The past, present and future of data warehousing

snowflake

HOW UNDERSTANDING FOUR DECADES OF TECHNOLOGY EVOLUTION CAN HELP YOU CHOOSE THE RIGHT SOLUTION TODAY

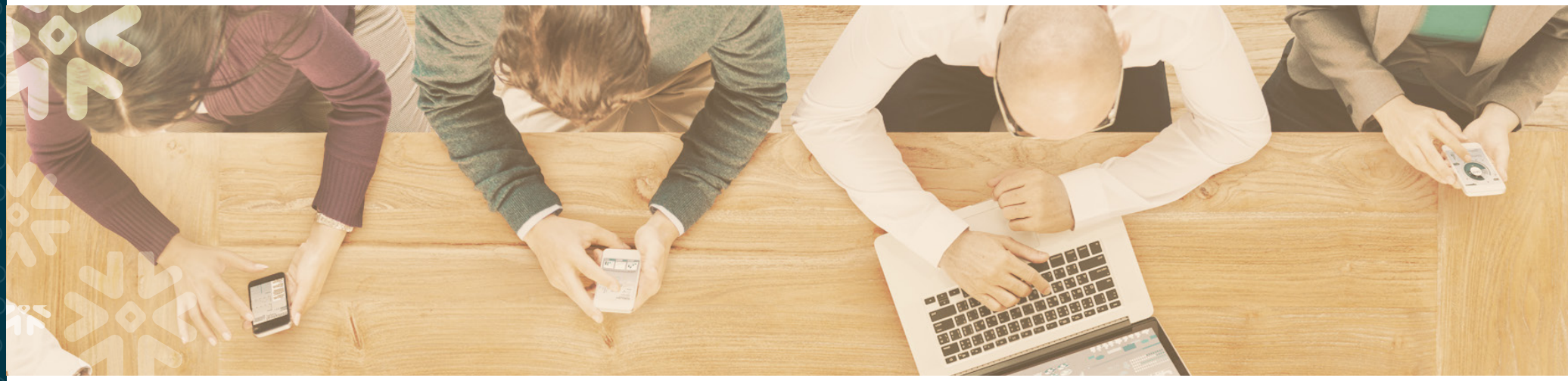
CHAMPION GUIDES

What's inside

snowflake

The need for the modern data warehouse

Every technology evolution tells a story



WHY THE DATA WAREHOUSE EMERGED

The forces that gave rise to data warehousing in the 1980s are just as important today. But from then until now, history reveals the benefits and drawbacks of the traditional data warehouse, the NoSQL options once seen as the replacement for data warehousing and, more recently, the cloud versions of these solutions. Each iteration solved an important problem but this evolution proves technology never caught up with the data demands of the enterprise.

More recently, "big data" hype has stirred healthy debate as to the relevance of the data warehouse. The emergence of NoSQL solutions in the past decade urged technology and data professionals to consider alternatives to the legacy data warehouse.

Understanding this history reveals why the data warehouse is even more relevant, not less, today. History also shows what a modern data warehouse must deliver: the best of legacy data warehouses, the best of NoSQL and big data systems, the true benefits of the cloud and much more.

The evolution of data warehousing also reveals the direct path you're on today: the need for a powerful, simple and affordable data warehouse built for the cloud to store and analyze all your data in one location.

In this ebook, we'll detail these key insights you can use today to champion modern data warehouse technology within your organization, helping it become a data-driven enterprise:

1. How traditional data warehousing defined the needs for today's modern data warehouse.
2. How the success and limitations of NoSQL systems such as Hadoop revealed the limitations of legacy data warehouses.
3. How the cloud revealed what's possible beyond "cloud-washing" an on-premise data warehouse or NoSQL system.
4. Why the modern data warehouse is the first solution to meet the data and analytics demands of today's enterprise.

The genesis of data warehousing

Why data warehousing emerged as the engine of data analytics

The 1980s ushered in the relational database. And it was good. Companies around the world used it to record what was happening in their organizations. The technology became so prevalent that enterprises began using relational databases to report on and analyze their business. However, IT groups soon discovered that a growing appetite for reporting and analytics had overwhelmed their transactional databases.

These new workloads were very different from the online transaction processing (OLTP) of relational databases. Enterprises started to offload these functions to separate databases optimized for reporting and analytics. Hence, the data warehouse was born.

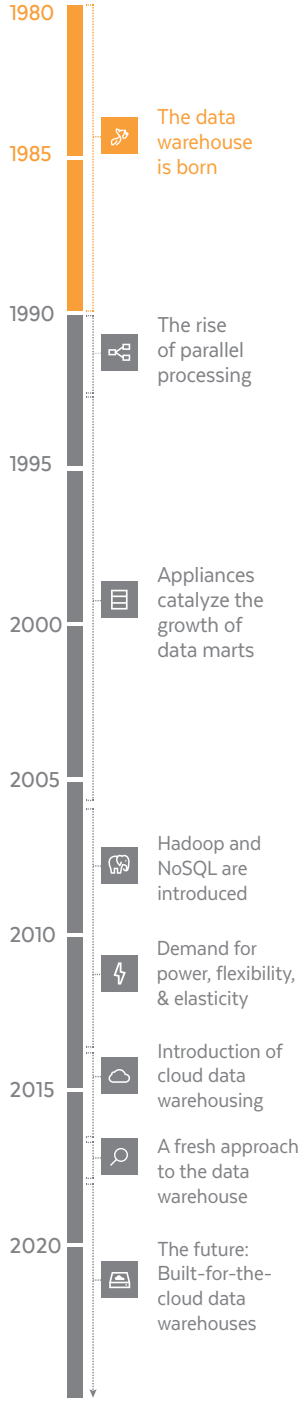
WHAT ARE SOME OF THE MAIN DIFFERENCES, AND SIMILARITIES, BETWEEN AN OLTP DATABASE AND A DATA WAREHOUSE?

