



# THE DATA WAREHOUSE: THE ENGINE THAT DRIVES ANALYTICS

How to reinvent your analytics with data warehousing built for the cloud



CHAMPION  
GUIDES

EBOOK

# TABLE OF CONTENTS

- 3** Introduction
- 5** Best of all worlds: Unlimited, cost-effective scalability and performance
- 6** Up, down and out: How a cloud data warehouse should scale
- 7** User profiles: Solutions for common demands
- 8** The big question: What about security?
- 9** But wait, there's more: Additional security factor?
- 10** Security and compliance: Meeting stringent industry requirements
- 11** Tools to get your data into the cloud
- 12** Large-scale physical data transfers
- 13** TCO: Budgeting and cost management
- 14** ROI: It's TCO and much, much more
- 15** Your 5-step journey to data-driven champion

# INTRODUCTION

It's never easy being a data management professional. Business users from all levels and departments want useful insights from all your organization's data. If that's not enough, they want that same level of insight from data that has yet to arrive, or just won't fit, into your traditional enterprise data warehouse. All the while, they expect smartphone response times to both simple and colossal queries.

Bombarded with demands, you're stuck in the middle, juggling legacy technology located on-premises, in the cloud, or both. Too much of your time is spent fixing operational snafus that shouldn't happen. But what if your organization could easily and affordably implement a solution that gives your business intelligence (BI) and data analytics users what they want, when they want? And what if you could shift from managing day-to-day issues to championing truly strategic technology and data initiatives?

## THE FAST PATH TO CHANGE

This guide contains the information you'll need to understand and succeed with modern cloud data warehousing. It will start you on a path to transform your company's data analytics with a cheat sheet on five key topics crucial to getting started with cloud data warehousing:

