MIGRATE TO THE CLOUD

The how and why of modernizing your data platform

snowflake

MIGRATION GUIDE

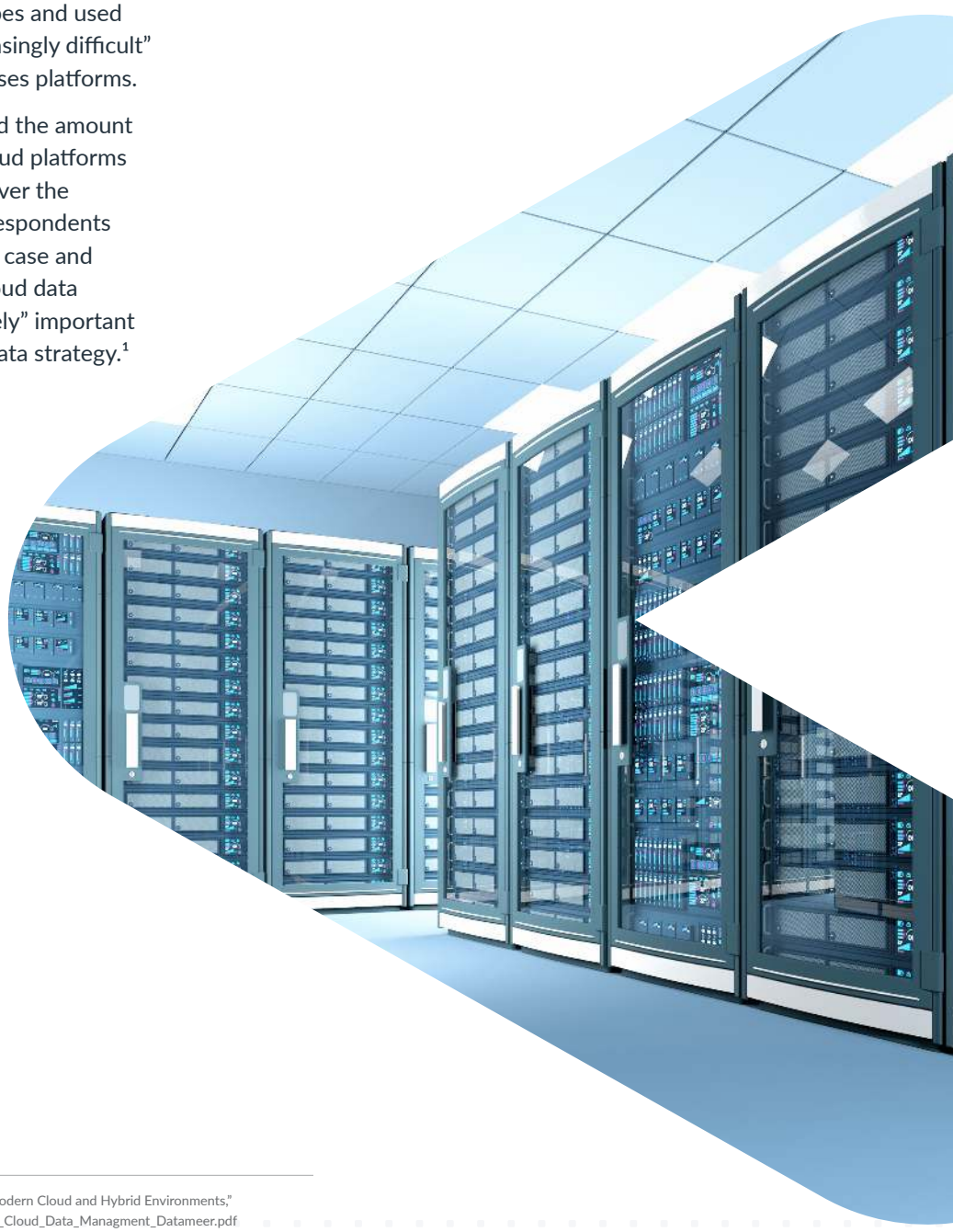
TABLE OF CONTENTS

WHAT THE MARKET SAYS

As the volume and diversity of data continue to grow exponentially, legacy data management systems no longer deliver on their purpose to easily and securely load and integrate data, enable rapid and democratized analysis, and make those insights available to all business users who need them. In a recent survey from TDWI, 75 percent of respondents said that as they adopted new data types and used new data sources, they found it "increasingly difficult" to capture and use data with on-premises platforms.

