



# The Data Warehouse: The Engine that Drives Data Analytics

WHY THE CORE OF DATA ANALYTICS IS CRITICAL  
TO DATA-DRIVEN INSIGHT



CHAMPION  
GUIDES



# Contents

3	Why organizations are falling behind in the analytics revolution
5	The path to transformative analytics is filled with roadblocks
8	The business implications of analytics roadblocks
10	Why the data warehouse is still relevant
11	The cloud is the natural place for the modern data warehouse
12	Built-for-the-cloud data warehousing as a service
16	The business benefits of implementing a cloud data warehouse
17	Case study: From implementation to “in the black” in less than a year
18	About Snowflake

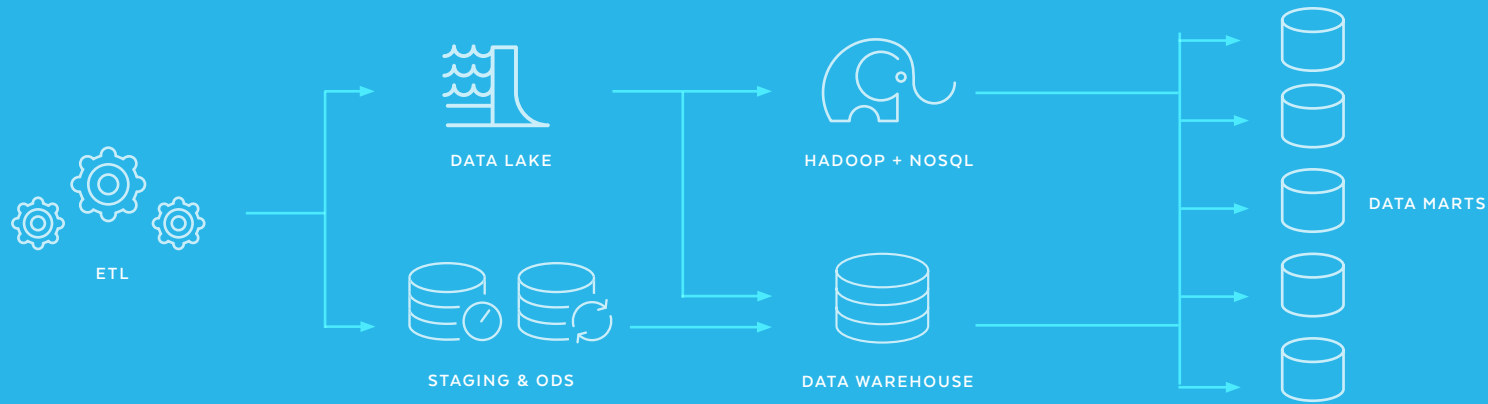
# Why Organizations Are Falling Behind in the Analytics Revolution

According to McKinsey, 86 percent of executives say their organizations have only been somewhat effective at achieving the primary objective of their data and analytics programs. Even more concerning, more than one-quarter say they've been ineffective.<sup>1</sup>

The fact that most data initiatives fail to meet the real needs of today's enterprises, or just outright fail to deliver any value at all, reveals two significant trends. First, enterprises are more focused on data analytics than ever before. Second, it's likely that most failures lie deep within an organization's data analytics platform.

Data analytics products continue to outpace the technology they rest upon. These vendors continue to deliver advanced but intuitive solutions that appeal to professionals of all levels, departments and business functions. That's the front end. On the back end, the same is happening with a number of nimble ETL (extract, transform and load) and ELT (extract, load and transform) vendors, which pull an enterprise's internal and external data into a central repository for data analytics solutions to do their work.

A TRADITIONAL CONFIGURATION FOR DATA ANALYTICS



What's left is where the problem lies. That central repository typically takes the form of a data warehouse built on legacy technology. Or, an enterprise has switched to a noSQL solution such as Hadoop. In more recent times, organizations have recognized the benefits, and many drawbacks, of both of these solutions and have responded with a hybrid alternative, using both systems in a lukewarm attempt to improve their data analytics. Enterprises are then forced to add another layer of complexity, data marts, which become a necessity for analytics with this model but are one of the big pains of data management.

Meanwhile, enterprises are collecting more data than ever before. And the volume, variety and velocity of data means that most of that data originates and is stored in the cloud. Petabytes of data are available, but that data goes stale in storage solutions outside traditional data warehouses because these legacy systems can't efficiently handle new, semi-structured types of data. Or, more

data is just too expensive to store in traditional systems that weren't built for the cloud. These systems, built when far less data was collected and far fewer users needed data access, still struggle under their legacy architecture and technology.

