snowflake®

# 3 WAYS DATA SCIENCE IS SHAPING THE FUTURE OF DATA WAREHOUSING

**The benefits of a self-adapting data warehouse**

CHAMPION GUIDES

# TABLE OF CONTENTS

# HOW DO WE USE ALL THIS DATA?

**As businesses become more data-driven, data warehouses play an increasingly important role, helping businesses store, process, harness, and manage their data to make informed business decisions. The underlying goal for any business leveraging a data warehouse is to reduce the time to value for that company's data. In other words, the goal is to shorten the amount of time it takes for a business to derive value from its data.**

While that goal has remained steadfast over time, achieving it has become more difficult. Data warehousing has become more complicated due to the rise of new data streams and the exponential growth of the data within those streams. At present, only 1%-2% of a company's data is considered actionable, and so is stored long-term, according to IDC.[1] Clearly, the challenge of loading, storing, and managing data is only going to grow larger.

As data sets have grown in complexity, so has the technology required to support a fully integrated data warehouse. But as technology advances to meet new data demands, it also creates new areas of opportunity for business growth and operational efficiency. The ability to access more data and process it more quickly will enable businesses to generate deeper insights, faster than ever before.

Today, for example, some technologies already enable automatic query optimization, while machine learning algorithms help automate a variety of once-manual functions. Advances in technology are even starting to let data warehouses tune themselves. This capability is accelerating the speed at which data warehouses deliver value to businesses. At Snowflake, Inc., we see these benefits firsthand, as our customers bring more data of varying types into Snowflake, and they get increasing value from that data at a faster pace.

With the advance of machine learning and availability of near-infinite storage and computing power in the cloud, we're headed toward an exciting new era: **the age of the self-adapting data warehouse.**

In the pages that follow, we'll explain how data warehousing evolved, the business drivers fueling the move to a self-adapting data warehouse, and the benefits that await data warehouse architects and business leaders who make the shift from legacy platforms.

[1] IDC's First Global StorageSphere Forecast Sees Installed Base of Storage Capacity More Than Doubling by 2023: https://bit.ly/2USrXeA

# THE INCREASING COMPLEXITY OF DATA WAREHOUSING

**As companies track data of varying types from an increasing number of sources and channels, it's no surprise that managing and getting value from the data has become more complex.**

Legacy data warehouse platforms have become more complicated by necessity. The amount of data a business must process today would seem unfathomable to a database administrator only a decade ago. Where we once talked about megabytes, gigabytes, and terabytes of data, more often today we talk about hundreds of terabytes, if not petabytes, of data. In the past decade alone, new and constantly evolving data streams, ranging from mobile and clickstream data to machine logs and other system-generated data, have overwhelmed these legacy systems. This, in turn, has required data warehouse architects and database administrators to develop highly specialized, platform-specific skills to create workarounds, custom tools, and new capabilities to handle the increased load.

Despite the increased complexity, leading businesses recognize the importance of data in making decisions, and they are investing more in time, tools, and people to manage their data. Technology investments are expected to increase at more than half (57%) of organizations worldwide in the next 12 months, with cloud computing and business process management as two of the top three areas targeted for greater investment by tech CIOs.[2] Recent IDC research highlights the rapid acceleration of data-driven business priorities, predicting that:

- **Worldwide revenue for big data and business analytics solutions will reach $260 billion in 2022, representing an 11.9% compound annual growth rate (CAGR) from 2017-2022[3]**
- **The installed base of storage capacity worldwide will more than double over the 2018-2023 forecast period, growing to 11.7 zettabytes (ZB) in 2023[4]**

[1] IDG. CIO Tech Poll: Tech Priorities 2019. https://bit.ly/2GrxN3R
[2] Revenues for Big Data and Business Analytics Solutions Forecast to Reach $260 Billion in 2022, Led by the Banking and Manufacturing Industries, according to IDC. https://bit.ly/2Mlkt3z
[3] IDC's First Global StorageSphere Forecast Sees Installed Base of Storage Capacity More Than Doubling by 2023. https://bit.ly/2USrXeA

# TOWARD THE SELF-ADAPTING DATA WAREHOUSE

Fortunately, computing power and memory have advanced to the point that machines can process much larger and more complicated data sets. Data warehousing can greatly reduce the time to value for a company's data and make data management processes more efficient.

As a result, many businesses are shifting to cloud-based data warehousing as a service, in which businesses can access and query their data in near real time, without incurring the costs of housing, managing, and maintaining an on-premises solution in a data center. This shift from fixed capacity, on-premises, servers to elastic cloud-based data warehousing as a service also transforms workload tuning activities, one of the critical parts of a successful data warehouse, from a manual in-house activity to part of the service offering.