1



PERFORMANCE  
AND SCALING

2



SECURITY IN THE  
CLOUD

3



MIGRATING DATA  
TO THE CLOUD

4

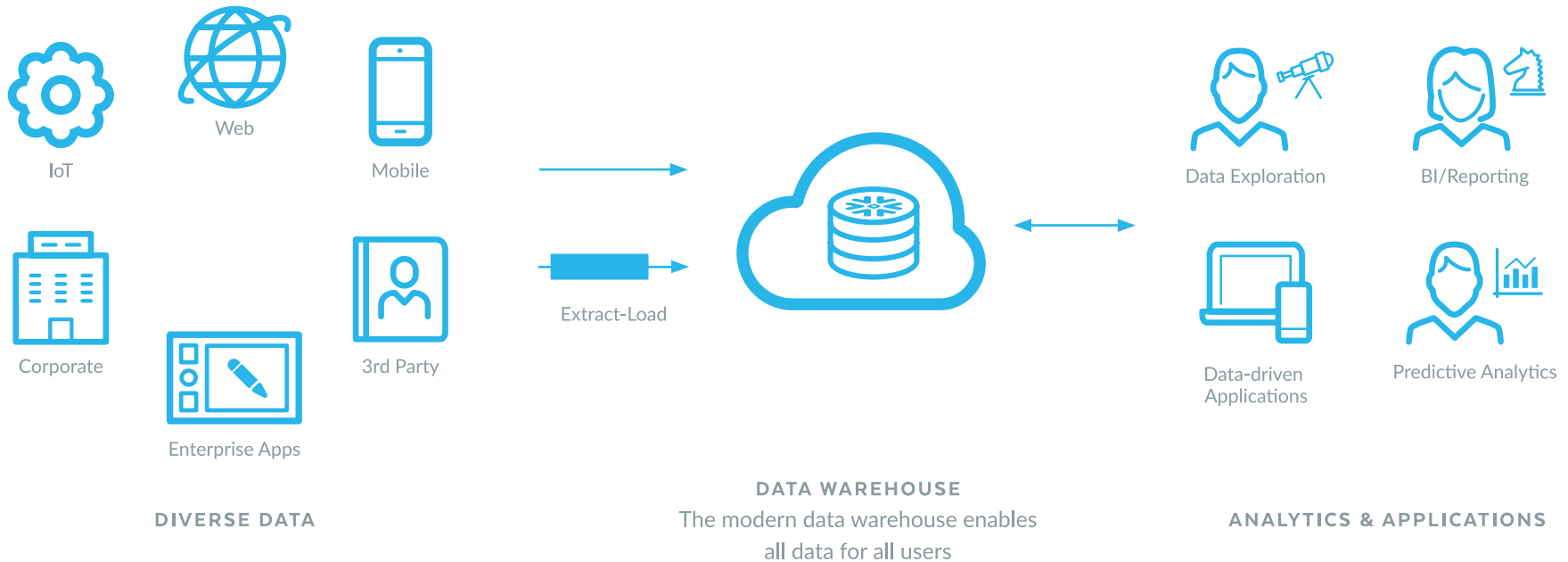


SAAS BUDGETING AND  
COST MANAGEMENT

5



EVALUATING RETURN  
ON INVESTMENT (ROI)



### GET YOUR DATA INTO A MODERN DATA WAREHOUSE BUILT FOR THE CLOUD

Free your data from legacy constraints with a modern cloud data architecture that delivers fundamental improvements in performance, concurrency and simplicity. Deliver all the insight from all your data with query times in seconds or minutes – not hours or days. What’s the result? You’ll have the time you need to work on the strategic stuff.

Believe the hype. In addition to many business and IT advantages, a cloud data warehouse that truly takes advantage of modern cloud architecture can increase performance by 200x for a tenth of the cost of legacy data warehouse systems located on-premises, or simply migrated to the cloud. Here’s what you need to know to get there.

# BEST OF ALL WORLDS: UNLIMITED, COST-EFFECTIVE SCALABILITY AND PERFORMANCE

Analyze vast amounts of varying data with speed

## TODAY, YOUR BUSINESS USERS WANT IT ALL:

- Easy access to very large and disparate data sets.
- Integration of flexible data types.
- The ability to drill down to all levels of data to get answers fast, no matter how involved the query.

User demands push legacy data warehouses beyond their limits. This means you're constantly juggling processes so the system doesn't crash. But too often, contention for limited resources forces you to reschedule jobs or kill a user's query entirely. When that happens, you're no one's champion.

## WELCOME TO LIMITLESS RESOURCE FLEXIBILITY

With the arrival of cloud architecture, you have a new way to think about scaling storage and compute: powerfully and cost-effectively. Cloud data warehousing can give you unlimited resources and the elasticity to access any scale of compute horsepower, 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. In addition, cloud storage can cost a fraction of the storage devices that currently live in your data center. But beware. Only a modern data warehouse built for the cloud, one that truly separates compute from storage, can effectively capitalize on everything cloud architecture can enable.

Here are some of the many issues a modern cloud data warehouse can alleviate:

- Competition between users' queries and data integration activities that degrade performance.
- Users forced to only analyze data subsets or aggregates to avoid further strain on the system.
- The complexity of data movement, refinement and transformation in legacy environments.

## TRADITIONAL VS. CLOUD-BASED ARCHITECTURE

### 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

# UP, DOWN AND OUT: HOW A CLOUD DATA WAREHOUSE SHOULD SCALE

Multiple ways to achieve 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 only pay only for what you use:

## 1 SPIN UP A NEW COMPUTE CLUSTER

Let's say you've got a workload that normally accesses the same amount of data all the time. You've chosen the right size cluster and it runs those queries in a timely manner. But what happens when users bump their queries from last month's data to include five years of data? You'd be better served with a new and bigger cluster.

However, adding more compute nodes to the existing cluster may not be the best solution. The good news is, with the right kind of cloud data warehouse, you can quickly define and spin up a new cluster in just a few minutes, flat.

## 2 KEEP A PRE-DEFINED COMPUTE CLUSTER IN SUSPENDED MODE, READY TO GO

If you have a regular event that requires a burst of compute resources, legacy architectures pressure you to build or add a new cluster from scratch every time you need it. This consumes time and money. On the flip side, the right cloud data warehouse can provide a pre-defined compute resource in suspended mode to switch on whenever you want. When users are done with the resource, you can park it in "sleep" mode until it's needed again.

Better still, you can configure the resource to go to sleep automatically after a predetermined time of no activity. Once the cluster turns off, the meter stops too, preventing you from paying for an unused resource.