And even though Hadoop and other noSQL systems have made it possible to ingest varying data types, getting the best analytics from that data is nearly impossible. These systems have left enterprises without the right foundation to turn their growing hunger for data into real business insight.

In this ebook, we examine the challenges enterprises have faced turning data into transformative business insights. You'll also learn about the technologies that are now available to enterprises to keep up with the analytics revolution. With a modern cloud data warehouse, you and your organization can be champions.

# The Path to Transformative Analytics is Filled with Roadblocks

By now, your organization was supposed to be doing transformative things with analytics. But architectural or technological roadblocks created by legacy data warehouses and noSQL systems continue to inhibit the data analytics necessary to constantly improve the experience of data users, streamline operations, evolve the experience of your company's customers and allow your enterprise to become and remain a market leader.

These frustrations fall into three main categories:

- Getting access to data
- Processing analytics consistently fast
- Human resource limitations

## GETTING ACCESS TO DATA

Analysts and business users alike often wait hours, if not days, for data to load and queries to run. To alleviate the burdens these workloads have on operations, you must often wait until odd times of night to load data into your data warehouses. However, for today's round-the-clock global businesses, there are no off-peak opportunities. You're nobody's champion when gridlock has you spending most of your time directing traffic.

This lengthy process ensures that the data in your warehouse is always less than current. And newer forms of semi-structured data, such as JSON, are just too hard to load into a legacy data warehouse. So the data remains in silos outside the data warehouse. Essentially, the results from queries will be inaccurate, stale or inadequate because of the limited and out-of-date information in the warehouse.

## PROCESSING ANALYTICS CONSISTENTLY FAST

The data sets in your data warehouse have wide application across your enterprise. But running concurrent queries, data loads and development jobs will slow most traditional data warehouses to a snail's pace, or worse. In fact, you may even have database administrators killing queries they deem less important so more important jobs can run.

When this happens, your data users may develop workarounds that often make things worse. Here are just a couple of common options, and their consequences, that develop in attempts to combat latency:

- **Siloing Data**
  - Workaround:** IT may leave data in its original location, outside of the data warehouse, to avoid waiting for slow ETL processes.
  - Consequence:** Users are constrained by limited query results and the inability to perform analytics on all of an enterprise's data from one repository. Or worse, they try to integrate data manually, which can create inconsistent views. Unless they know the data well, this can result in incorrect results if the data in these silos is not joined properly.

- **Using Spreadmarts**

**Workaround:** Users extract large sets of data, maybe overnight and park that data in Excel spreadsheets to run their analytics offline.

**Consequence:** When this practice becomes widespread, different users will extract, and save, different versions of the same data, and then develop different analytic queries and reach different conclusions. In this scenario, it's also common for the users to integrate other, ungoverned data sources with the extracted data, skewing the results even further. In the end, none of the results are auditable.

Every minute your business intelligence and analytics team spends waiting for data is a minute they can't spend learning from that data. As long as enterprises rely on data systems that make accessing data a slow and cumbersome process, users will accomplish a lower volume of analytical work. They will continue to find workarounds, which lead to even more data siloing. Your enterprise will be stuck with an incomplete or inaccurate picture of their data, while analysts won't have the time they need to find new insights or ideas in the data they have.