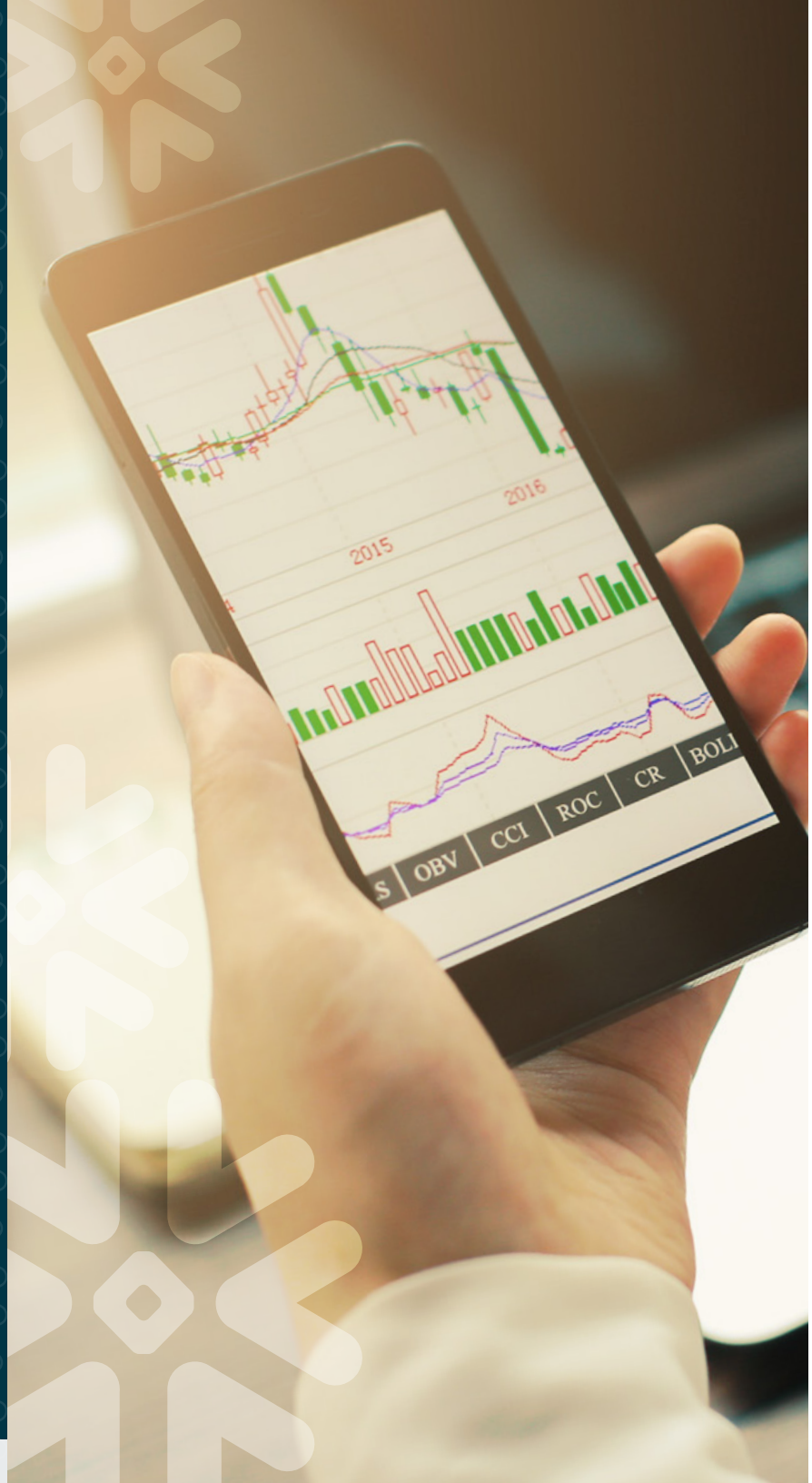
OLTP DATABASE

- A database designed for rapid storage and retrieval of small sets of current data records in support of transactions and interactions within an enterprise.
- Organizes data in tables and columns, and allows users access via structured query language (SQL).
- Designed to handle quick, real-time activity such as entering a customer name, recording a sale, and recording all accounting activity of that sale.
- Analysis is relegated to simple, static reports often driven by IT.

DATA WAREHOUSE

- A database designed to store and process historical data from multiple sources inside and outside the enterprise, including the cloud, for deep analysis.
- Organizes data in tables and columns, and allows users access via SQL.
- Optimized for loading, integrating and analyzing very large amounts of data.
- Designed for descriptive, diagnostic, predictive and prescriptive analysis.





WHAT BENEFITS EMERGED FROM STORING AND ANALYZING DATA IN A SEPARATE SYSTEM?