## 3 DYNAMICALLY SPIN UP AN ADDITIONAL CLUSTER TO HANDLE CONCURRENT USERS

The number of concurrent users and queries on a cluster can surge, developing a queue. A true cloud data warehouse can automatically scale concurrency by transparently creating a new cluster that load balances with the first.

The mirrored cluster has access to the same data as the original. When the load subsides and the queries catch up, the second cluster will automatically spin down. Ideally, you can determine this concurrency-building capability and cost by specifying the maximum number of clusters that can be provisioned automatically.

## 4 MANUALLY RESIZE AN EXISTING COMPUTE CLUSTER

If you want to keep tight control over compute resources and costs, manual resizing provides an alternative to auto-scaling.

Let's say you have an existing cluster spun up with four nodes. But you know there's a surge in data coming and you want to give that cluster more compute horsepower, but for a specific period of time. You can resize that existing cluster, specifying eight, 16 or 32 nodes, while it's still running. When the surge subsides, scale the cluster back to its original four-node configuration.

These are just some of the possibilities available with a data warehouse that truly takes advantage of cloud architecture. Its exponentially better performance, compared to on-premises and "cloudified" data warehouses, is ready to meet insatiable user demands for data volumes, velocity and variety unimagined just 10 years ago.

# USER PROFILES: SOLUTIONS FOR COMMON DEMAND

Warehousing strategies to get you out of reactive mode

USER



CFO

The CFO who needs an answer right now.

Spin up a new, ad hoc compute cluster fast so a finance analyst can run queries within minutes and without impacting other users. Your results come from the company's single source of data so you're confident it's correct.



SALES

The sales team that makes a big push at the end of every quarter.

Keep a pre-defined cluster for customer and sales data in suspended mode, ready to go. Activate this pre-defined resource at a moment's notice. When the end-of-quarter rush is over, easily put it back into sleep mode.



MKTG

The curious marketer looking for the most profitable customer journey.

When users start adding multiple dimensions to their queries and analysis, complexity escalates quickly. Having a separate cluster for the marketer will ensure that a complex, ad hoc query doesn't impact other users or workloads.



DATASCI.

The data scientist who suddenly wants to stress test her brilliant theory.

If the amount of data being analyzed jumps from one year to three, you can manually re-size an existing cluster to deliver the necessary compute power. When the data scientist is satisfied with her analysis, you can scale the cluster back to its original size to keep costs low until the next big test.



SUPPORT

The support team that wants to find out who has the best-case resolution scores.

If a manager typically looks at a particular data set, but a large group of team members suddenly hits the same data set with additional queries, you'll need to act fast – so fast that human intervention isn't even a possibility. But don't worry. Automatic concurrency scaling can spin up a second cluster, accessing the same data, and then load-balance the queries with the first cluster. When the flurry of queries subsides, the second cluster will automatically spin down.



ANALYST

The supply chain analyst diving deep into inventory turnover trends.

This analyst often works with a rolling six months of data. Suddenly he is running queries on data going back two years. When you get complaints that queries are super-slow, you can spin up a new, properly sized cluster in just a few minutes, or simply resize the existing cluster to meet the increased demand.

SOLUTION

# THE BIG QUESTION: WHAT ABOUT SECURITY?

The cloud can be more secure than on-premises solutions

For years, organizations have considered the cloud to be even more vulnerable than on-premises solutions even though hackers continue to breach major corporate data centers. Certainly, security remains the top concern for organizations migrating sensitive data to the cloud. If you carefully evaluate and choose cloud offerings that make security a priority, you can benefit from some of the best security defenses in the industry – better than what most organizations have in place to protect their legacy systems.

Getting cloud security right is critical. Here are a few of the top measures a cloud data warehouse should offer to allay stakeholders' security fears:

## 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. Whatever resources you dedicate to security for your on-premises data warehouse, a truly secure cloud data warehouse will enable you to redirect those resources to other strategic IT efforts.

In addition, an advanced 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.

Unlike legacy security architectures, encryption shouldn't impact query or load performance of a cloud data warehouse. However, it's rare for enterprise data centers to be encrypted to this high degree because of the cost, time and scarcity of expertise.

