpler and business impact more transparent.

CONCLUSION

Self-service data can be extremely valuable to any business, but there are challenges to managing performance, security, and costs—especially across multiple engines and cloud environments.

Snowflake solves these challenges with an elastic performance engine that lets domain teams use their preferred tools, languages, and cloud environments to increase productivity and unlock the value of their data.

To learn more about self-service data and how Snowflake can help your organization modernize its data architecture, visit [the Snowflake self-service site](#).





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](https://www.snowflake.com)



© 2022 Snowflake Inc. All rights reserved. Snowflake, the Snowflake logo, and all other Snowflake product, feature and service names mentioned herein are registered trademarks or trademarks of Snowflake Inc. in the United States and other countries. All other brand names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective holder(s). Snowflake may not be associated with, or be sponsored or endorsed by, any such holder(s).

CITATIONS

¹ bit.ly/3ErCElK

² bit.ly/3EpzKHc

³ Gartner, "Data and Analytics Worlds Collide: A Gartner Trend Insight Report," Carlie Idoine, June 17, 2021. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

⁴ Currently in public preview.

⁵ Snowpark native support for Python currently in private preview.

⁶ Currently in public preview.

⁷ Currently in public preview.

⁸ Gartner, "4 Data & Analytics Trends CFOs Can't Afford to Ignore," Jackie Wiles, September 30, 2021.