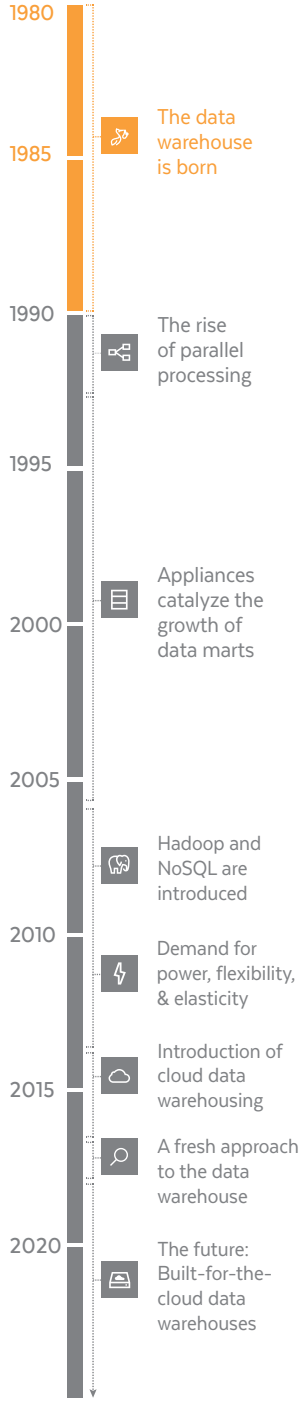
- Eased the burden of reporting from transactional systems.
- Produced more business-friendly data results.
- Provided access to a wider array of reports more quickly.
- Integrated valuable data from across the enterprise for richer insights.

WHAT WERE THE LIMITATIONS OF THESE EARLY DATA WAREHOUSE SYSTEMS?

- Getting data into the data warehouse was slow and problematic, often occurring only once a day and during off-peak hours.
- Having to plan and pay for the highest expected resource usage, when that period may only last a month, week or just one day of the year.
- These systems reached capacity quickly, impacting performance.
- Only the most technically astute could administer the data warehouse.
- The technology limited the breadth and depth of insight enterprises could act upon.

LESSON LEARNED

Data warehousing has demands and requirements that OLTP databases and systems not designed for those demands could not satisfy. Two decades later, "big data" systems such as Hadoop also came up short, which we'll discuss below.



Scaling the data warehouse

The rise of parallel processing

The amount of data and the complexity of workloads grew steadily throughout the 1990s, taxing the data warehouse beyond its limits. The early architectures for data warehousing weren't much different than OLTP databases: a big monolithic server attached to a big bunch of disk.

As data volumes grew, the architecture of the traditional data warehouse couldn't keep up. An individual server could only be so large before you experienced diminishing returns on performance at a rapidly escalating cost. The cost and complexity to upgrade to newer, faster server and storage options were also painful.

These factors led to the emergence of highly parallel architectures and massively parallel processing technology known as MPP systems. The transition to MPP required a major investment of research and development by database vendors.

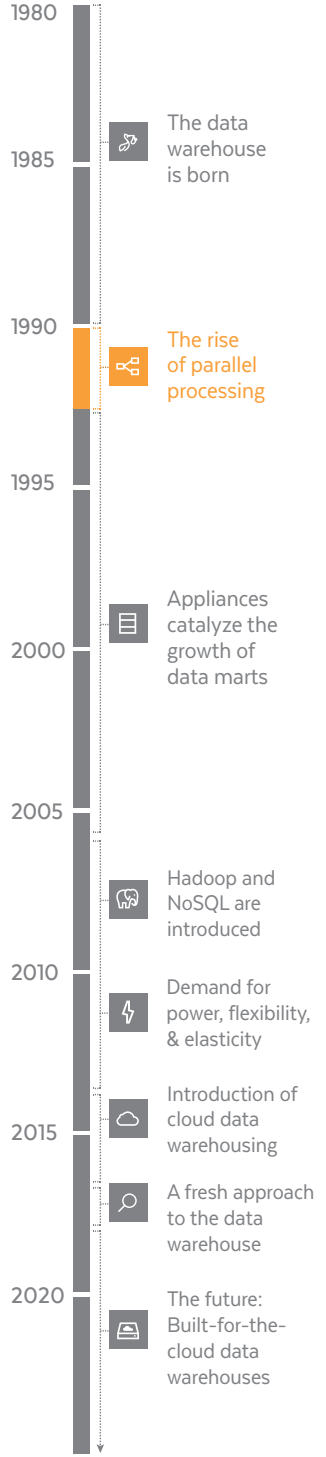
The first evolution of MPP was shared-disk: an MPP cluster of servers connected to a shared disk array. Shared-disk architectures retain simple storage management by centralizing data but with the tradeoff of a performance bottleneck between storage and compute. Next came the shared-nothing architecture: an MPP cluster without a central disk, with data distributed across the cluster. Shared-nothing architectures avoid the bottleneck between compute and storage but still carried the tradeoff of complicated storage management. For example, resizing the system requires redistributing and re-replicating data.

A MARGINAL STEP FORWARD

What were the benefits of this next iteration of the data warehouse? Enterprises could collect, integrate and query even larger data sets than ever before. But the complexity of configuring and administering these systems required exponentially more time than executing the actual analytic queries against the data warehouse. In addition, scaling for complex queries and more than just a few users could reduce performance to a crawl.

