


snowflake® talend

# BUILD A TRUE DATA LAKE WITH A CLOUD DATA PLATFORM

**A single source of truth that's secure, governed, and fast**

CHAMPION GUIDES

# TABLE OF CONTENTS

# THE DATA LAKE: INTENT VS REALITY

True data-driven organizations seek to extract all the insight from all their data to optimize every aspect of their business and better serve their customers. They collect and analyze more and more data from traditional sources, such as ERP, CRM, and point-of-sale systems, as well as from newer data sources such as weblogs, click-stream applications, IoT devices, and social media. In addition, organizations increasingly use data from third-party sources, or shared data sets, to enrich their own data and analytics.

However, it has historically been impossible or impractical for an organization to load all its data into a traditional data warehouse. Data from newer sources often arrives in semi-structured formats that require the data to be transformed and processed before it is loaded. Furthermore, the cost and complexity of storing large quantities of raw, unrefined data in a traditional data warehouse from an increasing number of sources is prohibitive.

The data lake emerged more than a decade ago. Its goal is to be a scalable, low-cost data repository for storing raw data from a diverse set of sources so it can be explored and refined. Then, subsets of the refined data are moved to other systems, including a data warehouse, to support high-performance analytics and reporting.

That simple idea is not so simple to implement in practice. Without the right technology and without high data quality and data governance, a data lake can easily become a data swamp—an isolated pool of data that's difficult to use, hard to understand, and practically inaccessible to most of the organization. The greater and more diverse the data in the data lake, the more significant the problem becomes, making it harder and harder to get meaningful insights and value from the data.

### THE PATH TO A BETTER SOLUTION

Building a successful data lake requires, from the very start, a design that embodies data stewardship, governance and security, and easy access to all data.

It's often assumed that a data lake requires technology that is deployed as an independent solution. However, a modern cloud data platform can bring advances in data warehousing, data management, and analytics solutions together to deliver on the original promise of a data lake.

Building a modern data lake requires technology that:

- Easily stores data in raw form
- Enables immediate exploration of that raw data
- Refines the data in a consistent and managed way
- Makes it easy to support a broad range of operational business reports and analytics



## WHAT'S DRIVING THE DEMAND FOR A MODERN DATA LAKE?

**At a macro level, the need to implement a modern data lake is driven by the intersection of three important trends:**

- SaaS applications are on the rise and data processing is moving to the cloud.
- Exponential amounts of data arrive at a fast pace, requiring business decisions to be made in real time before the data becomes stale.
- Data access controlled by IT teams is being replaced by self-service data access for business users.

But many of the data lakes that organizations have implemented have added new burdens to already overstretched IT teams. These projects lack the necessary flexibility, governance, and data management to keep up with newer forms of analytics.

# WHAT YOUR DATA LAKE SHOULD DELIVER

## A modern data lake should provide the following capabilities.

### DATA ACCESS FOR ALL USERS, REGARDLESS OF THE DATA SOURCE

Today's data comes from a wide variety of sources, including relational and NoSQL databases, IoT devices, and data generated by SaaS and enterprise applications. Bringing all this data together is a challenge for legacy platforms. As a result, different types of data are typically stored in different data platforms. That approach creates isolated islands of data, adding complexity and leaving potential insights hidden.

A key goal of a data lake is to bring together this data. However, many data lakes are implemented as their own data island, disconnected from other platforms such as data warehouses and data marts that support reporting and analytics.

The result is that only skilled data scientists and data engineers can access the data in the data lake. Other users must wait for those skilled users to prepare and export data for them. However, writing manual scripts to normalize this data into a standardized format for broader access can be time-consuming and costly. Relegating a data lake to only skilled users stifles an organization.

### GOVERNANCE, SECURITY, AND COMPLIANCE

With the explosion of data, more and more departmental line-of-business (LOB) users want access to data. Therefore, robust data governance is crucial to ensuring the right access is provided to the right data for the right usage. For example, each data type might contain information that must be managed and safeguarded in specific ways. Some data types also fall under stringent regulations required by healthcare legislation (HIPAA), PCI DSS, the EU's GDPR, and California's Consumer Privacy Act (CCPA).

