snowflake

Within Reach:
The End of Your Struggle for Data

EFFORTLESS DATA LOADING, DATA INTEGRATION AND DATA ANALYTICS

CHAMPION GUIDES

Contents

snowflake

The Data Struggle is Real

Technology and the clash between BI and IT

For over a decade, the explosion of data, and broad demand to analyze it, have created tension between data-hungry users of business intelligence (BI) systems and the resource-strapped IT teams that support them. Whether your organization uses a legacy on-premises or cloud data warehouse, or a noSQL solution, all of these approaches create organizational conflicts that impede the success of your users and your enterprise.

THE END IS NEAR

As an IT professional or BI user, your data struggle is real. Chief Information Officers (CIOs) and Chief Data Officers (CDOs) feel it, too. But the data struggle can end, and quickly. Organizations that choose modern, built-for-the-cloud data warehousing gain more than the resources needed to become truly data-driven. By providing easy access to enterprise data, you'll be a champion among your BI users and your IT organization, while helping to streamline operations and better serve your company's customers.

This eBook dives into how your organization can benefit from adopting data warehousing built for the cloud, with high-level advice on how to end your struggle with:

- Data loading
- Data integration
- Data analytics

A real-world snapshot follows, illustrating before-and-after differences with a data warehouse built for the cloud. This eBook closes with a summary of the additional, high-level benefits a modern data warehouse can offer.

The Struggle with Data Loading

Silos, capacity constraints, resource contention and more

Right now, your company's data is scattered across hundreds or even thousands of data silos. That data stays where it is due to the extreme difficulty of moving it into a single physical or logical location, and the prohibitive cost to store it there. This reality creates a multitude of downstream problems for IT organizations:

- **CAPACITY PLANNING**: Let's say you're going to bring in a new data source that requires 10 or even 100 terabytes of storage. If that exceeds your system's current capacity, how difficult and costly will it be to expand?
- **PLANNING FOR NEW DATA SOURCES**: It's impossible to forecast, with pinpoint accuracy, which new data sources your organization may need, and how much data they'll want to consume. For example, you may start looking at data from a social media feed, or a third-party market research firm, or both.
- **RESOURCE CONTENTION**: On the flip side, you may have the storage capacity to hold all your data. But limited (shared) compute resources means there may not be enough hours in the day to load it all. In traditional warehouse architectures, data loading can only be done off-peak to avoid degrading query performance for BI users during normal business hours. But for global enterprises, there's no such thing as off-peak.
- **COMPLICATED TRANSFORMATIONS**: Data transformation poses another challenge in traditional warehouses, and yet another impediment to getting all your data into one location. The complexity of handling flexible-schema data types from a myriad of sources often causes IT to leave data in silos.

Even worse, most traditional platforms don't understand semi-structured data types. IT teams need to load them first into a noSQL environment, outside of the warehouse, and then write and run complex programs to flatten the data into a relational format. Only then can the data be loaded into the data warehouse, where it can be integrated with structured data.

In other words, loading the raw data into the warehouse is simply not an option. Instead, you're stuck with a sub-optimal data pipeline that requires too much time to massage the data into a useful form.

How Cloud Data Warehousing Can Streamline Data Loading

Unlimited, inexpensive resources and infinite flexibility

Modern, usage-based, cloud data warehousing allows you to skip the painful upfront capacity planning exercise because it can give you:

- Unlimited, inexpensive data storage and compute resources on demand.
- The elasticity to access these resources at any scale without impacting query or development and testing activity.

With the right cloud solution, you can avoid the legacy problem of overprovisioning for peak demand, and then getting stuck with an underutilized system the rest of the time. In the cloud, you'll pay only for what you need – by the month, week, day or hour.

SEMI-STRUCTURED DATA? NO PROBLEM!

The ability to land all your structured and semi-structured data of any size in one place, and not worry about, is unthinkable with traditional on-premises or cloud-washed data warehouse solutions. Yet, it's a core capability for a modern data warehouse. You can load semi-structured data into the warehouse directly, use SQL extensions to join it to structured data, and still get the optimized query performance you need.

What's more, once the data, structured or semi-structured, lands in a modern cloud data warehouse, it's immediately ready for query by people who just need access to raw data.

THE UPSHOT

With the right cloud data warehouse, your data loading processes can execute exponentially faster compared to a traditional on-premises or cloud-washed data warehouse. Data that took days to load can now be in your warehouse in minutes.

The Struggle with Data Integration

Data repositories, future overhead and the elusive "single version of the truth"

OK, you did it. You managed to get data into your traditional warehouse. Now you must prepare it for analysis. The next step, integrating data from multiple sources, entails another set of challenges for both your company's IT organization and BI users.