#### HUMAN RESOURCE LIMITATIONS

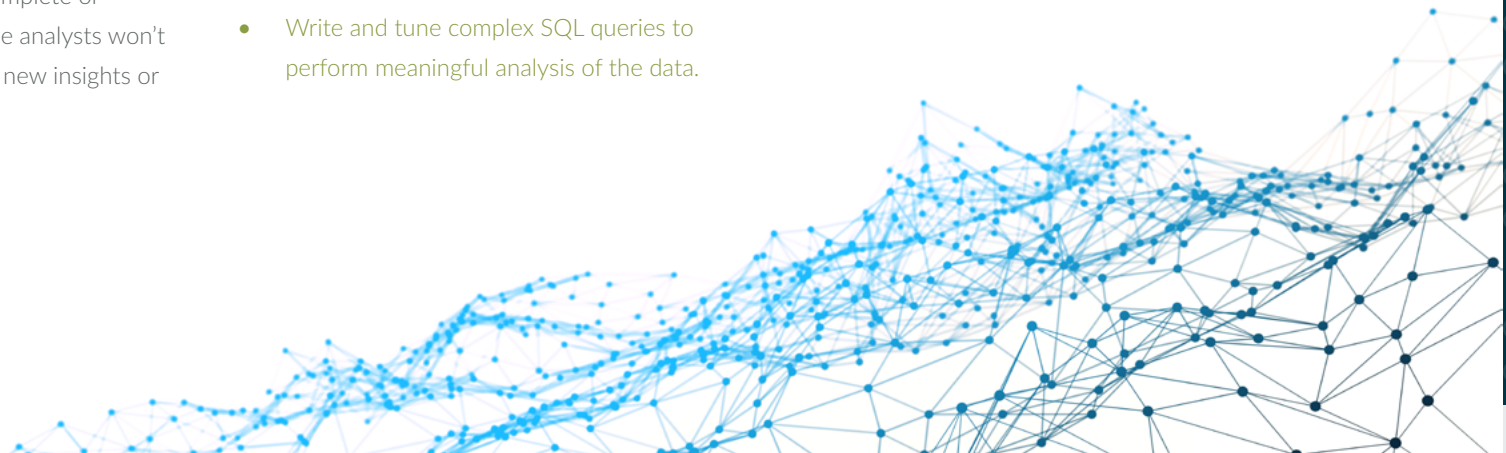
For all the talk of the “democratizing data,” enterprises still need technical staff to run their legacy, on-premises or cloud data warehouses. IT and data analytics professionals with all levels of skills remain the most sought-after human resource in an organization. Global IT staffing provider Modis estimates a 12 percent growth in IT jobs by 2024 – nearly double the projected growth of all other industries.<sup>2</sup>

With a traditional data warehouse, enterprises need staff to:

- Load and transform data into the right structure for a SQL database.
- Manage the database and the hardware required to maintain the database.
- Assist business users with their data “self-service.”
- Help business users develop and execute visualizations for more complex analysis.
- Write and tune complex SQL queries to perform meaningful analysis of the data.

The continued use of legacy data warehouse technology has given rise to an additional and more recent human resource problem. The rise of semi-structured data from sources such as mobile devices, social media platforms and machine data has revealed the limitations of the traditional data warehouse. In response, many enterprises have turned to noSQL systems such as Hadoop to build a “data lake.” However, querying these data lakes is vastly more complicated than running a SQL query in a relational data warehouse, and it requires different skills.

So, for Hadoop and other noSQL solutions, the human resource problem is even more acute. These systems have only appeared in the past 10 years. And they haven't magically produced the results that the initial hype predicted. Highly skilled professionals proficient with these solutions are integral to a Hadoop project becoming moderately successful. More importantly, IT professionals with these skills are hard to find due to the recency of these systems and their shortcomings.



Enterprises with noSQL data repositories need advanced data scientists who can write MapReduce queries in more complex languages, such as Java. IT professionals with these MapReduce skills are in far shorter supply than professionals with SQL experience. This has led to employees skilled in these areas being difficult to recruit and expensive to retain.

To add yet another challenge, both traditional data warehouse solutions and noSQL systems emerged before the cloud. On-premises or “cloudified” versions of these solutions will never reduce the cost, complexity and headache of efficiently getting all your data in and all your insight out. All the while, access to these systems will still be limited to the few technical team members. Those team members will only push answers, not access, to many more people across your enterprise who work with data and rely on data to perform their jobs.



# The Business Implications of Analytics Roadblocks

The technology constraints that legacy data warehouse and noSQL solutions present are no secret. The headache of constantly tweaking, tuning, fixing and maintaining a so-so system is a stark reminder of these constraints. And all your efforts are likely to provide only incremental improvements to maintain a system that reached its potential long ago. But what about the business impediments that this technology creates?

## HOW RESOURCE SCARCITY INHIBITS ANALYTICS

The challenges caused by ineffective data systems can be boiled down to two words: resource scarcity.

Your resource challenges inhibit your ability to dive deep into data exploration, business intelligence, reporting and predictive analytics. Without the time, money and human intelligence needed to devote toward generating analytics insights, you're unlikely to develop new ideas, products or processes from your data. Rather, you'll be stuck simply running basic reports, while your competitors transform their businesses with complex analytical processes such as machine learning, predictive analytics and data mining.