## SECURE BY DESIGN



AUTHENTICATION



DATA ENCRYPTION



ACCESS CONTROL



EXTERNAL VALIDATION

# BUT WAIT, THERE'S MORE: ADDITIONAL SECURITY FACTORS

Exploring encryption key management and multi-factor authentication

There are many additional details to consider besides the encryption cipher. One of the most important is key management.

## KEY MANAGEMENT

Key management governs the lifecycle of the encryption keys, which includes the generation, storage, distribution, use and disposal of the keys. Ideally, a key hierarchy is utilized such that the root keys encrypt secondary keys – about one per data partition. These, in turn, encrypt even more granular keys such as one per table.

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.**

Both mechanisms are necessary to manage the complete lifecycle of encryption keys per the highest industry standards. A cloud data warehouse that provides this functionality nearly eliminates the responsibility of customers implementing encryption configuration and management with a legacy system. Even if key management wasn't available with your legacy data warehouse, consider it mandatory to operating securely in the cloud.

## MULTIPLE SECURITY FACTORS

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. Together, these factors prove that users are who they say they are – a strong measure to ensure only authorized parties gain access to data in the cloud.

# SECURITY AND COMPLIANCE: MEETING STRINGENT INDUSTRY REQUIREMENTS

## The importance of third-party verification

Industry-specific standards provide an additional layer of assurance for data security concerns. If you already ensure your legacy data warehouse meets security compliance, or it's something you weren't previously accustomed to, a cloud data warehouse provider should comply with the following standards:

### 1 SOC 2:

The American Institute of CPAs (AICPA) has developed the SOC 2 report. The purpose of a SOC 2 report is to evaluate an organization's information systems relevant to security, availability, processing integrity, confidentiality or privacy.

### 2 HIPAA:

Protected Health Information (PHI) is subject to the privacy and security rules under the Health Insurance Portability and Accountability Act (HIPAA). Cloud service providers storing PHI must adhere to HIPAA regulations for:

- Security and privacy
- User access tools
- Encryption
- Data location
- Return of data
- Contingency planning and disaster recovery
- Service Level Agreements (SLAs)

For most cloud infrastructure providers, such as Amazon Web Services (AWS), properly managing PHI data requires additional security controls detailed in AWS's business associate agreement (BAA). These controls often extend critical protections, such as encrypting data at rest, to exceed HIPAA specifications because cloud service providers want to reduce their risk. Healthcare organizations that store their data in modern warehouses built for the cloud reap the benefits of these additional protections.

### 3 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. For card information stored in a cloud warehouse, the vendor must:

- Build and maintain a secure network
- Protect cardholder data
- Maintain a vulnerability management program
- Implement strong access control measures
- Regularly monitor and test networks
- Maintain an information security policy

## THE IMPORTANCE OF THIRD-PARTY VERIFICATION

Some regulations allow providers to "self-attest" as proof of compliance. To ensure your data has the highest security, expect a cloud data warehouse vendor to use an independent provider to conduct penetration tests. A "pen test" is an attempt to evaluate IT infrastructure security by safely trying to exploit vulnerabilities that can exist in operating systems, service and application flaws, improper configurations, or risky end-user behavior. This type of security validation provides further assurance the security of a cloud data warehouse vendor meets industry standards and your organization's expectations.

# TOOLS TO GET YOUR DATA INTO THE CLOUD ...

... from wherever it resides now

Getting data into any data warehouse has historically been a slow and tedious process. That means business analysts may not always have the most current data. In addition, your enterprise, like many others, may be moving major applications to the cloud. Now you have to contend with accessing that data in the cloud, too.

But there is good news. Many new tools now exist to help you migrate existing, on-premises application data into a cloud data warehouse. Popular choices include graphical UI pipeline tools that allow you to extract, transform and load (ETL) data from both cloud services and traditional on-premises database sources into your modern cloud warehouse.

## THE BIG MOVE: BULK-LOADING FILES

A bulk-load approach can work best for your initial transfer if you have many terabytes of data that live on storage devices within your enterprise. The data is unloaded into thousands of manageable files and bulk loaded in parallel into the cloud data warehouse. Easily-deployed, supplemental resources to speed up loading should be integral to a modern cloud data warehouse.