No surprise that the same survey found the amount of data management conducted on cloud platforms in any organization will at least triple over the next three years—with 64 percent of respondents selecting analytics as the intended use case and a whopping 86 percent considering cloud data management "extremely" or "moderately" important to the success of their organization's data strategy.[1]

With some upfront planning and consideration, migrating your data analytics solutions to the cloud can lead to big payoffs for your organization, especially if you're moving your legacy data solutions—including databases, data warehouses, and data lakes—to a cloud-built data platform. This ebook provides a roadmap for just such a migration.



[1] "Cloud Data management: Integrating and Processing Data in Modern Cloud and Hybrid Environments," datameer.com/wp-content/uploads/2019/05/TDWI_BPReport_Cloud_Data_Managment_Datameer.pdf

MIGRATION FRAMEWORK—APPROACHES, STRATEGIES, AND REQUIREMENTS

There are many reasons organizations choose to embrace cloud computing. But most organizations need a plan: something to grab onto that shows them what the future looks like. Not everyone is in the same place regarding analytic capability or cloud maturity. Therefore, give careful consideration to how much legacy code you want to move, and how fast you want to move it, from your on-premises environment to a pure cloud infrastructure.

THE FOUR MOST COMMON MIGRATION SCENARIOS

The type of migration you embark on will significantly influence your migration strategy. Here are four potential paths many organizations take to migrate their legacy data solutions to a cloud data platform:

1. OLTP for operational reporting and analytics

This is extremely common. Many organizations use online transaction processing (OLTP) systems, such as SQL Server, Oracle, or MySQL, for basic reporting and analytics. While this might work as a short-term solution, the reporting needs of the business compete with the operational needs, overtaxing a fixed resource and slowing performance for both. A truly elastic cloud data platform eliminates this problem. It's pretty easy. As discussed later, take your existing transactional schema, which is usually in third normal form (3NF), and move it, as is, to the cloud. This removes the reporting workload from the existing system and houses the data in a platform built for analytics, eliminating performance bottlenecks and, in some cases, giving your operational data store new life.

2. Appliance-based data warehouse

Many of the on-premises appliance vendors are sunsetting these legacy systems. More importantly, their customers want to escape the performance, cost, and other limitations of these systems that can't address the ever-changing data analytics needs of the modern enterprise. Appliance-based data warehouses also require huge upfront costs, typically in the form of capital expenditures (CapEx).

On the flip side, a data platform built for the cloud is designed to grow with you from zero upfront costs, thanks to a pay-as-you-go model, representing an operational expenditure (OpEx). This removes the guesswork of planning for your biggest day of consumption and then overpaying for an underutilized system for the other 364 days of the year. Similarly, if you need to expand your analytics unexpectedly during the year, you are hamstrung if you have a system that can't dynamically adjust to meet your needs. A cloud data platform removes that worry, giving you the ability to scale instantly. Another benefit is near real-time access to data. Since on-premises appliances are a fixed resource, data management teams for those systems create overnight load windows to make data available for the next morning. Today's cloud-built technologies allow you to segment workloads and load data 24/7 without impacting query processing, speeding the time to value of your data.