LESSON LEARNED

Parallelism is a critical tool to addressing scale, and scaling additional computing resources more effectively. But parallelism didn't completely solve the problem of getting data-driven insights to users, and quickly. In addition, the complexity of these architectures created new problems.



The data warehouse in a box

Data warehouse appliances attempt to democratize data access

As data warehousing grew in importance in the late 1990s and early 2000s, more and more parts of the business demanded access to data for reporting and analytics. These growing demands on data warehousing led to increasingly complex systems, and lots of time spent configuring, maintaining and deploying an overburdened data warehouse.

To try and relieve some of the load on the data warehouse, enterprises began setting up additional copies of data in the central data warehouse. This gave smaller groups inside the enterprise their own sandbox, or data mart: a subset of the data warehouse to serve a specific group or line of business (LOB).

APPLIANCES CATALYZE THE GROWTH OF DATA MARTS

Supporting this trend, IT vendors started offering preconfigured data warehouse appliances to simplify data mart set-up. Customers could get up and running, without the headache of purchasing and configuring individual servers and installing and configuring database software.

The data warehouse appliance enabled a significant expansion of data access across the enterprise. However, these devices also created a much more complicated data environment to manage: multiple data repositories, housing multiple copies and versions of data, spread across many different physical locations.

WHY DID ENTERPRISES EMBRACE THE DATA WAREHOUSE APPLIANCE?

- Reduced system administration.
- Most appliance vendors used MPP architectures for high performance.
- Users achieved faster query performance over non-appliance solutions.
- Faster path to go-live by eliminating complex, piecemeal processes to purchase, install and configure hardware and software.

WHAT DID ENTERPRISES SEE AS THE DRAWBACKS?

- Proprietary platforms made migration off the system difficult.
- No flexibility around configuring the ratio of storage vs compute.
- Scaling up the system required a “forklift” replacement and migration project.
- They often served as data marts, creating separate repositories of inconsistent data.

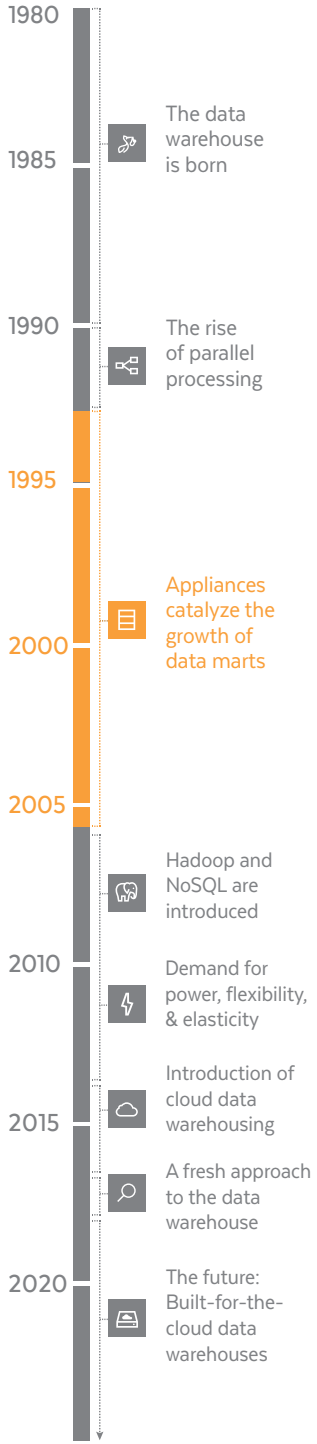
THE PITFALLS OF LOW STORAGE LIMITS

In addition to the inherent drawbacks of these appliances, the growth in data meant enterprises would often exceed the finite available storage of the appliance, which could be as low as 10TB of storage. When enterprises reached this limit, it was near impossible to simply increase the storage unless the organization undertook a complicated, disruptive migration to a larger hardware configuration.

This constraint, which plagues on-premises data warehouses of all types, remains a significant issue for enterprises today. Cloud, which we’ll discuss later, emerged as the first real opportunity to enable unlimited storage and scalability without the high cost and headache of scaling seen with traditional data warehouses.

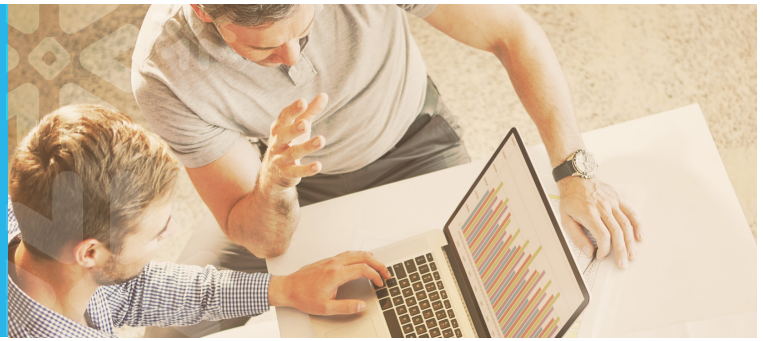
LESSON LEARNED

Making a data warehouse easier to get up and running, and use, is critical to supporting broader access to data insights. But that solution shouldn’t lead to fragmented data silos, generating significant data inconsistencies and requiring significant resources to manage.



