snowflake | talend

Build a True Data Lake with a Cloud Data Warehouse

A SINGLE SOURCE OF TRUTH THAT'S SECURE, GOVERNED AND FAST

CHAMPION GUIDES

snowflake | talend

What's inside:

snowflake | talend

The data lake: Intent versus 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 logs, web applications, Internet of things (IoT) devices and more.

However, that's only possible with a single repository to easily and efficiently store all your data and make it useful. But it's not possible or practical to load all data into a traditional data warehouse. Data from newer sources often arrives in semi-structured formats, which require further transformation and processing before loading. Further, the cost and complexity of storing large quantities of raw, unrefined data in a traditional data warehouse from an increasing number of sources would be prohibitive.

The data lake emerged more than a decade ago to solve this problem: How to create a scalable, low-cost data repository for storing raw data from a diverse set of sources to explore and refine that data. Then, move subsets of the refined data to other systems, including a data warehouse, to support high-performance analytics and reporting.

That simple idea is not so simple in practice. Without the right technology and without proper data quality and data governance, a data lake can all too easily become a data swamp—an isolated pool of data difficult to use, hard to understand and practically inaccessible to most of the organization. The greater and more diverse the data put into the data lake, the more significant the problem becomes, making it harder and harder to get meaningful insights and value from that 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 the data. It's often assumed a data lake requires different technology, deployed as its own independent solution.

However, advances in data warehouse and data management technologies now make it possible to deliver on the original promise of a data lake, but without divergent technology that falls way short and results in a new data silo. Building a modern data lake requires technology to easily store data in raw form, provide immediate exploration of that raw data, refine it in a consistent and managed way, and make it easy to support a broad range of operational business reports and analytics.

Demand for the 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 every pace, requiring business decisions be made in real-time, before the data becomes stale.
- IT-controlled data access is migrating to self-service access for business users.

IT teams feel the brute force of this rapid, pervasive change. But many of the data lakes organizations have implemented have only 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.

snowflake | talend

What your data lake should deliver

THE ESSENTIALS FOR AN INDUSTRY-LEADING ENTERPRISE

BIG DATA THAT DOESN'T COST BIG

Today's data comes from a jumble of sources including relational and NoSQL databases, IoT devices and data generated by SaaS and enterprise applications. Bringing all this data together remains 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 a very small number of 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 extremely time-consuming and costly. Relegating the data lake to only sophisticated data scientists and developers will stifle 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 may contain information that must be managed and safeguarded in specific ways. Some data types also fall under stringent regulations required by the healthcare legislation (HIPAA), Payment Card Industry Data Security Standard (PCI DSS), and the European Union's (EU) General Data Protection Regulation (GDPR). The types of information that may require specific governance include:

- Credit card information
- Social Security numbers
- Date of birth
- Address
- 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 secured at all times. This must happen when the data is stored inside a data platform, and during data loading and integration. Achieving these standards 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 changes to data to support important compliance requirements.

A CONSISTENT AND RELIABLE VIEW OF DATA

More data is good, but not at the expense of data quality. Sophisticated data scientists need access to raw data before it has been cleansed and standardized in order to explore and experiment with that data. Yet, the rest of the organization needs a consistent, reliable view of data for their reporting and analytics.

The ultimate 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, with auditable metadata about where the data originated and when it was loaded. Data standardization and other components of data quality are necessary to allow IT to deliver self-service for line of business (LOB) users, while metadata enables data subsets to be easily located, profiled and discarded (or cleansed) if the quality is poor.

SERVE ALL YOUR DATA USERS

The number of data analysts, data scientists and other professionals wanting data-driven business insights is growing by leaps and bounds. Many of these knowledge workers are scattered across all 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 data available when that data resides in a data lake creates ever-growing 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 the combination of a data platform and a data management solution that eliminates complexity while ensuring properly managed and secured access. For example, the data platform needs to bypass manual deployment steps, avoid the need for laborious tuning and optimization and support unpredictable workloads and concurrency. Without those capabilities, self-service is impossible for IT to support.

snowflake | talend

Building a true data lake

YOUR FOUR PILLARS

snowflake | talend

STEP 1

How to unite diverse data sources

FLOW ALL YOUR DATA INTO A SINGLE REPOSITORY

If your company is not a yet a data-driven company, it's likely transforming into one by relying on a diverse set of data sources: NoSQL and relational SQL data stores, SaaS and legacy applications, and even IoT data. These data sources have different formats, data models and structures. You'll need modern platforms and tools that make managing and consolidating all this data as simple as possible.