MAKING SENSE OF THE RAW DATA

If you're in IT, you need to figure out how to enable BI users to get value from your raw data. You'll need to connect the disparate sources so users can get the results they want.

In the traditional data warehouse, you'll have data from multiple sources, such as enterprise databases and your CRM system and general ledger, to name a few. To get a cohesive, cross-functional picture of your organization, you'll need to integrate and rationalize all of those pieces into a "single version of the truth."

But your assumptions about the queries BI users will want answers to will significantly influence the data integration plan. Second-guessing BI users' needs is difficult at best. What often happens is that users come up with more questions than IT organizations can possibly anticipate. Depending on how you've transformed and integrated the data, you may have to go back and re-engineer everything. That takes time and money.

PLANNING FOR UNKNOWN OVERHEAD

The way you integrate data will have downstream performance implications for the data warehouse and how fast BI users' queries can be fulfilled. For example, let's say you combine certain data sets to form a customer health metric. This key performance indicator (KPI) becomes extremely popular with support and account teams and is a fixture in the performance dashboards they consult.

The upside of your effort is that you've created a valuable new KPI to help run the business. The downside is that you've also created a form of "technical debt": The KPI's popularity will drive future demands on system performance, concurrency and the metric's underlying data preparation, which happens inside the warehouse. All of this complex integration work takes compute and storage resources in addition to everyday "background" overhead such as backups, and other standard administrative functions.

Altogether, it's a lot of overhead. Can your traditional data warehouse cope with it? Can you?

CONCURRENCY AND THE ILLUSION OF "A SINGLE SOURCE OF THE TRUTH"

Data integration is further challenged, simply put, because the numbers must add up. Every day, at world-class companies, equivalent reports pulled from the theoretically same data set can project vastly different results.

It's extremely difficult to maintain a single version of the truth in a traditional, on-premises or cloud-washed data warehouse. Such a system will constrain attempts to create a single, unified view of complex datasets. Its inability to quickly scale impedes large groups of concurrent users to efficiently access and analyze these data sets. The ability to scale also emerges whenever you execute complex joins that integrate the data. The more data sets you have to join, the more complex the join conditions will be, requiring ever-greater amounts of compute power. This often leads to attempts to offload the work into separate data marts with their own resources. This, in turn, results in different versions of the truth as new data silos evolve and are out of sync with the main repository.

How Cloud Data Warehousing Can Transform Data Integration

A single version of truth for all your data

The unlimited resources of a modern data warehouse can eliminate the inherent complexity and struggle of data integration. A modern cloud data warehouse can allow your organization to build a unified data environment where everyone can work from a single version of the truth.

Infinite scalability also eliminates worry over the impact of the future overhead caused by today's data integration strategy, or even the need to extensively plan for it. In a data warehouse built for the cloud, you're able to execute more complex queries on demand and at speeds faster than many highly tuned traditional systems. If you need more compute resources, they're available automatically or on demand with the click of a button.

In part, this means you have the capability to execute much of the complex integration using SQL views rather than having to hard code your assumption into brittle ETL code. This allows you to more easily iterate and test business rules and assumptions in preparing the data for consumption while minimizing the time needed to re-engineer if something changes.

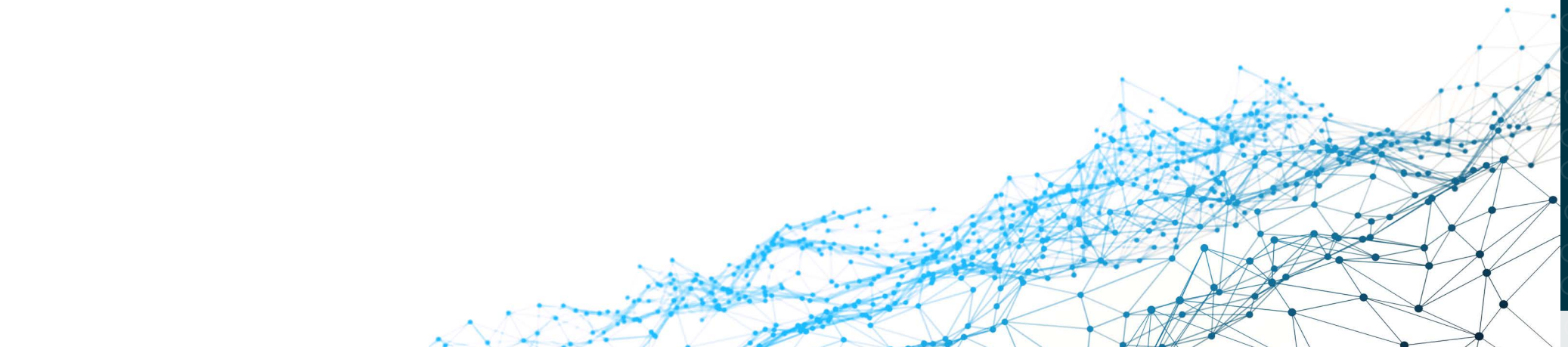
THE UPSHOT

With the right data warehouse, you can eliminate the most time-consuming data integrations, allowing the IT team and BI users to focus on their core work of respectively enabling and analyzing data.