The types of information that may require specific governance include:

- Credit card information
- Social Security numbers
- Dates of birth
- Addresses
- Email addresses
- IP network information
- Geolocation coordinates

Data in a data lake is not exempt from these requirements. Data governance ensures that data access and usage is managed, tracked, and secure when the data is stored inside a data platform and also during data loading and integration. Achieving this governance requires data management and governance tools that provide stringent access control, encryption of data both at rest and in motion, and auditable records of data access and data changes to support compliance requirements.

## A CONSISTENT AND RELIABLE VIEW OF DATA

Having more data is good, but not at the expense of data quality. In order to explore and experiment with data, data scientists need access to raw data before it has been cleaned and standardized. However, the rest of the organization needs a consistent, reliable view of data for reporting and analytics.

A data lake must support both exploration and the more rigorous demands of business reporting and analytics. That requires data to be standardized into the right format and have auditable metadata about where the data originated and when it was loaded. Data standardization and other components of data quality enable IT teams to provide self-service data access for LOB users, while metadata enables data subsets to be easily located and profiled or discarded (or cleansed) if the data quality is poor.

## SELF-SERVICE FOR ALL DATA USERS

The number of data analysts, data scientists, and other professionals wanting data-driven business insights is growing quickly. These knowledge workers are scattered across various departments of an organization, and all of them need data access. Supporting broad access to data in a data lake is impossible without the right approach and technology. Attempting to apply cumbersome traditional methods for making the data in a data lake available creates delays that frustrate and hinder users.

Self-service access has become essential to give growing numbers of users access to data. Making self-service possible requires a modern cloud data platform that eliminates complexity while ensuring properly managed and secured access. For example, the platform should not require manual deployment steps, should avoid the need for laborious tuning and optimization, and should support unpredictable workloads while ensuring concurrency. Without those capabilities, it is impossible for IT teams to provide self-service.

# USING A CLOUD DATA PLATFORM TO BUILD A TRUE DATA LAKE: FOUR ESSENTIAL CAPABILITIES

1. Unite Diverse Data Sources
2. Safeguard Through Governance
3. Ensure Data Quality
4. Enable Self-service

Using a cloud data platform that supports structured and semi-structured data at unlimited scale, you can build a true data lake that unites diverse data sources, provides governance, ensures data quality, and enables self-service for all users.

# 1. UNITE DIVERSE DATA SOURCES

## FLOW ALL YOUR DATA INTO A SINGLE PLATFORM

Data-driven companies rely on a diverse set of data sources: NoSQL and relational SQL data stores, SaaS and legacy applications, IoT data, and maybe unstructured data such as audio and video files. These data sources have different formats, data models, and structures, so you'll need a modern cloud data platform that makes managing and consolidating all this data as simple as possible.

Achieving that simplicity begins with getting all your data consolidated into a single location, and the cloud is the ideal integration point.

A modern cloud data platform makes it possible to implement a data lake to store diverse data in native form, at low cost. That makes it possible to bring together data from diverse sources without creating a new data island.

To be the right foundation for a data lake, a cloud data platform should do the following:

- Load and analyze raw data immediately, without requiring parsing or transformation prior to data loading. It should also enable you to query the data immediately after loading it into the data lake.
- Handle structured and semi-structured data. Both data types should happily coexist. Ideally, you should be able to create simple tables in the cloud data platform and stream structured and semi-structured data without manual coding or any other manual intervention.
- Use native SQL queries against structured and semi-structured types without requiring additional programming, and use the Schema-on-Read approach for the semi-structured data.
- Separate its compute and storage capabilities, so organizations can store massive volumes of raw data cost-effectively while deploying only the computing capacity needed.