Resource scarcity caused by inefficient data systems prevents the process of innovating products and business practices with analytics. There is simply no way to conduct experiments on your data and use those experiments to generate new data. Instead, you're stuck at step one, letting ideas pile up and wishing you could do something about them.

## THE INABILITY TO EXPAND DATA ACCESS TO ALL USERS

The strains caused by exponential growth in data and the needs of a growing user base remain critical for many enterprises relying on a data warehouse as the foundation of their analytics.

Due to the intricacies and glacial speed of scaling up most data warehouse solutions, sales, finance and C-suite stakeholders, who are excited by the possibilities offered by deep analytics, will have to wait months or even years to begin to gain insight from all your available data.

These stakeholders will continue to rely on IT and skilled data analysts instead. The few who have the ability to access the data warehouse will provide answers to a small subset of stakeholders they can effectively support. Of course, this means that these few analysts will spend large parts of their day running basic queries to extract data and create visualizations to present that data, rather than performing more advanced data analyses.

### SPENDING MORE TIME ON INFRASTRUCTURE THAN DATA

Many organizations spend far more time and effort maintaining the infrastructure of their data and data systems than they do actually learning from the data itself.

These infrastructure needs involve maintaining servers, monitoring system performance and fixing bugs as they arise. Often, this involves making difficult decisions about which data is important to load and which queries can or cannot be stopped in the interest of system performance. Depending on the size and complexity of your platform for data analytics, some organizations may have several full-time employees whose sole job is to monitor and maintain data center infrastructures.

Depending on your internal protocols and current data warehouse architecture, you may also spend huge amounts of time extracting, transforming and loading your data into your data warehouse. For most organizations, ETL can involve multiple steps when it comes to cleansing and shaping the data to meet the system's needs.

All of this time spent on infrastructure, database management and ETL (the complexity of which is often a result of infrastructure) adds up. That means your data team is left playing traffic cop, while missing out on the chance to get their hands on significant analytics projects and strategies that help advance your enterprise and your career.



# Why the Data Warehouse is Still Relevant

Given all of the challenges posed by the data warehouse, is it still a relevant architecture for today's era of analytics?

The answer, surprisingly, is yes.

As was the case when the data warehouse emerged in the early '90s, enterprises need a centralized place to house and run queries on all of their data. And the rise of noSQL systems such as Hadoop have failed to replace the data warehouse as the dominant model.

Data lakes offered great potential because of their ability to easily load virtually any type of data. But unlike the organizational elegance of a data warehouse, running a query against a disorganized and ungoverned data lake is somewhat akin to trying to find a specific size, color and brand of shirt in a vast discount store amid thousands of unmarked bins.

From three decades and two vastly different approaches of enabling effective data analytics, the modern data warehouse has emerged. It represents the power of traditional data warehousing, the flexibility of big data platforms and the elasticity of the cloud at a fraction of the cost of previous solutions.

# The Cloud is the Natural Place for the Modern Data Warehouse

The benefits of modern cloud data warehousing continue to eclipse the initial concerns of moving an organization's data out of its on-premises data center.

This migration has definitely solved many of the core manageability problems posed by on-premises data warehouses systems. Now, your organization can rely on cloud-based providers to handle these needs via managed services. But not all cloud-based data warehouses offer the same architecture and technology to completely leverage the benefits of the cloud.

In fact, many cloud-based data warehouses still pose some of the same roadblocks to enabling innovative analytics that on-premises systems pose. They really are just "cloudified" solutions, or even simply hosted versions of on-premises data warehouses. These systems still require large teams to handle ETL and database management.

For example, these cloud-based systems offer the same traditional architecture of having data and compute on the same node or cluster. This means that scaling up to handle growing demands requires a cumbersome process to redistribute data before

the new resources can be used. The more popular option is to do nothing and keep playing traffic cop to provide meager performance for every user and job in the data warehouse. All the while, enterprises remain stuck waiting to realize their vision of an analytics-powered future.

While the cloud has addressed some of the challenges posed by on-premises systems, to truly gain from the analytics revolution, enterprises need a data warehouse designed specifically for the cloud – designed to overcome the challenges described above.

# Built-for-the-cloud Data Warehousing as a Service

With the right features, functionality and architecture, built-for-the-cloud data warehousing as a service offers great potential for enterprises looking to overcome the challenges and roadblocks they've faced due to other traditional data warehouse and open source systems.