A HIGH-PERFORMANCE DATA LAKE IN A CLOUD DATA WAREHOUSE

Simplicity is key when building a data lake. Achieving that simplicity begins with getting all your data consolidated into a single location. The cloud is the ideal integration point for that data.

Modern cloud data warehouse technology makes it possible to implement a data lake inside a data warehouse to store diverse data in native form, at low cost. That makes it possible to implement a data lake that brings together data from diverse sources without creating a new island.

To be the right foundation for a data lake, the modern cloud data warehouse should:

- Load and analyze raw data immediately, without requiring parsing or transformation prior to 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 warehouse and stream either structured and semi-structured data, without hand coding or any manual intervention.
- Write SQL queries against structured and semi-structured types without additional programming. A modern cloud data warehouse should natively speak SQL and realize schema-on-read for the semi-structured data.
- Separate compute and storage so organizations can store massive volumes of raw data cost-effectively, while deploying only the computing capacity needed.

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

REQUIREMENTS FOR DATA LOADING

Data is constantly changing. As it comes in, data may need to undergo the same steps of being profiled, normalized, aggregated and cleansed. Accelerating data loading is an essential step in uniting diverse data sources. Best practices for creating a solid platform for data ingestion include:

- Connect to various data sources easily, without major programmatic scripting, so you can collect all your data from wherever it's located.
- Batch and streaming ubiquity—Handle historical and real-time data loading and process data pipelines as they come in.
- Scale with volume and variety—Quickly onboard new data sources, such as data from third-party data providers, web clickstreams, social media and smart devices.

MANAGING THE FULL DATA LIFECYCLE

To build the ultimate data lake, you also need data integration tools that allow you to run all parts of the data lifecycle: profiling, aggregating, normalizing, and cleansing the data. With that, you can shorten the time to value by making it easier to collect and organize all this disparate data.

snowflake | talend

STEP 2

Safeguard through governance

USE METADATA AND DATA LAKE ARCHITECTURE TO YOUR ADVANTAGE

With all the data flowing into your data lake, you're now co-mingling data from across the enterprise and potentially from both inside and outside of the enterprise. And you have a wide range of departments that now want to access that data, which may include sensitive 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 also see only the data they're permitted to access. The data lake solution should enable effective governance by:

- Adding context to metadata to provide information such as: Where is the data coming from, who touched that data and how, and what are the relationships between various data sets?
- Curating the data through stewardship and preparation by people who can accurately qualify it.
- Enabling a collaborative data governance process. Instead of authoritative top-down governance, the data lake solution should offer a collaborative approach that allows IT and the business to ensure information stored in the data lake is accurate. In this way, you can build structured reliability into the data lake, making it a trusted, single source of truth.

ENABLING GOVERNANCE IN THE CLOUD DATA WAREHOUSE

Putting in place these controls needs the support of the underlying data platform. To support use as a data lake, a cloud data warehouse must do the following:

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

Metadata generated within a cloud data warehouse can also support governance requirements. For example, a cloud data warehouse can generate metadata about data loaded into the data warehouse and make that data accessible to queries.

From data integration to data quality

BUILD A NIMBLE, PRODUCTION-READY DATA LAKE

Supporting business intelligence (BI) is often the purpose behind building a data pipeline. But it's often a struggle if you try to attach your BI stack to a poorly designed data lake. A data lake without data quality makes delivering a self-service environment difficult if not impossible. Without the right cleansing operations on the data, it doesn't provide accurate business intelligence, which can be detrimental to an organization.

A modern data lake, built on a modern cloud data warehouse 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 located and integrated in a single system, your users can accomplish their BI tasks and be confident of their analytic results. In addition, a high-performance cloud data warehouse solution provides users with the scalability and flexibility they need for data lake exploration, without being forced 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 offers:

- Pervasive data quality and data masking as part of the solution.
- Consistent operationalization to increase data quality and agility.
- A single platform for all use cases and personas, to increase productivity and collaboration across teams.