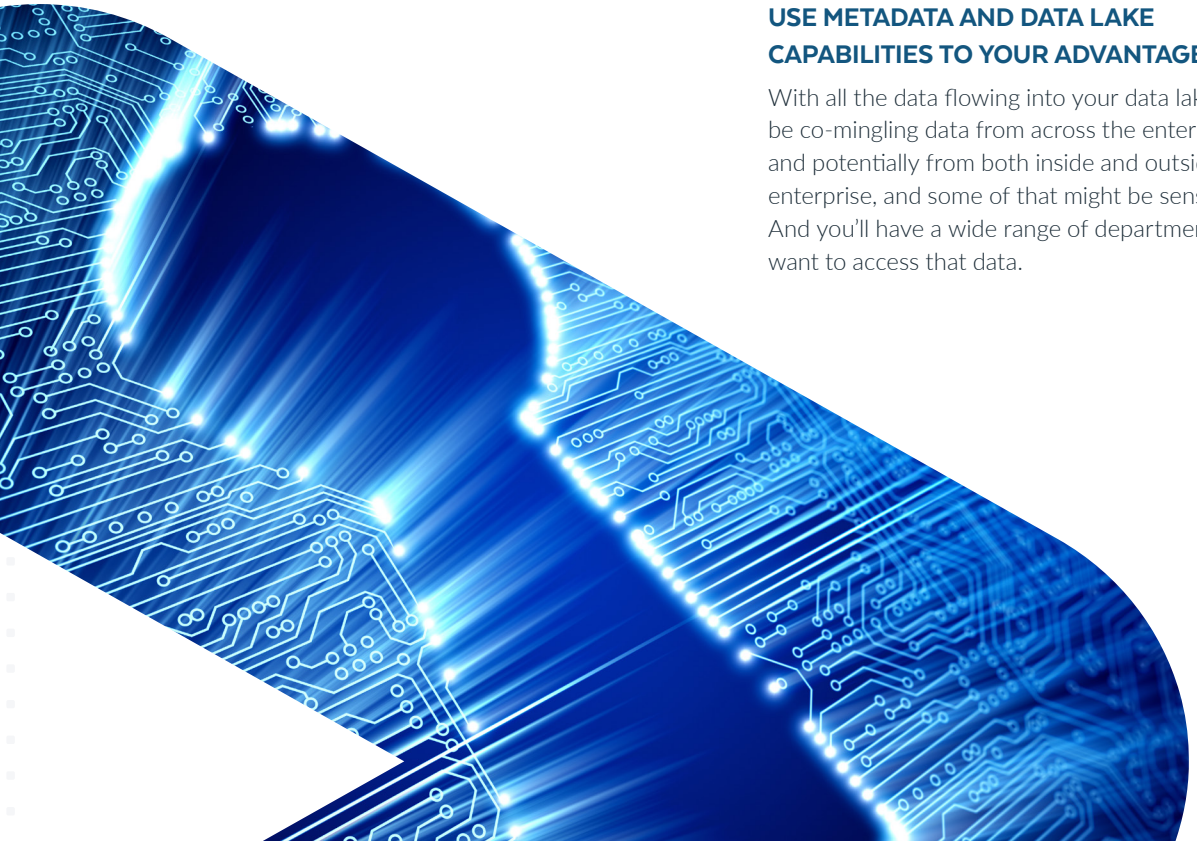
In addition, a modern cloud data platform should be addressable via a fully native connector to maximize multi-tenant functionality, thus supporting parallel data loading. No matter how many tables and data sources there are, an optimized connector ensures the data lake ingests data at maximum rate.

## ACCELERATE DATA LOADING

Data is constantly changing. As it comes in, data may need to be profiled, normalized, aggregated, and cleansed. Accelerating data loading is an essential step in uniting diverse data sources.

Best practices require that a solid platform for data ingestion be able to do the following:

- Connect to various data sources easily, without major programmatic scripting, so you can collect all the data from wherever it's located.
- Provide batch and streaming ubiquity to handle historical and real-time data loading and process data as it comes in.
- Scale for data volume and various data types to quickly onboard new data sources, such as data from third-party data providers, web click-streams, social media, and smart devices.



### MANAGE THE FULL DATA LIFECYCLE

To build the ultimate data lake, a cloud data platform should provide data integration tools that allow you to run all parts of the data lifecycle: profiling, aggregating, normalizing, and cleansing the data. With those capabilities, you can shorten the time to value by making it easier to collect and organize disparate data.