3. Data marts

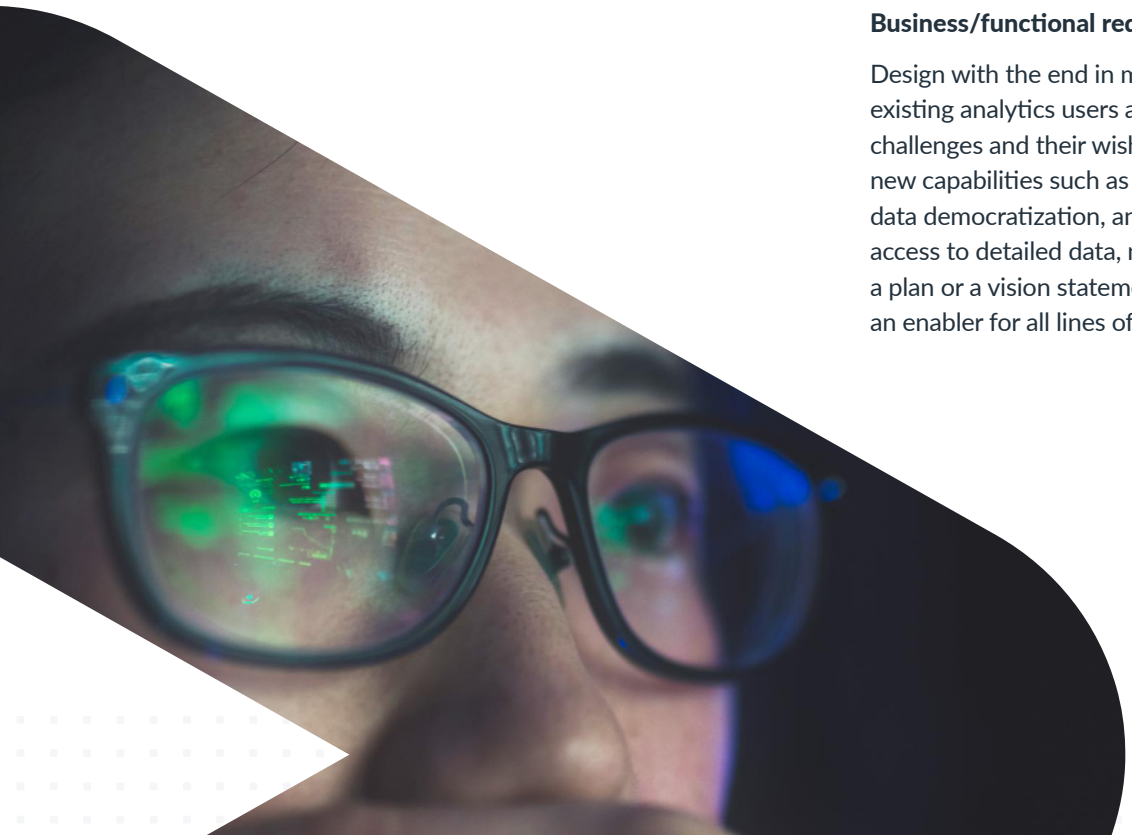
Most organizations suffer because a single source of truth is always out of reach. They have data siloed in many repositories. They may have tried to federate access across these repositories but quickly realized the cost to create and maintain that access wasn't feasible. They need a centralized repository to eliminate these barriers to getting all the insight from all their data. A cloud-built data platform becomes an obvious choice for data consolidation because it's ACID-compliant (transactionally consistent), can be partitioned/segmented logically without replicating, and can scale computing resources up and down, and on-demand. Whether you're a Kimball or Inmon fan, you need a platform that separates compute from storage and allows end users the greatest flexibility to access data sets with enterprise tools that leverage ANSI SQL and other languages that enable advanced analytic workloads.

4. Data lakes

Data lake initiatives have become a proverbial black hole: easy to get data in, complex to get data out. Many enterprises have realized that on-premises Apache Hadoop infrastructures are costly and complex and don't meet their analytics and concurrency requirements. Fortunately, there is a path forward. Leveraging a "layered" or "zones" approach is a great way for an enterprise to identify data sets it can comfortably move to the cloud. This method makes it easy to show the movement of data in a controlled and secure manner from an on-premises environment into a cloud storage infrastructure such as Azure Blob Storage, AWS S3, or Google Cloud Storage.

STRATEGY REQUIREMENTS

Migrations aren't much different than most IT projects, which means they usually begin with requirements. Defining requirements often cross multiple boundaries, since a cloud migration strategy can be an executive-level decision. Without executive buy-in, your project will be limited in scope and be labeled as "shadow IT." This might work for some lines of business to get started. Eventually, however, everyone needs to be on the same page, from the architecture team to the security team to the chief data officer and even the CFO.



Organizational strategy

For many organizations, the cloud is in their DNA and even written into their mission statements. Others that have been around awhile may know they need to modernize, but not at the expense of changing too quickly. The business benefits of the cloud are hard to ignore—more agility, lower costs, and deeper analytics, for starters. But your cloud migration project should move at a pace consistent with your corporate objectives. Determine whether the project is the "tip of the spear," a way to get the company moving toward the cloud in the right direction, or whether it is part of a larger cloud initiative that would allow sharing best practices and technical resources. One of the critical success factors will be determining which data sets are ready to leave your data center first and which data sets will follow later to the cloud.

Technical strategy

Every strategy starts with the same question: "What are we trying to do?" For technology people, this can be defined in the requirements. There is an old adage that says, "You can have it fast, good, or cheap. Pick two." Using a combination of agile development and taking advantage of what the cloud offers, that adage is more antiquated than accurate.

Business/functional requirements

Design with the end in mind. Discuss goals with existing analytics users and understand their current challenges and their wish lists. The cloud allows for new capabilities such as near real-time data access, data democratization, and next-level analytics with access to detailed data, not just aggregates. Create a plan or a vision statement that highlights being an enabler for all lines of business (LOBs), with the

appropriate security controls and tools. Use these goals to align IT and LOBs to help set the vision of securely getting accurate information to the right people at the right time.