Data warehousing and the big data explosion

In stomps the elephant: Hadoop and NoSQL



In the late 2000s, the cost of creating and storing data rapidly declined. As a result, new applications emerged, generating exponentially more data than ever seen previously. Social media applications attracted millions of users, capturing every click, view, share and post. Facebook has two billion users and generates nearly one petabyte of data each day. Other applications now generate huge amounts of log data or machine data.

NEW SEMI-STRUCTURED DATA TYPES

The data explosion spawned new structures for data beyond the traditional and rigid relational formats. These structures that emerged were flexible and could express more complex relationships, including semi-structured data such as JSON, Avro and XML. Since existing data warehouses weren't designed to handle these types of data, new NoSQL systems emerged such as Hadoop, an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. Originally, NoSQL meant "No SQL", which evolved to "not only SQL".

Eventually, people realized they needed to use SQL on those NoSQL systems to leverage the tools and skills they already had. That demand led to the creation of a plethora of SQL and SQL-like interfaces for NoSQL systems.

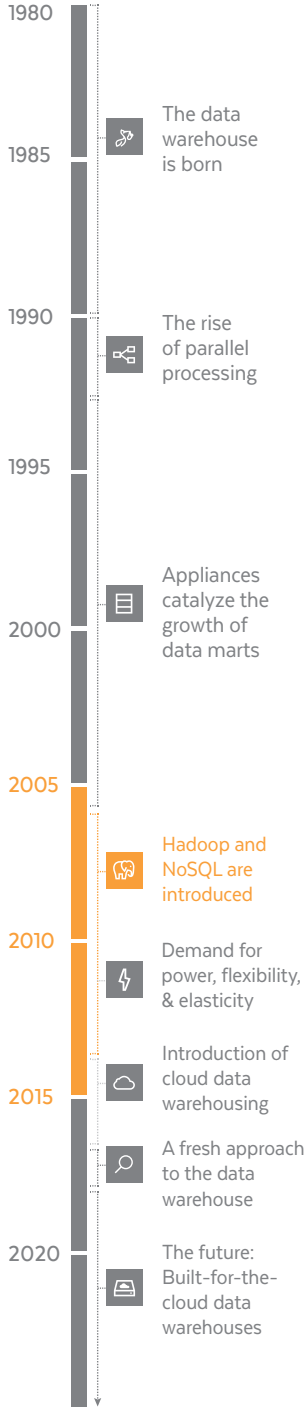
These developments led many to question whether the data warehouse would become irrelevant, replaced by NoSQL systems. However, SQL support and performance for these systems has largely remained inadequate, and the complexity of these systems is still high.

WHAT WAS HADOOP DESIGNED FOR?

- Collecting large amounts of varying data types in their raw forms and storing them in a single data storage repository often called a data lake.
- Easily scaling storage but at low cost.

WHAT WASN'T HADOOP DESIGNED FOR?

- Optimized analytics
- High concurrency
- Easy access to query the data using standard SQL



The upside? Hadoop revealed the traditional data warehouse had not evolved fast enough to address this new tsunami of data. Hadoop and other, NoSQL systems addressed the challenge of getting all types of relational and non-relational data into a single repository – something traditional data warehouses were never designed for.

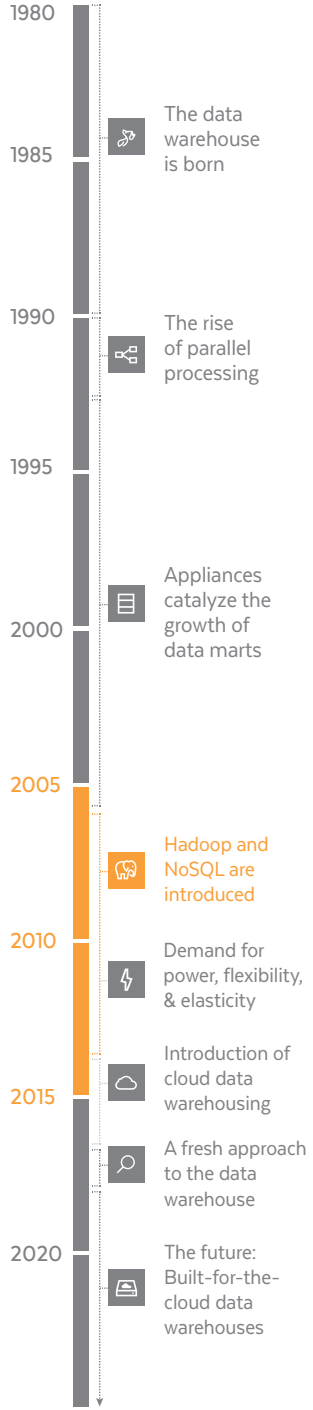
However, once IT stored the data in Hadoop, enterprises were left wondering how to transform this data into a format for fast, easy access by business people to perform deep but efficient analysis. Data warehousing was built upon this premise decades before, and the need still remained.

HADOOP DIDN'T REPLACE THE DATA WAREHOUSE

Ultimately it became clear that Hadoop couldn't replace the data warehouse. If the data warehouse had become obsolete, everyone would be writing Java programs in MapReduce or data frames in Spark. We learned that Hadoop is a toolkit, not an off-the-shelf solution designed for analytics. Data loading with Hadoop is fairly straightforward. But getting insight from data stored in Hadoop requires highly technical individuals with unique skills, which are in short supply.