Only the modern data warehouse built for the cloud can offer these defining qualities to revolutionize data analytics:

- Exponential improvements in performance
- Zero management
- Infinite scalability, up and down
- Unlimited concurrent users without an impact on performance
- Natively handled petabytes of structured and semi-structured data

Built-for-the-cloud data warehousing as a service provides the engine that enterprises need to perform deep analytics on all their data to evolve their business. The benefits of the modern data warehouse fall into a few core categories.

**ELASTICITY AND SCALABILITY**

To offer ideal scalability and elasticity without sacrificing performance, the built-for-the-cloud data warehouse as a service combines the best aspects of shared-disk architecture and shared-nothing architecture.

Shared-disk architectures simplify data management by centralizing data, but with the tradeoff of a performance bottleneck between storage and compute. On the other hand, shared-nothing architectures avoid the bottleneck between compute and storage, but with the tradeoff of complicated data management — resizing the system requires redistributing and re-replicating data.

The ideal data warehouse as a service advances on these two approaches with a brand-new architecture, separating storage resources from compute resources to preserve speed and performance. Each query that users execute, ETL routine they run, or development job launched, would use a separate compute cluster against a single copy of the data. Each query or job would run independently of the others and could scale up or down, on the fly. In other words, users would get the speed they need while always being sure that the data they're viewing is accurate and current. No silos.

**TRADITIONAL ARCHITECTURES**

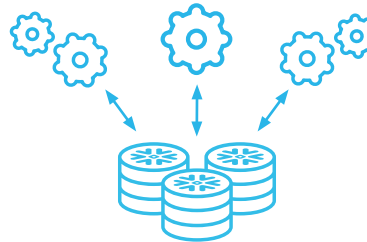


**SHARED DISK**  
SHARED STORAGE  
SINGLE CLUSTER



**SHARED NOTHING**  
DECENTRALIZED LOCAL STORAGE  
SINGLE CLUSTER

**BUILT-FOR-THE-CLOUD ARCHITECTURE**



**MULTI-CLUSTER, SHARED DATA**  
CENTRALIZED SCALE-OUT STORAGE  
MULTIPLE, INDEPENDENT COMPUTE CLUSTER

Additionally, because the solution is offered as a service, you only pay for the resources you use, when you use them. Enterprises that rely on a traditional on-premises or “cloudified” data warehouses must plan in advance for the maximum amount of storage and compute they may need on any day of the year.

The downside of the traditional architecture is that you must pay for all of the storage and compute resources that go unused for the other 364 days of the year. On the flip side, if your organization’s needs exceed your planned storage and compute resources, the time and money to scale up is prohibitive. And if you do scale in the traditional way, you’ve increased again the resources that go unused for the rest of the year.

A true built-for-the-cloud “as a service” approach enables enterprises to instantly scale up and down the amount of storage and compute they need from the warehouse. This means enterprises always have exactly the resources they need without wasting time and money better used to develop innovative analytics.



### DATA VARIETY

The built-for-the-cloud data warehouse can handle traditional structured data and make loading and querying semi-structured data quick and simple.

A modern built-for-the-cloud data warehouse should automatically identify key value tags and nested array structures within semi-structured (or flexible-schema) data, and make them visible to analysts via SQL. This should happen without relying on internal resources to design and develop rigid data structures upfront, and then use ETL tools to transform the data before loading it into the warehouse. This approach dramatically reduces the amount of time and resources enterprises would need to transform and load semi-structured data. It also makes the data immediately available to data analysts via SQL, and SQL-based analytics and BI tools — something that noSQL systems struggle with.

### SPEED

The built-for-the-cloud data warehouse automatically optimizes for speed.

For optimal querying speed, a data warehouse built for the cloud gives users the ability to automatically spin up and size an MPP cluster for any sized query, ETL job or development activity. When the job or query finishes, the modern data warehouse spins down or turns off the compute resources so your company doesn't let the meter keep running on unused resources. This also means the modern, built-for-the-cloud data warehouse can enable an unlimited number of concurrent queries, without affecting performance.

For each individual query or load, a modern data warehouse should also minimize or eliminate the need for mundane performance-tuning tasks, such as indexing and distribution keys, which are required with legacy technologies. It should also include an intelligent query optimizer that automatically gathers statistics on data characteristics as the data loads in real-time, and stores that information in a metadata engine. It should then use advanced optimizations to dynamically determine the most efficient path to the data. This process should consider the query and data structures and the MPP resources available at the time of query execution. This eliminates the time normally required for manual analysis and tuning, and also saves an enormous amount of space (e.g., for indexes, etc.), and therefore money.