## INCREMENTAL UPDATES

Once you've brought your data over, you'll want to capture incremental changes going forward. Expect to use one of the many modern ETL or ELT (extract, load, transform) tools for data transfer, available in commercial software and open source options. Additional options include the newer change data capture (CDC) and advanced data replication tools. Depending on the nature and source of the data, newer, cloud-based data streaming tools and platforms can also be useful. These tools are often well suited for incremental migrations rather than bulk loads.

# LARGE-SCALE PHYSICAL DATA TRANSFERS

When your data has snowballed into something really big

If you have hundreds of terabytes or even petabytes of data, it's sensible to use physical devices from cloud providers to migrate your data. These ruggedized and secure appliances are about the size of an old, school milk crate. You can load large amounts of encrypted data from your on-premises data center and ship the appliance back to the cloud services provider, which then uploads the encrypted data to a staging area in the cloud.

This method allows the transfer of about a petabyte of data per week to the cloud. You can then easily move the data over to your cloud data warehouse provider.

Data transfer appliances are often faster and more cost effective than trying to push all that data to the cloud via the Internet. They're a logical choice if all of your data would take longer than a week to upload to a cloud data warehouse.

## THE EXABYTE OPTION

Companies today generate so much data it's often measured in exabytes. (A single exabyte is equal to one billion gigabytes.) Some cloud service providers now offer a "drive-up" option for this level of data transfer. Literally speaking, it's a device housed in a large shipping-like container attached to a semi-truck. With ten of these trucks, you can have an exabyte of data moved to the cloud in six months.



# TCO: BUDGETING AND COST MANAGEMENT

## Planning a move to the cloud business model

Sure, moving your data warehouse to the cloud will save you the significant expense of buying, maintaining and securing an on-premises system. But how do you budget for an entirely different cost model?

### SAAS IS NOT JUST A TECHNOLOGY SHIFT

Switching from on-premises to a cloud solution also changes how an organization records and budgets for the purchase. An on-premises solution is often a large, up-front capital expense that's treated as an asset and depreciated over time. A cloud solution is a much smaller, up-front transaction and treated as an operating expense – deducted every month from an organization's corresponding revenues. And after the initial purchase of either solution, they both require additional but vastly different expenses to continue operating that solution.

Why should an IT person care? IT managers must show that a new technology meets an organization's technology needs and its budget. If the typical process to account for that purchase changes, then IT must present that new model. What does this all mean? At the heart of any solution purchase is a total cost of ownership (TCO) – a financial estimate that details the costs of a system over its expected lifetime.

### TCO: CLOUD VS. ON-PREMISES

One thing is for sure. With a TCO estimate, the first-year expense for a cloud-based solution should be a fraction of the cost of buying for on-premises. But when estimating for five, 10 or more years of ownership, a debate often ensues, prompting, "Will the cost of the cloud solution eventually exceed an on-premises alternative?"

With larger or enterprise-sized solutions, it's unlikely the cost of cloud and on-premises will ever converge. The initial and ongoing expenses of an on-premises solution will continue to cost more than the aggregate usage charges of a cloud-based solution.

The reality is, after the huge, upfront costs of an on-premises purchase, an organization must factor in the ongoing maintenance, repair and eventual replacement of its IT infrastructure. This includes building and maintaining one or more data centers, which can cost many millions of dollars.

In addition, there are annual maintenance and support fees for on-premises systems that can exceed 20% of the initial purchase price of the appliance and/or data warehouse software. There may be fees for financing the purchase, as well. Finally, the salaries of the highly specialized people who support a data center and all related functions can easily exceed 50% of the TCO. This is a substantial cost component even when comparing it to the large, up-front purchase of an on-premises solution.

In short, a simple but exhaustive TCO exercise for a cloud-based data warehouse will reveal not just the technological advances, but also the monetary value of switching to a cloud model.