Non-functional requirements

Take stock of what often are called the "ilities": policies related to service level agreements (SLAs) between IT and the business, the security requirements for protecting your data, usability requirements of your end users, and many other essentials. Common topics include:

- **Security**
- **Reliability**
- **Performance**
- **Maintainability**
- **Scalability**
- **Usability**

With the cloud, don't be afraid to include some aspirational requirements. Once you have a solid list, label the items as either nice-to-have or must-have. Be wary of analysis paralysis, and choose an implementation window and development methodology—agile, waterfall, and so forth—that is comfortable for your organization. Pay particular attention to your high-availability and disaster-recovery (HA/DR) requirements. A cloud-built data platform can save your organization from having to design elegant but expensive solutions to meet the needs of the business.

THREE APPROACHES FOR IMPLEMENTING YOUR MIGRATION

Now it's time to figure out what type of migration would make sense. You have options, which include lift and shift; lift, "improve" and shift; and full redesign. The steps you've taken prior to this stage, such as aligning the migration with your organization and defining the technical and business strategies, will drive your chosen migration strategy.

Lift and shift

Most would consider this to be the safest and most straightforward way to do a migration. The plan is simple: Everything you do with the existing system should be exactly the same in the new system, with minimal changes. A lift-and-shift strategy is a good one if:

- **Requirements are narrowly defined (very few new requirements)**
- **Time to implementation is critical (you need to get off the old system ASAP)**
- **Your new system has all of the features and functions of the old system**
- **Your ecosystem of surrounding tools (ETL, BI, system management) requires minimal or no changes**
- **The migration is not the centerpiece of your cloud migration strategy (see below)**

The last item on that list is debatable, because many cloud migrations require technical changes and changes to the culture of an organization. If your first initiative doesn't provide more than just the same features and functions as your old system, can your organization view it as a "win"? Sometimes, just showing you can migrate without risk to the business, while improving performance in some way, can be good enough for your stakeholders.

Lift, "improve," and shift

This is by far the most popular approach. The concept is simple. As you're converting assets, look for opportunities to streamline or improve the data pipeline, how data is organized, when data is transformed, and how data is accessed. Then, find ways to take advantage of new capabilities/functions in the system to which you're migrating. The point isn't to change any of the core functionality of the system but simply to take advantage of the opportunity to simplify or streamline.

The benefit here is to show some improvements over the existing process without breaking things and introducing too much risk. The executive team can use the migration as a proof point to the business. Even though this represents a major shift in IT philosophy, it does so without negatively impacting performance and provides additional business benefits, including:

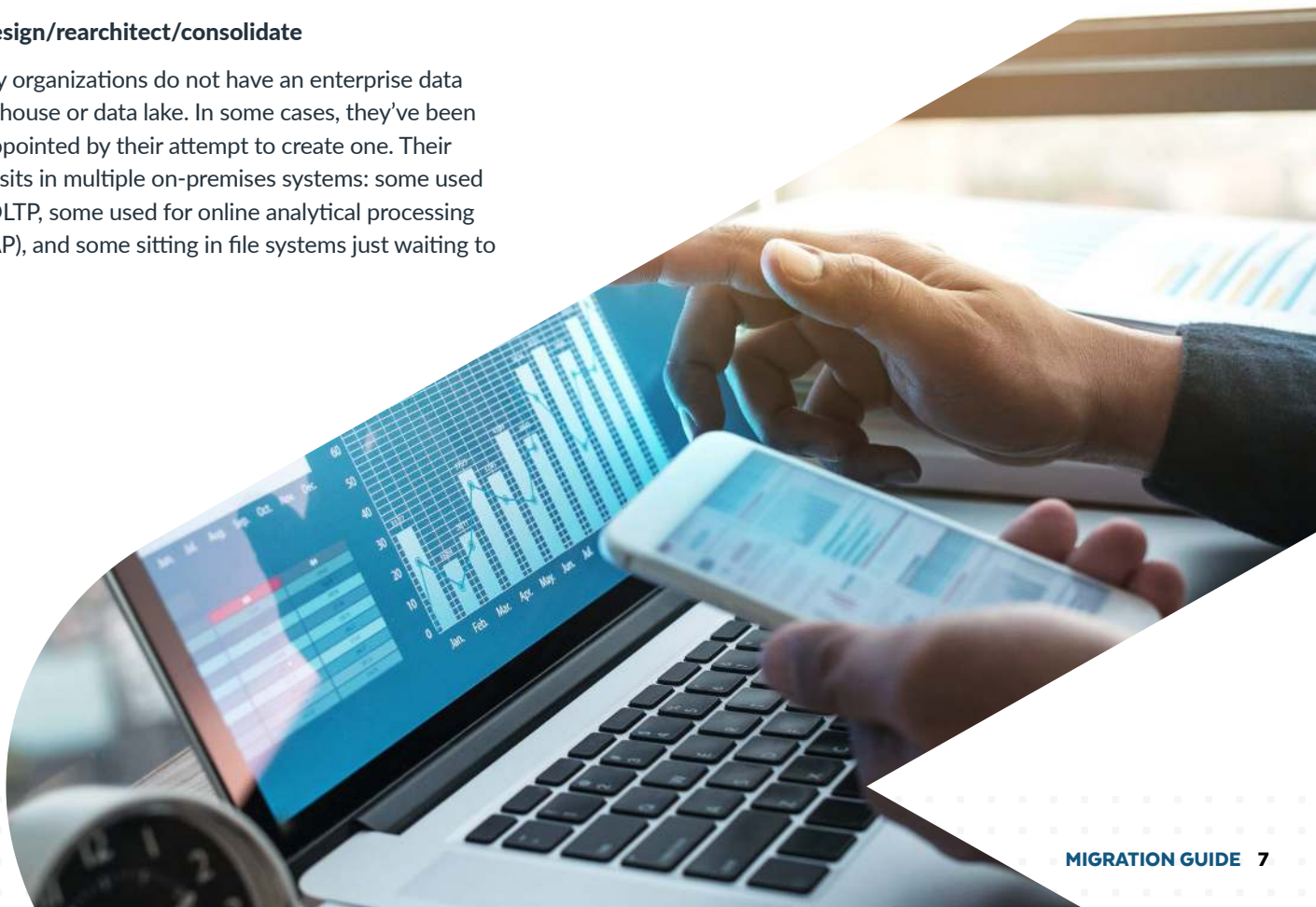
- **Faster access to more data**
- **More-granular data analytics**
- **Better performance on individual queries/reports**
- **No contention for near-unlimited computing resources**

Redesign/rearchitect/consolidate

Many organizations do not have an enterprise data warehouse or data lake. In some cases, they've been disappointed by their attempt to create one. Their data sits in multiple on-premises systems: some used for OLTP, some used for online analytical processing (OLAP), and some sitting in file systems just waiting to be analyzed. Changing platforms is viewed as an ideal time to re-architect, or architect for the first time, a fully functional data platform capable of scaling with the business.

The platform must meet the requirements outlined in the above sections: it must handle multiple types of data, and it must allow end users to employ their favorite tools and languages, such as SQL, to create a data democratization strategy for all LOBs. In some cases, the cloud data platform even creates a new opportunity for modern data sharing across and outside an enterprise. Projects such as this are a great way to consolidate infrastructure and get a better handle on contracts, security, and shadow IT, while producing incremental results for the business.

Most consolidation efforts start by combining multiple data sets onto the new cloud platform to show their analytic value. The next step would be to gradually restrict access to the legacy systems, while growing the capability of the cloud data platform and establishing quick wins. Many of these initiatives also focus on creating new revenue streams, shrinking or eliminating data pipelines, and consolidating disparate data repositories.



PART TWO

PLANNING YOUR MIGRATION

Executing these steps, and in this order, is not a necessity. Depending on the scope of your migration, you may need more or fewer steps. The key is to design a framework and core elements of the plan that you can work from. Assess your internal skill sets, don't be afraid to leverage the best practices outlined by strategic vendors, and consider partnering with migration experts.

STEP 1: DETERMINE THE SCOPE

Stating the obvious, no two migrations are the same, and rarely is the end state well understood. The goal is to create a plan that aligns with the goals of the business, provides capabilities in the shortest reasonable timeframe, and sets you on the path for incremental improvement. Your end state could be getting a single workload into the cloud within one month, or it could be migrating your entire analytics platform by the end of the year. It's reasonable to plan for a one-year ROI, which you can even accelerate under certain scenarios.

STEP 2: DOCUMENT THE "AS IS"

This isn't the most glamorous part of a migration, but it's likely one of the most critical. You'll need to communicate both internally and externally, and up and down the reporting chains, regarding the current "as is" implementation. Assets to migrate include but are not limited to:

- **All sources that populate the existing systems**
- **All database objects (tables, views, users, and so forth)**
- **All transformations, with schedules for execution or triggering criteria**
- **A diagram from the interaction of systems/tools**

STEP 3: DETERMINE THE APPROACH AND ASSEMBLE THE IMPLEMENTATION TEAM

Multiple options exist here. We've outlined above the most common approaches, but there are combinations of these. You could choose to implement one method to get to initial capability and another as you approach full production. Creating high-level milestones at this step is a good way to segment when a capability will be available, and which requirements you'll satisfy via release schedules.