In addition, you know you won't run into storage constraints or have a business unit that needs special accommodations because they have more data than anyone else. In a modern data warehouse built for the cloud, one that capitalizes on true cloud architecture, there are no limits.

STRUCTURED AND SEMI-STRUCTURED DATA, TOGETHER

By handling semi-structured and structured data in one system, you can apply integration logic within a modern cloud data warehouse instead of trying to apply pre-logic before the data arrives. With data integration entwined with data loading, transformation occurs in a way that's scalable and highly visible. It's easy to apply business logic, filters and rules in the modern cloud warehouse, as opposed to integrating data in a segmented legacy system that is inherently more complex and less transparent.



The struggle with data analytics

One word says it all: Slow

Legacy data warehouse technology is the source of your organization's struggles with data loading and data integration. With data analytics, the limitations of legacy technology are also clear. Performing analytics on data in an on-premises or cloud-washed warehouse can be painfully slow.

THE LIMITATIONS OF TRADITIONAL APPROACHES

First, users of an on-premises warehouse always face the issue of speed. Do I save money at the expense of slower disk technology, or, do I spend more for better performance? No matter the decision, the limits of infrastructure are always an issue with an on-premises data center. In the cloud, there are no limits.

Second, queries to a legacy system suffer because limited compute resources often become overloaded. The capacity for concurrent queries is quickly reached, thus slowing response times for everyone. Once a server hits its limit, response times for large, complex queries and support for increasing numbers of concurrent users will, at best, be problematic. Worse case? IT gets "fined" for missing its SLAs.

Again, it's important to distinguish between a warehouse built for the cloud and one that is merely cloud-washed. Despite residing in the cloud, the performance of a cloud-washed system is not much better than an on-premises data warehouse. Whether single- or multi-tenant, cloud-washed systems are not born in the cloud. They are ported versions of legacy architectures originally designed for on-premises environments and therefore unable to take full advantage of a modern cloud infrastructure.

DIMINISHED OPTIMIZATION

Both on-premises and cloud-washed data warehouses require a significant investment in query optimization. For on-premises, a team of database administrators (DBAs) is doing all of the:

- Optimization work
- Research
- Query profiling
- Managing of the performance indexes and partitioning schemes

With a cloud-washed data warehouse, there's still a team of DBAs managing performance behind the scenes. This correlates with legacy warehouse technology masquerading as a cloud data warehouse, offering little fundamental improvement.

THE UPSHOT

Whether your analytics run on a traditional, on-premises system or on legacy technology transplanted to the cloud, the underlying technology is incapable of executing as fast as a data warehouse built for the cloud.

How Cloud Data Warehousing can Transform Data Analytics

Unlimited resources to support unlimited scalability and concurrency

A true cloud data warehouse is delivered as software-as-a-service (SaaS), offering unlimited scalability up, down and out (concurrency) on demand. Scaling in either direction is effortless but only if warehouse compute and storage resources are truly separate. In addition, there must be a third, sophisticated metadata layer that orchestrates all of the work.

With this type of architecture, your organization can choose any size compute cluster to handle any query, data loading or dev/test job. You're not forced to dump data from a tightly coupled compute node before resizing. This also means you can scale down the compute power when a job is done, paying only for what you use.

NO CONTENTION OR DATA INCONSISTENCY

A modern data warehouse should similarly scale for infinite concurrency so there's no contention for compute resources. No longer will you need to worry that a spike in concurrent users, or any other compute activity, will grind query response times to a halt.

Finally, as previously described, by separating compute from storage, all users can access the same, single copy of the data. This eliminates the chance of data inconsistency, which occurs when multiple user groups copy the same data to data marts to speed query performance but use different rules.

THE UPSHOT

A modern data warehouse allows you to shift your focus from system management to pure analysis. It's a radically different and radically better way to manage the data your BI users depend on.

How Amino Eliminated its Data Struggle

Consumer healthcare search platform reduces analytics processing time from a week to less than an hour

With Snowflake, we went from waiting for seven days to doing anything with our data and analytics, to just under an hour."

— Bobby Chowdary, head of data engineering, Amino

COMPANY: Amino's healthcare database translates health insurance claim information into meaningful insights for consumers.

WHO: Bobby Chowdary, head of data engineering

DATA STRUGGLE: With a dataset of 215 million people, 900,000 providers, and over five billion patient and doctor interactions, about 90 percent of Amino's use cases involve ad hoc querying. This requires "full table scans, all the dataset, all the time," Chowdary says.