THE FALLOUT? ENTERPRISES FOLLOWED ONE OF THREE APPROACHES:

- Keeping their traditional data warehouse
- Trying to work within the constraints of Hadoop
- Deploying both a data warehouse and Hadoop to advance closer to data nirvana: the data-driven enterprise.



The best, and worst, of both worlds reveal the future

Enterprises demand power, flexibility and elasticity

Data warehouse and NoSQL vendors continue to evolve their offerings, a testament to the varied methods today's enterprises demand to better serve their customers, streamline their organizations and lead their industries. Overall, these demands fall into the following categories:

POWER

Users want the benefits of what the traditional data warehouse embodied, and more. They want to perform today's analytics without the constraints of yesterday's offerings. They want the compute resources and concurrency to easily and affordably execute all sorts of analytics, including data exploration, BI and reporting, ad hoc analysis, event-driven analytics, predictive analytics and data-driven applications.

FLEXIBILITY

NoSQL systems created the ability to load all of an enterprise's data into a single repository. These data lakes largely overcame the challenge of housing the multitude of varying data types. Yet, these data sets remained in their original, raw formats. As discussed above, this created significant challenges to integrate and analyze that data.

Enterprises now want the best of both worlds: The flexibility of easily loading diverse data into a system that doesn't require transforming that data first. They also want a system that can handle changing data, while also having high-performance access to that data with the skills and tools they already have.

ELASTICITY

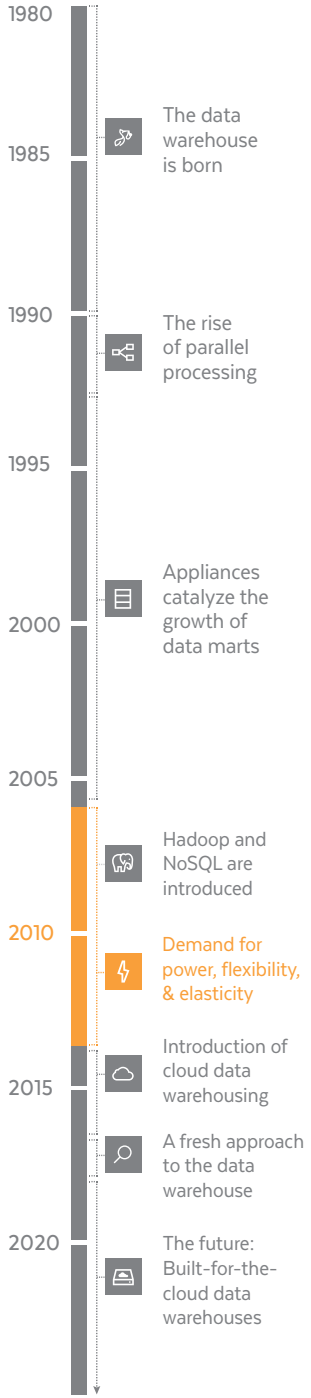
Many technology vendors have simply moved their on-premises data warehouse or NoSQL solutions to the cloud and offer it as a service. This shift changed nothing about the underlying architecture of these solutions. All of the on-premises constraints remained:

- Pre-defined and limited compute and storage resources
- Competition between user queries and data integration activities
- Compute and storage located on the same cluster, preventing you from scaling one without also scaling the other
- A host of other architectural impediments

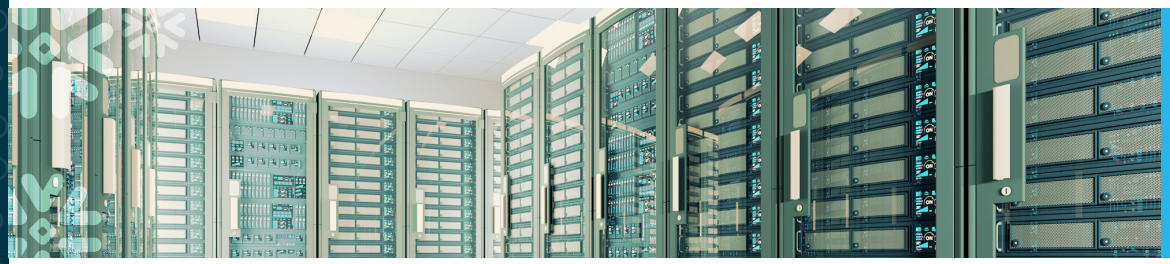
Enterprises still needed a better, affordable and elastic architecture to meet their ever-changing needs for data storage and compute resources. They also needed to optimize their costs and avoid pre-configuring their data warehouse for peak performance that may only last a month, week or day of the year. They needed to scale up, down and out (concurrency), automatically or with the click of a button, and pay only for the resources they used.

LESSON LEARNED

A modern data platform has to support traditional (structured) and non-relational (semi-structured), forms of data. It should also affordably meet the on-demand business needs for storage and compute. Finally, it should allow people to use tools and skills based on SQL, the dominate query language in enterprise environments.



The cloud revolution



Data warehousing and cloud computing

Today's big revolution is cloud computing. At first, cloud was used for low-priority workloads such as testing and development. Its initial value proposition was immediate access to storage and compute without purchasing hardware.