STEP 4: DIAGRAM THE "TO BE"

Once you get your arms around what you want to migrate and when you want capabilities available, you can begin to document the "to be" architecture. Be aware that there is no "one chart to rule them all." You'll be communicating to technical, business, and executive audiences, and each will want different levels of detail on the initial operating capability (IOC) and final operating capability. Figure 1 illustrates a "to be" example of modern cloud data platform architecture.

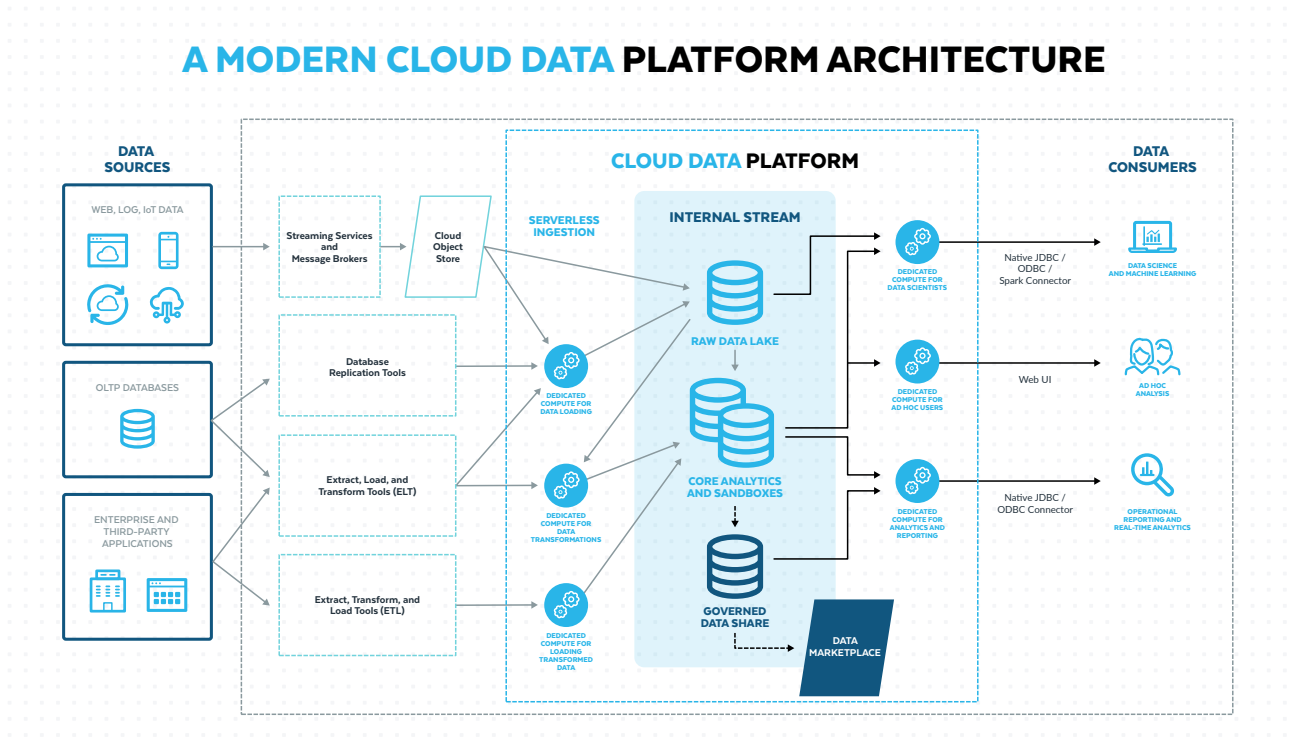


Figure 1: Design your modern cloud data platform architecture to represent your technical, business, and executive needs.

STEP 5: PLAN FOR YOUR DATA LOAD AND "SIZING"

One of the most challenging aspects of migrating to the cloud is moving data and changing your paradigm to take advantage of the elastic resources of the cloud. There are three pieces to the puzzle:

1. Initial load

The initial load can be challenging based on data volumes and security requirements. Work closely with your security team and the LOBs that own the data to make sure you don't have to go through a tokenization/obfuscation process before moving data into the cloud. Then decide which data sets are okay to move, taking into account regional and industry-specific regulatory and data privacy compliance standards such as PII, PCI, and HIPAA. Regarding the volume of data, as networking gets better and you can move terabytes of data into the cloud, this issue starts to go away.