"We often had to ask our data scientists to stop doing what they were doing so we could run our batch jobs on the enormous dataset. This caused a lot of cross-functional headaches across the board and, most importantly, we weren't able to get to the information we wanted quickly and reliably."

THE NEED: Amino had an important batch job running on a Hive Hadoop cluster. The job analyzed all historical claims data from 2012 to the current date to create a holistic view of patient and doctor interactions. The job took roughly seven days to run on the Hadoop cluster. Amino needed the insight much faster.

THE SOLUTION: Snowflake, a SQL data warehouse built for the cloud from the ground up.

RESULTS: "Snowflake helps us reduce times and get to decisions very quickly," Chowdary says. "When we ran the batch job on Snowflake, we were surprised to see it complete in under an hour."

"That was a game-changer for us. Because this particular dataset is core to what Amino does, it's extremely important that we were able to hash that out and produce the dataset with minimal downtime."

BONUS: Since Amino handles protected health information (PHI), Amino is subject to stringent HIPAA data security compliance. "Snowflake also eliminates a lot of our operational and security headaches," Chowdary says. "Snowflake provides security for data at rest and on the wire, which is ideal because Amino handles sensitive information. Snowflake enables us to be HIPAA compliant. In addition to security, backups and user access are also automated."

Beyond the Data Struggle

More benefits from modern data warehousing

You've just read about the many advantages that built-for-the-cloud data warehousing can deliver to fundamentally improve data loading, integration and analytics, putting an end to your data struggle. You've read about how Amino now runs full-table queries in a fraction of the time required by its Hadoop cluster. The figure below provides an overview of the game-changing benefits of data warehouse technology built for the cloud.

LIMITLESS RESOURCE FLEXIBILITY

Cloud-built data warehousing can give you unlimited resources and the elasticity to access any scale of compute horsepower and data storage, paying only for what you need – by the month, week, day or hour. With cloud, you can avoid the legacy problem of overprovisioning for peak demand and getting stuck with an underutilized system the rest of the time.

A SMARTER ARCHITECTURE THAN A TRADITIONAL WAREHOUSE

The most advanced cloud data warehouses comprise separate layers for storage, compute and services. The services layer is a critical differentiator. It understands how your data is formatted (structured and semi-structured), it includes an optimized security subsystem, and it contains a metadata store with the statistical information about the data needed to automatically optimize workload performance. Most importantly, the services layer handles all the transaction management across the virtual clusters, allowing a consistent set of operations against the same data at the same time.

THE UPSHOT

A cloud data warehouse that truly takes advantage of modern cloud architecture can increase BI performance by up to 200x for a tenth of the cost of legacy data warehouse systems located on-premises or merely migrated to the cloud.

MULTIPLE CHOICES FOR PEAK PERFORMANCE

Cloud data warehousing can give you unlimited compute resources dynamically and without lag time. There are multiple ways to easily scale up, down and out (concurrency) to meet demand and pay only for what you use.

DATA ENCRYPTED IN TRANSIT AND AT REST

The right cloud data warehouse can protect data in transit and at rest, whenever it is sent over a network or stored on disk. This includes data files persistently stored, query results and the content of a local disk cache. A truly secure cloud data warehouse allows you to redirect resources dedicated to on-premises security to other strategic IT efforts.

FAST DEPLOYMENT AND AUTOMATIC UPGRADES

A cloud-built data warehouse can go live in weeks or just a few months, depending on the size of the project and the migration strategy from on-premises to cloud. Your organization can see benefits quickly, with lower upfront investment. With a modern cloud data warehouse, you can expect incremental updates every month, without service disruption.

Find out why Snowflake was independently ranked as the #1 cloud data warehouse

READ THE REPORT

GET THE POWER AND FLEXIBILITY TO MEET ALL YOUR BI USER DEMANDS

- The CFO who needs an answer right now.
- The sales team that makes a big push at the end of every quarter.
- The curious marketer looking for the most profitable customer journey.
- The data scientist who suddenly wants to stress-test her brilliant theory.
- The support team that wants to find out which engineer has the best-case resolution scores.
- The supply chain analyst diving deep into inventory turnover trends.

About Snowflake

Snowflake started with a clear vision: Make modern data warehousing effective, affordable and accessible to all data users. Snowflake delivers the performance, concurrency and simplicity needed to store and analyze all of an organization's data in one location. Because traditional on-premises and cloud solutions struggle with this, Snowflake developed a new product with a new built-for-the-cloud architecture that combines the power of data warehousing, the flexibility of big data platforms and the elasticity of the cloud at a fraction of the cost of traditional solutions. Snowflake: Your data, no limits.

Visit **snowflake.net**