The second major development is the emergence of self-adapting data warehouses. With self-adapting data warehouses, tuning is not just part of the service offering, but can be increasingly automated with machine learning. The scale and volume of data isn't a limiting factor. Instead, it's an advantage, helping discern patterns more quickly and delivering the right data to the right person, at the right time.

In this way, data warehousing will resemble other smart applications, which incorporate data-driven, actionable insights into the user experience itself. As Snowflake Inc., VP of Engineering Sameet Agarwal said, "I don't want to run my own software. It's much better to treat my data warehouse like an iPhone. With my iPhone, I let Apple update my software on the back end. I go to bed and charge my phone, and 'magically' it's updated. That's a far better user experience." Likewise, data warehousing as a service has advanced to the point that upgrades and patches to the system are completely invisible to the end users.

More broadly, the move to a self-adapting data warehouse complements an industry-wide shift toward agile methodologies, in which businesses extract value faster through regular deployment of incremental improvements. In the world of software, agile methods help developers get to market with a minimum viable product (MVP) faster. In the world of data warehousing, self-adapting systems help businesses get insights faster, with less manual intervention. As the data volume grows, and more queries are executed, the self-adapting data warehouse will deliver exponentially better results based on the additional inputs it has to analyze.

As we enter the age of the self-adapting data warehouse, three key benefits await those businesses that make the shift.

# BENEFIT #1: ENJOY GREATER SPEED AND AGILITY

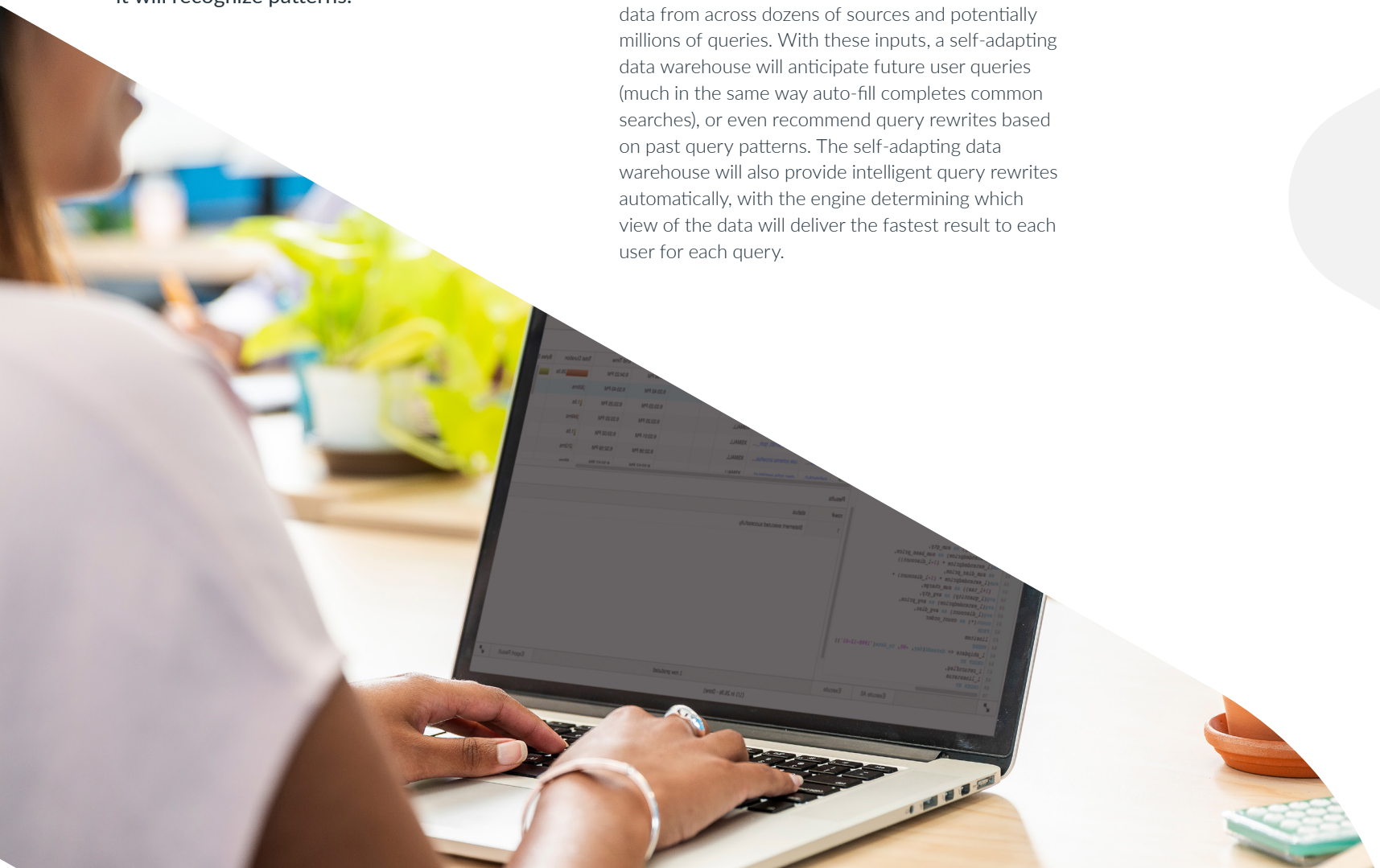
The self-adapting data warehouse will load, process, and present data in the best possible way for each user without human intervention, delivering more business value faster.