TIP: Many organizations receive data externally from partners and vendors (Salesforce, for example). It might be prudent to dump these data sets into a cloud object storage service to keep them from becoming classified as on-premises assets. If data is coming over the Internet, it should be okay to secure it in the cloud.

2. Ongoing updates

Each source of data, the ETL logic, and integration will dictate the methods used for updating data in your cloud data platform. Some sources and ETL tools support change data capture (CDC) strategies. Others might support all inserts, while others might require a full refresh. There is no one-size-fits-all approach.

3. Planning for usage and storage

Most organizations execute a proof of concept (POC) and go through a return on investment (ROI) exercise before executing a migration. At this phase, it's usually a good idea to revalidate the usage plan and work the operational side of the equation with regards to how to monitor availability and how to govern usage of the system. Because some cloud data platforms provide the ability to scale up and down, turn resources on and off, segment workloads and resources, and auto-scale both processing power and storage, the model changes from time spent slicing a fixed resource (limiting your business user access) to allocating resources based on business need and value. You no longer have to do a big planning exercise to handle your largest workload and leave the system underutilized for the other 364 days of the year.

STEP 6: CONVERT ASSETS

This step refers to defining data assets you may need to convert. These include data definition language (DDL), role-based access control (RBAC), and data manipulation language (DML) used in scripts. The good news is that most relational databases leverage the ANSI-SQL standard. Most of the changes will revolve around ensuring DATE and TIMESTAMP formats are converted correctly and the SQL functions used to access those are checked for compliance. (Not all vendors implement functions the same way.) Some cloud data platforms simplify DDL by eliminating the need to partition and index, so your DDL becomes much cleaner (less verbose).

STEP 7: SET UP YOUR "TO BE" ENVIRONMENT AND TEST CONNECTIVITY/SECURITY

It should be no surprise that you'll have to complete your networking, proxy, and firewall configurations during your migration strategy. It usually helps to have a chart or two outlining what ports and URLs you will need to access. You will also want to work with your security group to download and install any drivers (ODBC, JDBC, and so forth) or support software such as a command line interface (CLI), which most DBA-type developers prefer to use when interacting with a modern cloud data platform. You will also want to set up your account parameters such as IP whitelisting and role-based access control before opening up the environment to larger groups.

STEP 8: TEST THE PROCESS END TO END WITH A SUBSET OF DATA

Once you have connectivity worked out for all tool sets, it's best practice to test your process from end to end for both functionality and performance. If you're coming from an existing system, you'll have the advantage of knowing what your SLAs are with each LOB and your requirements for each step in the process—load, transform/aggregate, query execution, and so on. It's always best practice to select test cases and data sets critical to early success. This is also an opportunity to implement the "improve" part of the migration, if that's the methodology you chose.

STEP 9: MIGRATE THE DATA AND TEST PERFORMANCE

Before going live, you'll want to rerun some or all of your performance tests from Step 8 to ensure the system is configured for individual query performance and for concurrency. You'll want a champion from each of the critical lines of business to test accessing the system, making sure their tools work, and ensuring they get the performance you expect for them. Validate that you're getting the same calculated results from the old to the new system.

STEP 10: RUN YOUR EXISTING AND NEW SYSTEMS IN PARALLEL

When replacing an on-premises enterprise data warehouse or data lake, it's easy to dual-load the systems and run them in parallel. From there, gradually move users or groups onto the new system, trying not to disrupt business operations. Focus on the groups that expressed the most challenges or concerns during the planning and requirements phase, because they will be the most receptive and probably become the most vocal advocates about success.

STEP 11: PICK YOUR CUTOVER DATE

You can run the systems in parallel for a while, but once your end users experience the new performance benefits, they will never want to go back. You'll likely want to run the systems in parallel for at least one major reporting cycle—a week, a month, or a quarter. Once you pick an official cutover date, continue to dual-load the systems in case you run into a problem. Once everyone seems happy, it's time to pop the champagne and retire the old system.