As enterprises considered the merits of cloud beyond easy access to cheaper resources, cloud software solutions emerged. However, these first-generation, managed services used the same technology and architecture enterprises used inside their on-premises data centers. The only major change was shifting the management of the technology from the enterprise to a cloud software provider.

WHAT WERE THE BENEFITS OF THESE EARLY CLOUD SOLUTIONS, INCLUDING EARLY CLOUD DATA WAREHOUSING SOLUTIONS?

- Minimized the huge, upfront purchase of a traditional, on-premises solution.
- Depending on the solution, some or all of hardware and software maintenance transfers to the data warehouse vendor.
- The time to implement could be much less without having to purchase and install an on-premises solution.

WHAT DO ENTERPRISES MISS OUT ON WITH A DATA WAREHOUSE NOT BUILT FOR THE CLOUD?

- Effortless and infinite scaling of cloud compute and storage resources.
- Independent scaling of compute and storage so you only pay for what you use.
- True elasticity to scale up, down and out, automatically or with a simple click of a button.
- Support for an infinite number of concurrent users without competing for resources.
- Loading and querying data simultaneously without degrading performance.
- Transferring the responsibilities of sizing, balancing and tuning the data warehouse, and enabling security, to the cloud vendor.

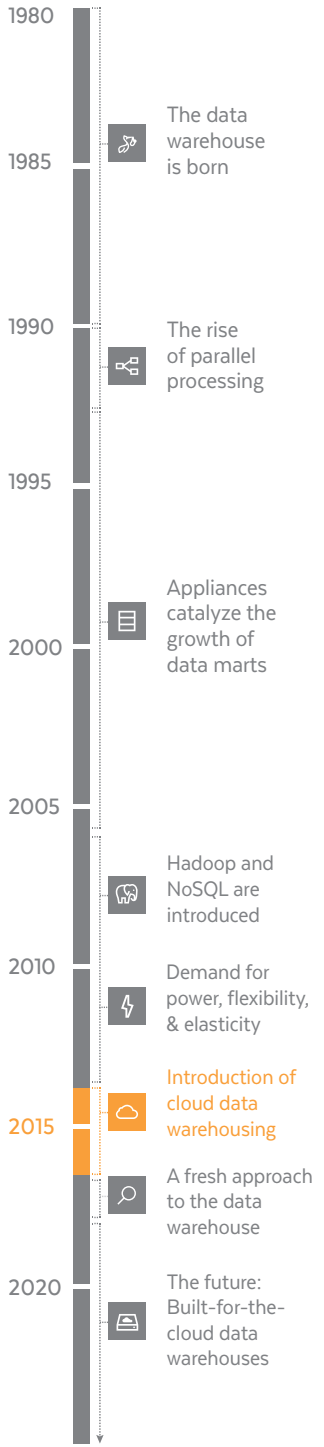
CLOUD: TODAY'S CHOICE FOR DATA ANALYTICS

As the cloud matured, it became clear it was a new development platform, one with unique capabilities. Fast forward to today: The cloud is now considered a viable option, and often the default option, for a wide array of workloads, including data analytics.

The cloud transformed the notion of what's possible when architecting and building a data warehouse. Traditional vendors understandably took the most direct path by simply moving their on-premises data warehouse solutions to the cloud. Their customers experienced a few benefits. However, these vendors had only scratched the surface of what's possible with a true, built-for-the-cloud data warehouse architecture.

LESSON LEARNED

Most data warehouse and NoSQL offerings are simply copied to the cloud, or, "cloud washed" versions of existing options. To take advantage of cloud, a solution must be built for the cloud, from the ground up.



The data warehouse brought full circle

Four major impacts converge

In recent years, four major impacts have converged to both necessitate and enable a fresh approach to data warehousing that truly enables the data-driven enterprise:

- **History:** Data warehousing and NoSQL technologies both paved the way for the modern data warehouse.
- **Data:** The amount of data generated in the last two years has eclipsed all other data created in previous times. And it's coming in all forms.
- **Demand:** Enterprises want the simplicity, concurrency and affordability to store, integrate and analyze all useful data by all business data users, around the clock.
- **Cloud:** The architecture to efficiently scale unlimited compute and storage, enabling data analytics for unlimited, data-driven insight.

KEY CONSIDERATIONS FOR CHOOSING A DATA WAREHOUSE THAT MEETS THE NEEDS OF TODAY'S ORGANIZATIONS

So what are the implications for today's data platform buyers? Data warehousing is more relevant than ever. SQL has proved to be a critical requirement for any data platform, and SQL remains at the core of data warehousing. History has shown that NoSQL options will not eliminate the data warehouse.

Existing cloud and on-premises architectures have reached their limits. Therefore, data warehousing must evolve to keep up with demand. A crucial part of that evolution is the need to develop a new architecture that can support unlimited growth in data volume, workload intensity and concurrency.