### MANAGEABILITY

Data warehousing as a service eliminates the need to deploy a great deal of human and capital resources. This frees up time and talent to focus on generating insights from data analytics and business intelligence.

The built-for-the-cloud data warehouse as a service eliminates the need to purchase, run and maintain the many elements of a data center that traditional on-premises solutions require. But most cloud data warehouse vendors manage only some of the aspects of installing, managing and tuning their solutions. These alternatives typically take the form of IaaS (infrastructure as a service) or PaaS (platform as a service).

Only a true SaaS (software as a service) built-for-the-cloud data warehouse provides all hardware and software as part of its service, including all aspects of managing those resources. Typically included in the service are software and hardware upgrades, security, availability, data protection and performance optimization.

All of this adds up to more of your IT and data analytics people having more time to actually find, explore and experiment with data. And with this extra time, you'll gain the kinds of knowledge and insights necessary to transform your products, business practices and operations.

## SECURITY

Security remains the top concern for organizations migrating data to the cloud. Here are a few of the industry-standard measures a cloud data warehouse should offer:

### 1. Encrypting data in transit and at rest

If an unauthorized user gains access to your data they must not be able to read it. Period. The modern cloud data warehouse should 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.

In addition, a modern cloud data warehouse solution should use the latest industry-standard encryption algorithms. The Advanced Encryption Standard, AES, with 128-bit keys, is the minimum best practice for symmetric encryption. For enhanced security, the most robust cloud data warehouses use AES-256.

### 2. Key management

For any data warehouse, you should limit the amount of data covered by an individual encryption key and limit the time the key is used. This is an industry best practice delivered through *key rotation* and data *rekeying*.

- **Key rotation** is a method to periodically generate a new encryption key to protect newly inserted data.
- **Rekeying** is the ability to go back to previously stored data, re-encrypting it with freshly generated, new encryption keys and then disposing of the old encryption keys.

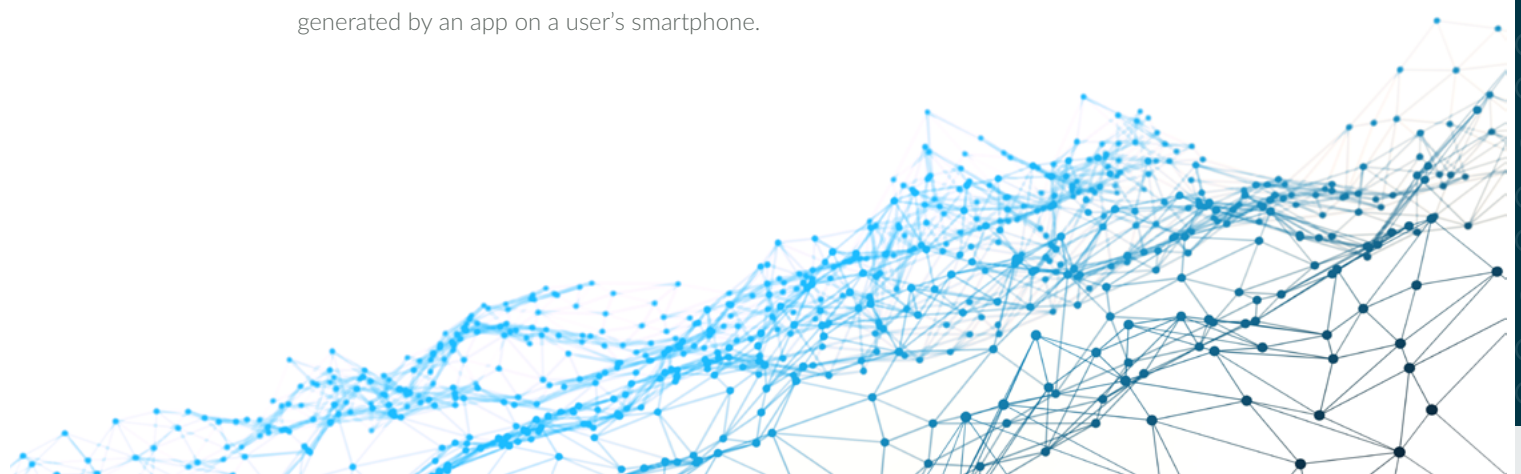
### 3. Multi-factor authentication

As a best practice, anyone accessing data in a cloud data warehouse should do so using multi-factor authentication (MFA). After logging in with a username and password, the user will need a second authentication mechanism. This can be a random code generated by an app on a user's smartphone.