The combination of a modern cloud data warehouse with an enterprise solution for data management provides a solution that meets these needs. Data is managed with a consistent approach throughout the data pipeline, and it resides in a unified data platform that supports and simplifies data management and governance.

A SUPERIOR APPROACH TO DATA TRANSFORMATION

The data management platform serving the data lake should offer the ability to develop customized transformations. This contrasts with being limited to a library of pre-defined 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 warehouse supports transforming data in-database, and it provides low-cost storage independent of compute resources. This supports the ability to leverage the scale and horsepower of the data warehouse and accelerate data transformation.

Enable self-service

ACCESS GOVERNED DATA THROUGH HIGHLY 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 deliver abilities 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.
- Scale the operationalization of data-driven decision making, to allow as many users as possible across the business to participate.

LIMITLESS SCALABILITY IN THE CLOUD DATA WAREHOUSE

A modern cloud warehouse provides a new way to think about scaling storage and compute resources. Modern cloud data warehousing offers unlimited resources and the elasticity to 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.

Limitless scalability allows your data lake to handle any data volume, workload and concurrency. 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. Multiple ways exist to easily and dynamically scale up, down and out (concurrency), without lag time to meet peaks and valleys of demand:

- Scale out storage to store large diverse datasets in an affordable manner independent of computing cycles.
- Immediately scale compute resources up or down to address surges in activity, such as users querying much larger than normal data volumes.
- Easily resize an existing compute cluster if you want to keep tight control over compute resources and costs.
- Keep a predefined compute cluster in suspended mode, ready to go if you have a regular event that requires a burst of compute resources.
- Automatically spin up an additional cluster to handle concurrent users, avoiding performance degradation as concurrency increases. When the load subsides and the queries catch up, the second cluster will automatically spin down.

With all these options, you'll pay only for the resources you use but without the admin headache that exists in legacy technologies when trying to add more resources on the fly.

BUILT-IN SECURITY FOR YOUR DATA LAKE

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

What are the characteristics of a modern cloud data warehouse that ensures security?

- Implements standards-compliant security protocols.
- Ensures granular access control to operations and data.
- Provides this technology to thousands of customers, by default.
- Always up to date with best-of-breed security practices, including use of AES-256 encryption and compliance with SOC 2 Type 2 standards.
- Certified to meet the demands of industry-specific requirements such as HIPAA and PCI DSS.

Turn the tide with predictive insight

NAVIGATE BUSINESS CHANGE BY DISCOVERING NEW PERFORMANCE INDICATORS

Companies are always looking to identify what will drive business in the future. They want to move from a reliance on lagging indicators to leading indicators. The data lake can help expedite predictive and prescriptive insight. You can bring in nearly any internal or external data source, central or tangential to the business, to be tapped for potential insight.

A data lake built on a modern cloud data warehouse, and comprising various integrated data sources, enables users to quickly obtain new, transformative business insights. From there, business users and data scientists can delve into deeper exploration, without having to connect to new data sources or execute complicated data transfers. They get easy access to data sets never thought possible before.

FULFILLING THE MANDATE FOR PREDICTABILITY

The goal of operationalizing predictive analytics is top-of-mind for today's data-savvy business leaders. A modern data lake can simplify access to an enormous amount of data used to unlock an organization's predictive and prescriptive power, to answer questions such as:

- When will a certain piece of capital equipment break down?
- Where will we get the most return on investment (ROI) from advertising placements on any given day?
- How can we optimize truck routes around a coming storm?
- How many packages of peanuts should be stocked on a particular flight?
- And millions more.

With a modern data lake, the opportunities for predictive discovery run deep.

Find out more

MAKE YOUR DATA LAKE NAVIGABLE, GOVERNED AND FAST

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 warehouse integrated with a modern data management solution to equip your company with unfathomable levels of new insight.

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

About Snowflake

Snowflake is the only data warehouse built for the cloud. Snowflake delivers the performance, concurrency and simplicity needed to store and analyze all data available to an organization in one location. Snowflake's technology combines the power of data warehousing, the flexibility of big data platforms, the elasticity of the cloud, and live data sharing at a fraction of the cost of traditional solutions. Snowflake: Your data, no limits. Find out more at **snowflake.net.**

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.