A self-adapting data warehouse will analyze larger data sets—data sets that were previously out of reach—more quickly, without the months of preparation and planning required for legacy on-premises platforms. For example, a major payments provider loading one trillion rows of semi-structured data per day wanted to analyze 15 years of web log data. In the past, this would have been impossible. With a self-adapting data warehouse, it's possible to load the data and begin analyzing it immediately, without the need to spend time tuning.

# BENEFIT #2: ANTICIPATE PATTERNS FASTER AND OPTIMIZE QUERIES

The more data a machine algorithm is exposed to, the faster, and more accurately, it will recognize patterns.

That said, it takes a lot of data to reach a statistically significant, accurate conclusion. Data warehouses have an advantage in this arena, since they see data from across dozens of sources and potentially millions of queries. With these inputs, a self-adapting data warehouse will anticipate future user queries (much in the same way auto-fill completes common searches), or even recommend query rewrites based on past query patterns. The self-adapting data warehouse will also provide intelligent query rewrites automatically, with the engine determining which view of the data will deliver the fastest result to each user for each query.



# BENEFIT #3: AUTOMATED DATA ORGANIZATION & OPTIMIZED WORKLOADS

As any database administrator can attest, determining how to index and organize the data for greatest congruence, performance, and accessibility is a challenge in legacy platforms. Getting performance and value out of a data warehouse historically took a lot of time and manual effort for planning and maintenance.

Self-adapting data warehouses will automate the organization of the data, eliminating the manual work required to design the optimal organization of the data. Eliminating these complex, error prone, and manual tasks creates an opportunity for database administrators and architects to focus on broader strategic data management initiatives.

A self-adapting data warehouse will also optimize workloads. In the past, when multiple users accessed a very large data set simultaneously, legacy data warehouses could not easily optimize all queries and response times. But a self-adapting data warehouse with an underlying machine learning algorithm can determine how best to organize compute resources for the various workloads. This includes automatically scaling compute to support increased concurrency when there is an unexpected surge in usage, maintaining a consistent level of service. A self-adapting data warehouse will organize data and the compute power automatically so that it provides the best performance for each user and maximizes the number, and type, of workloads that can execute concurrently.

# CONCLUSION

The rapidly growing scale and variety of enterprise data, once a concern among data warehouse experts, is quickly becoming a key asset for reducing time to value. By leveraging large-scale data sets and query history, the self-adapting data warehouse will recognize patterns faster, anticipating user needs and optimizing queries, views, and workloads.

The move to a self-adapting data warehouse represents the next phase in the evolution of data warehousing. But at its foundation, it's a return to the real value of any data warehouse–helping businesses derive more value from their data, faster.

## ABOUT THE AUTHOR

Kent Graziano is the Chief Technical Evangelist for Snowflake Inc. He is an award-winning author, speaker, and trainer in the areas of data modeling, data architecture, and data warehousing. He is an Oracle ACE Director - Alumni, member of the OakTable Network, a certified Data Vault Master and Data Vault 2.0 Practitioner (CDVP2), expert data modeler and solution architect with more than 30 years of experience.

## ABOUT SNOWFLAKE

Snowflake shatters barriers that prevent organizations from unleashing the true value from their data. Thousands of customers around the world mobilize their data in ways previously unimaginable with Snowflake's cloud data platform — a solution for data warehousing, data lakes, data engineering, data science, data application development, and data exchange. Snowflake provides the near-unlimited scale, concurrency, and performance our customers in a variety of industries want, while delivering a single data experience that spans multiple clouds and geographies. Our cloud data platform is also the engine that drives the Data Cloud — the global ecosystem where thousands of organizations have seamless and governed access to explore, share, and unlock the potential of data. Learn how you can mobilize your data at **snowflake.com.**