## 2. SAFEGUARD THROUGH GOVERNANCE AND SECURITY

### USE METADATA AND DATA LAKE CAPABILITIES TO YOUR ADVANTAGE

With all the data flowing into your data lake, you'll be co-mingling data from across the enterprise and potentially from both inside and outside of the enterprise, and some of that might be sensitive data. And you'll have a wide range of departments that want to access that data.

With the range of possible use cases, there are numerous data governance aspects to consider to ensure users not only see fresh and accurate data, but they also see only the data they're permitted to access. A data lake solution should enable effective governance by providing the following:

- Metadata that identifies where the data came from, who touched the data and how, and what relationships exist between various data sets
- The ability to curate the data through stewardship and preparation by people who can accurately qualify it
- A collaborative data governance process, instead of authoritative top-down governance, to enable IT and business teams to ensure information stored in the data lake is accurate, making the data lake a trusted, single source of truth

### ENABLE GOVERNANCE WITH A CLOUD DATA PLATFORM

Putting in place governance controls requires the support of the underlying data platform. To support its use as a data lake, a cloud data platform must:

- Ensure robust data encryption and key management for all data
- Provide and enforce granular access control configured by user and by role
- Maintain records of actual and attempted data access

Metadata generated within a cloud data platform can also support governance requirements. For example, a cloud data platform can generate metadata about data and make the data accessible for querying.

### BUILD SECURITY INTO YOUR DATA LAKE

Security is always of paramount importance, particularly when personal information is stored in a data lake. Properly managed, the security capabilities provided by a modern cloud data platform can be a much more effective and less expensive option than attempting to manage a security infrastructure yourself.

A modern cloud data platform ensures security by:

- Implementing standards-compliant security protocols
- Ensuring granular access control to operations and data
- Providing security by default
- Always using up- to- date best-of-breed security practices, including the use of AES-256 encryption and compliance with SOC1/SOC 2 Type 2 standards
- Being certified to meet the demands of industry-specific requirements such as HIPAA, FedRAMP, and PCI DSS

## 3. ENSURE DATA QUALITY

### BUILD A DATA LAKE THAT SUPPORTS BI

Supporting business intelligence (BI) is often the goal behind building a data pipeline, but that's often a struggle if you attach your BI stack to a poorly designed data lake. A data lake that does not ensure data quality makes delivering a self-service environment difficult if not impossible. If a data lake doesn't perform the right cleansing operations on the data, it can't provide accurate BI.

A modern data lake, built on a modern cloud data platform, is fast and nimble. It offers a single, integrated system for easily storing and accessing vast amounts of data. It's the place where data is selectively exposed to BI professionals and many other users from across the organization.

With all your diverse data sources and metadata integrated in a single platform, users can accomplish their BI tasks and be confident in analytics results. In addition, a high-performance cloud data platform provides users with the scalability and flexibility they need for data lake exploration, without forcing them to move data to a different system to get great performance.

### UNIFY DATA MANAGEMENT

To make your modern data lake a single source of enterprise data, you need a unified data management framework that:

- Ensures data quality and provides data masking
- Provides consistent operationalization to increase data quality and agility
- Supports all use cases and personas to increase productivity and collaboration across teams

A modern cloud data platform that includes an enterprise data management solution meets those needs. Because all the data resides in a unified data platform, it is managed consistently throughout the data pipeline, which simplifies data management and governance.

### USE A SUPERIOR APPROACH TO DATA TRANSFORMATION

The data platform serving the data lake should offer the ability for you to develop customized transformations. This contrasts with being limited to a library of predefined transformations that are plugged in—the typical approach of legacy data integration technology. Legacy approaches don't give you the flexibility you need to deliver the necessary structure for your ever-changing data schemas.

An infinitely scalable cloud data platform supports transforming data in-database, provides low-cost storage independent of compute resources, and even offers governed access to third-party data or analytics services through secure data sharing. These functionalities enable you to leverage the scale and horsepower of the platform, accelerate data transformation, and ensure data governance and regulatory compliance.