### 4. Third-party verification

A cloud data warehouse provider should comply with the following standards, depending on the type of data you intend to store:

- **SOC 2:** The purpose of a SOC 2 report is to evaluate an organization's information systems relevant to security, availability, processing integrity, confidentiality or privacy.
- **HIPAA:** Protected Health Information (PHI) is subject to the privacy and security rules under the Health Insurance Portability and Accountability Act (HIPAA).
- **PCI:** Payment card industry (PCI) compliance is adherence to a set of specific security standards developed to protect credit card information during and after a financial transaction. All card brands require PCI compliance.



# The Business Benefits of Implementing a Cloud Data Warehouse

How far your enterprise is able to go with data analytics is largely determined by the engine powering those analytics.

As competitors find more effective ways to store data and ramp up their analytics, lagging organizations will fall behind, not just in terms of innovation, but in terms of revenue and profitability as well. The unfortunate but hard-and-fast rule is that data analysts spend 80 percent of their time retrieving and organizing data, and the other 20 percent actually performing data analytics. How much closer would your organization be to a data-driven business if that ratio was reversed?

If you're stuck running slow queries to generate basic reports, how are you ever going to take advantage of data mining, predictive analytics, machine learning or whatever the next horizon of analytics turns out to be? After all, the knowledge you gain from an uninhibited approach to analyzing your data may be the intelligence you need to transform your business.

"Insights often live at the boundaries," McKinsey says. "Just as considering soft data can reveal new insights, combining one's sources of information can make those insights sharper still."<sup>3</sup>

In the years ahead, the businesses that will thrive are those that adapt quickly to the changes around them. To see those changes clearly, you need a solid platform to drive your data analytics. So, become the champion to help transform your organization into the data-driven enterprise.

ENTERPRISES THAT DRIVE DECISIONS WITH DATA WILL SERVE THEIR CUSTOMERS, STREAMLINE THEIR OPERATIONS AND LEAD THEIR INDUSTRIES.

The built-for-the-cloud data warehouse as a service provides that foundation. By maximizing speed, minimizing costs and freeing up human resources, enterprises can gain the time and flexibility they need to more effectively capitalize on the value extracted from their data.

# Case Study: From Implementation to “In the Black” in Less than a Year

With the right data warehouse as a foundation, a future fueled by powerful analytics is easily imaginable. But perhaps a simple story of speed to value and ROI best illustrates the transformative value of the built-for-the-cloud data warehouse as a service.

PDX, a SaaS analytics provider to thousands of pharmacies, experienced many of the frustrations of its legacy data warehouse. After a proof of concept, it opted to implement a built-for-the-cloud data warehouse with all of the speed, scalability and manageability described above.

With a quick implementation, the company combined its new data warehouse with an analytics tool and began developing a data analytics service for its pharmacy clients. This service offered pharmacies insights about their customers not possible with PDX's previous data warehouse.

Within eight months, this service had grown so successful that PDX's SaaS solution was “in the black.” In other words, in less time than it takes to implement a legacy data warehouse system, this company had its built-for-the-cloud data warehouse up and running and leveraging a conglomerate of data sources from a single location, while making a profit.

**WANT MORE INFORMATIVE CONTENT TO HELP YOU TRANSFORM YOUR ORGANIZATION INTO A DATA-DRIVEN ENTERPRISE?**

[Visit Snowflake's Resource Library](#)

# 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](http://snowflake.net)

## CITATIONS

- <sup>1</sup> Brad Brown and Josh Gottlieb, "The Need to Lead in Data and Analytics," McKinsey, 2016. <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/the-need-to-lead-in-data-and-analytics> (retrieved 12/22/16)
- <sup>2</sup> "Forward Thinking: A Look Ahead at Tech Jobs," Modis, 2016. <http://www.modis.com/it-insights/infographics/top-it-jobs-of-2017/> (retrieved 1/17/2017)
- <sup>3</sup> Helen Mayhew, Tamim Saleh and Simon Williams, "Making data analytics work for you – instead of the other way around," McKinsey Quarterly, 2016. <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/making-data-analytics-work-for-you-instead-of-the-other-way-around> (retrieved 12/22/16)