Another critical component is support for diverse data: structured and semi-structured data. Ease of use matters. Without it, all potential data users can't have access to effective data analytics. "Why not cloud?" has become the default question but answering it effectively requires a solution built for the cloud

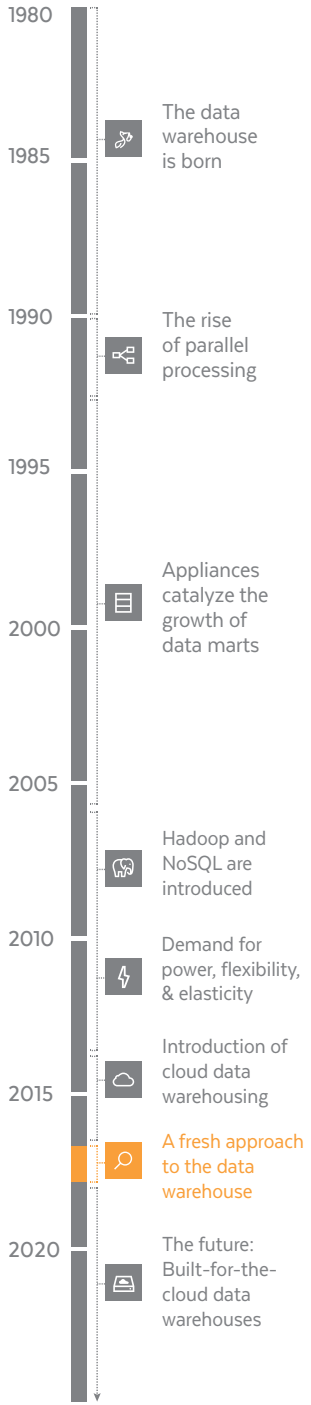
THE MODERN DATA WAREHOUSE: BUILT FOR THE CLOUD

The innovations of traditional data warehousing and subsequent NoSQL systems span four decades. As technology vendors iterated over and over, a number of themes emerged that were as important at the birth of the data warehouse as they are today:

- Enterprises still strive today for what they wanted 30 years ago: data-driven insight.
- Each solution paved some of the path to enabling deeper and wider data analytics.
- The demand for better technology has always exceeded current offerings.
- Enterprises want all their data in one location, with secure access by all users.
- Four decades of innovations have crystallized the architecture and technology needed to enable the data-driven enterprise.

LESSON LEARNED

History reveals the modern data warehouse must leverage the architecture and affordability of the cloud, the flexibility of NoSQL technology and the power of the traditional warehouse.



SIX QUALITIES THAT DEFINE A DATA WAREHOUSE BUILT FOR THE CLOUD:

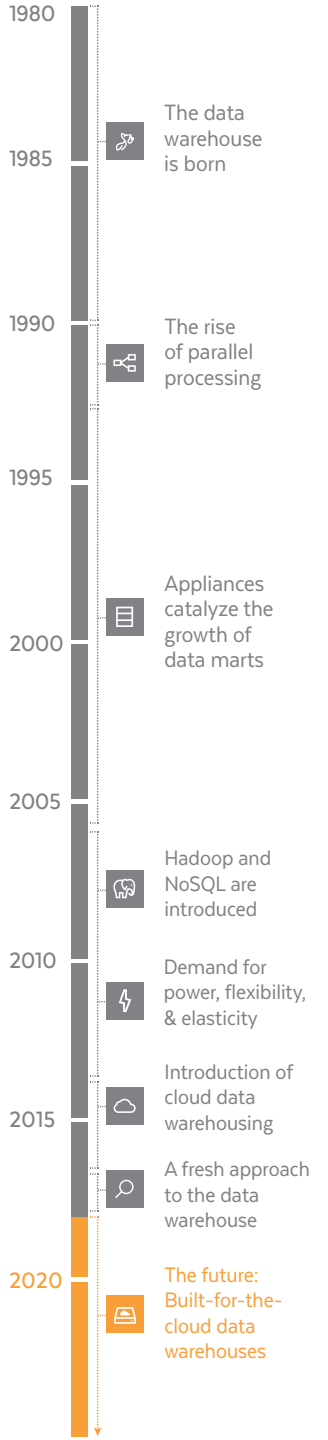
- **Complete SQL database:** Supports the tools millions of business users already know how to use today.
- **Zero management:** Reduces complexity with built-in performance so there's no infrastructure to tweak, no knobs to turn and no tuning required.
- **All your data:** Analyze multiple petabytes of structured and semi-structured data (JSON, XML, Avro) to quickly extract critical insight.
- **All your users:** An architecture that supports an unlimited number of concurrent users and applications to gain access to the data warehouse without eroding performance.
- **Pay only for what you use:** Scale storage separate from compute, up and down, transparently and automatically.
- **Data sharing:** Direct, governed and secure data sharing within seconds so enterprises can easily forge one-to-one, one-to-many and many-to-many relationships.

From the past, to the present, to your organization's future

The built-for-the-cloud data warehouse enables the data-driven enterprise

Architectures of solutions from the past can't be re-engineered to deliver all these capabilities. A single solution is possible only with the invention of a brand new data warehouse architecture built from the ground up for the cloud.

Armed with the knowledge of the evolution of the modern data warehouse, you and your organization can be champions for a data-driven enterprise. Give your organization the architecture and technology of a modern data warehouse - one built for the cloud.



Time to move forward

A path for modern data-driven organizations

BECOME A CHAMPION TODAY WITH MODERN CLOUD DATA WAREHOUSING

Data warehousing has been rethought and reborn in the cloud for the modern, data-driven organization. Find out how you can challenge the status quo and become an IT champion that creates the future, giving users the benefits they dream of with data warehousing built for the cloud.

Take the next step toward writing the next chapter of your organization's data warehousing history.

Visit www.snowflake.net