## 4. ENABLE SELF-SERVICE

### ACCESS GOVERNED DATA THROUGH SCALABLE COMPUTE AND STORAGE RESOURCES

Your modern data lake will reach its maximum potential only if you can get data into the hands of more users. But doing so requires striking a balance. Your goal isn't necessarily to give everyone self-service access to everything in the data lake. Your goal is to have the controls and governance in place that allow self-service where it makes sense. To make sure you can deliver the right type of self-service at the right point, your data lake solution should have the ability to:

- Automate and eliminate complex deployment and configuration so users can rapidly access the resources and environments they need in order to work with data
- Make data accessible by deploying easy-to-use tools for LOB users who are making business decisions using data from the lake
- Implement governed self-service that offers general access to corporate information without chaos or risk
- Allow as many users as possible across the business to benefit from data-driven decision-making

## SCALE INSTANTLY WITH A CLOUD DATA PLATFORM

A modern cloud platform provides a new way to think about scaling storage and compute resources. It offers unlimited resources and elasticity so you pay only for what you need—by the month, week, day, or hour. You'll enjoy the same flexibility with storage, as well, scaling up or down almost instantly to meet user demands.

Instant scalability allows your data lake to handle any data volume and workload while providing concurrency to support any number of users during periods of peak demand. You can deliver fast performance to a larger number of users and provide multiple views of data tailored to specific needs—all within a single platform and without lag time—by:

- Scaling out storage to store large diverse data sets in an affordable manner independent of computing cycles
- Immediately scaling compute resources up or down to address surges in activity, such as users querying much larger than normal data volumes
- Easily resizing an existing compute cluster if you want to keep tight control over compute resources and costs
- Keeping a predefined compute cluster in suspended mode, so it's ready to go for a regular event that requires a burst of compute resources
- Automatically spinning up an additional cluster to handle concurrent users and avoid performance degradation and then spinning down that cluster when the load subsides

With all these options, you'll pay only for the resources you use, without the administration headaches that exist when you try to add more resources on the fly using legacy technologies.

# EXPEDITING PREDICTIVE AND PRESCRIPTIVE INSIGHT

Companies are always looking to identify what will drive business in the future. They want to move from relying on lagging indicators to benefitting from leading indicators. A data lake built on a cloud data platform can help you expedite predictive and prescriptive insights, because you can bring in nearly any internal or external data source, whether it's central or tangential to the business. Then, business users and data scientists can perform deep exploration to discover new performance indicators and fulfill the mandate for predictability, without having to connect to new data sources or execute complicated data transfers. For example, a modern data lake can answer questions such as the following and millions more:

- When will a certain piece of capital equipment break down?
- Where will we get the most ROI from advertising placements on any given day?
- How can we optimize truck routes for an upcoming storm?
- How many packages of peanuts should be stocked on a particular flight?

With a modern data lake, the opportunities for predictive discovery are endless.

# FIND OUT MORE

If you want to give your organization a data lake that's deeper and broader than ever before, while maintaining essential control and governance, you can do so today. The right data lake platform is the modern cloud data platform that provides a data management solution to equip your company with new levels of new insight.

Find out more about how to build the ultimate data lake today. Visit snowflake.com and talend.com.


## ABOUT SNOWFLAKE

Snowflake's cloud data platform shatters the barriers that prevent organizations of all sizes from unleashing the true value from their data. Thousands of customers deploy Snowflake to advance their organizations beyond what was possible by deriving all the insights from all their data by all their business users. Snowflake equips organizations with a single, integrated platform that offers the only data warehouse built for the cloud; instant, secure, and governed access to their entire network of data; and a core architecture to enable many other types of data workloads, including a single platform for developing modern data applications. Snowflake: Data without limits. Find out more at **snowflake.com**.

## ABOUT TALEND

Talend is a next-generation leader in cloud and big data integration software that helps companies become data driven by making data more accessible, improving its quality and quickly moving it where it's needed for real-time decision making. Talend's open-source, native, and unified integration platform, Data Fabric, enables customers to embrace new innovations and scale to meet the evolving data demands of the business.