# ROI: IT'S TCO AND MUCH, MUCH MORE

## Evaluating the ROI for your cloud data warehouse

TCO and ROI (return on investment) are the two sides of the same coin – they are inextricably linked. A TCO estimate reveals how much a cloud data warehouse will cost, and how much it will save an organization compared to an on-premises alternative. But those savings represent only one aspect of ROI. There are many other factors to consider when thinking about the ROI differences between on-premises and cloud.

Your ROI estimate should additionally include:

### 1 DEPLOYMENT TIMEST

A cloud-based data warehouse can go live in weeks or just a few months, depending on the size of the project and the migration strategy to the cloud. An organization can see benefits much sooner with a cloud solution and with a lower upfront investment. But beware, fast deployment estimates pertain only to a true SaaS (Software-as-a-Service) data warehouse. Solutions based on IaaS (Infrastructure-as-a-Service) or PaaS (Platform-as-a-Service) approaches can take much longer. Time saved directly translates into money saved.

### 2 SOFTWARE UPGRADEST

On-premises and “cloudified” on-premises solutions take a standard “waterfall” development approach to functionality updates. To enable the annual or biannual update, IT must usually take the system down or place it in maintenance mode – more lost time and money. To avoid this, IT may anchor to a specific version of the software, which creates a whole other set of headaches

With a modern cloud data warehouse, the upgrades should originate from an agile DevOps approach – incremental updates every month that avoid any disruption to customers.

### 3 THE PEOPLE FACTOR

As previously described, the number of people who maintain an on-premises, enterprise data warehouse and supporting infrastructure can be an enormous expense. Depending on the cloud alternative, an organization can significantly reduce or nearly eliminate this expense depending on the level of functionality, automation and vendor management of the solution.


### 4 PAY ONLY FOR WHAT YOU USE

An on-premises data warehouse forces you to buy enough storage space and compute horsepower to handle demand on your busiest day of the year. What about the other 364 days? With the right cloud data warehouse, there's tremendous opportunity to pay only for what you use, when you use it. Additionally, the cost of the actual storage and compute resources should be significantly lower with a cloud solution thanks to cloud economies of scale.

These factors and more should be part of your ROI calculations when comparing your on-premises option to a cloud data warehouse. If you consider more than one cloud data warehouse vendor for your next purchase, don't forget to evaluate the differences among those alternatives since no two data warehouses are the same – on-premises or cloud.

# YOUR 5-STEP JOURNEY TO DATA-DRIVEN CHAMPION

A cloud data warehouse cheat sheet

1	2	3	4	5
<b>SCALABILITY &amp; PERFORMANCE</b> 	<b>SECURITY &amp; COMPLIANCE</b> 	<b>DATA TRANSFER</b> 	<b>TCO</b> 	<b>ROI</b> 
<ul style="list-style-type: none"><li>➤ Compute separate from storage</li><li>➤ Scaling up, down and out on the fly</li><li>➤ Analyze large, disparate data sets</li><li>➤ Unlimited concurrency</li></ul>	<ul style="list-style-type: none"><li>➤ Encrypting data in transit and at rest</li><li>➤ Industry-standard encryption algorithms</li><li>➤ No impact to query or load performance</li><li>➤ Key management: rotation, rekeying</li><li>➤ Third-party compliance verification</li></ul>	<ul style="list-style-type: none"><li>➤ Existing migration tools available</li><li>➤ Bulk loading</li><li>➤ Incremental updates</li><li>➤ Large-scale physical transfers</li><li>➤ The exabyte option</li></ul>	<ul style="list-style-type: none"><li>➤ Moving to a cloud model</li><li>➤ From CAPEX to OPEX</li><li>➤ More than a technology shift</li><li>➤ Cloud vs on-premises</li></ul>	<ul style="list-style-type: none"><li>➤ Time to deploy</li><li>➤ Software upgrades</li><li>➤ The people factor</li><li>➤ Pay for what you use</li></ul>



# FIND OUT MORE

BECOME A CHAMPION TODAY WITH MODERN CLOUD DATA WAREHOUSING

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

Stop juggling problems. Take the first step of taking control of your data analytics operation. Visit [snowflake.com](https://www.snowflake.com)

Want more informative content to help you transform your organization into a data-driven enterprise? Visit [Snowflake's Resource Library](#)



© 2019 Snowflake. All rights reserved.

---