PART THREE

TAKING ADVANTAGE OF YOUR NEW CLOUD DATA PLATFORM

Now that your migration is complete, it's time to start taking advantage of the new capabilities. One strategy is to look for ways to improve performance and get data into the hands of end users more quickly, or distributed to more users. Some options include:

DETERMINE LOB AND END-USER NEEDS

You compiled an initial list of requirements from the LOBs earlier in the process. Now it's time to revisit that list and start a dialog with the LOBs, educating them on what's possible with the new system. Some of these new capabilities include but are not limited to:

- **More data—access to detailed records reaching back years, not just months or weeks**
- **Different data types—structured and semi-structured (JSON, AVRO, XML, PARQUET, ORC)**
- **Cleaner data**
- **Better formatting/modeling—changing schemas from 3NF to star or other data models**
- **Faster performance—many organizations use summary tables or materialized views to improve performance**

ACCESS TO CROSS-BUSINESS-UNIT DATA

Along with the LOB-specific data sets, most organizations also want access to the latest and greatest information from other business units and sources. Some modern cloud data platforms provide several methods for controlling data access while still enabling curated data sets to other end users. In some cases, organizations share or receive data sets via FTP with other organizations. Moving data this way is time-consuming and expensive, especially as data volumes and frequency increase. Some modern data sharing features associated with modern cloud data platforms eliminate the need to transfer and transform data, streamlining the process and reducing ETL cost and complexity. They are also more secure and enable you to share live, governed data that complies with data privacy regulations.

OPTIMIZE (RETHINK) YOUR LOAD STRATEGY

Typically, on-premises data warehouses have load windows to execute in batch overnight, making yesterday's data available for analysis the next morning. With today's technology, you can load data and query data without contention, opening the possibility to load data 24/7, and providing data access sooner to end users. You can spin resources up and down instantly, which allows you to load data as fast as you want, and at the same price point no matter the compute resources you spin up.

For example, let's say it takes four hours to load 1 TB of data every night. In some cases, you're using one node for this workload. With near-linear scalability, if you execute the same workload with a two-node cluster, it will load twice as fast but cost exactly the same. Double the cluster size again, and it gets done twice as fast again at exactly the same cost.

ANALYZE SOURCES FOR SIMPLIFICATION/STREAMLINING

In some cases, the tools and systems you were using in the past may not be necessary in the future. Today's modern cloud data platforms can handle new data types and process data more efficiently, blurring the lines between OLTP, OLAP, and data lakes. Large organizations tend to have many tools, all bought for specific functions or capabilities.

Now is a good time for an ETL/ELT tool rationalization strategy. Some of these vendors have upgraded their products to work in a cloud environment. Do not try to standardize on a single tool. Instead, make technical recommendations for accessing specific sources, or, if you are implementing a data lake strategy, use tools appropriate for accessing different zones or layers.

You should also analyze the tools in the context of how quickly they can move data from your original system to your new platform. Most organizations are moving from extract, transform, and load (ETL) strategies to extract, load, and transform (ELT) and stream-processing strategies to take advantage of on-demand scaling. By separating compute resources from storage resources, you have the ability to load data without impacting query performance, making near real-time data processing at scale a reality. You just need to figure out how much change your organization can absorb at one time.

IDEMPOTENT LOADING

Never heard of it before? This isn't a new term, but it has been coming up in lots of conversations related to data processing. The goal is simple. You want to continually load data into your system, but if something goes wrong along the way you don't want to get confused about what has been processed and what hasn't. Idempotent loading means that it doesn't matter if you load a file or record once or 10 times, you end up with the exact same result. This is pretty powerful and can be achieved via a combination of the COPY command and transformation using the MERGE commands found in modern data platforms. It's especially good for streaming data with updates or deletes, but you can implement it in other situations as well.

ZERO-COPY CLONE, TIME TRAVEL, AND UNDROP

Some modern cloud data platforms have the ability to query "back in time" between the previous 24 hours and 90 days. Many organizations use this capability to transform the ELT pipelines. So, if something goes wrong in step 45 of a 50-step process, you don't have to start over. You use time travel to set yourself up to a known state (step 44 that executed correctly), fix step 45, and continue processing. Most legacy platforms would force you to reload or recover the original table and start over. If you need a snapshot of your database or table before doing an update, use a zero-copy clone function to make a copy of either but without duplicating the data. That's a big deal, especially if you're currently doing large ETL jobs to pull terabytes of data into data marts for data science or test/QA teams to work with. And finally, did you accidentally drop a production-level database or table while doing the nightly change? No worries. UNDROP works really well in that situation, and magically all of the data and schema reappears.

CONCLUSION

MORE AND MORE DATA IS BORN IN THE CLOUD

Enterprises now realize there are efficiency and performance advantages to storing, analyzing, and sharing data in the cloud. This approach removes numerous steps for enterprises, and removes the chaos that can result from siloed, duplicated data that becomes disconnected from its original source. A carefully planned migration can lead to significant advantages over a conventional data warehouse or data lake, including more capabilities at lower cost.

LEARN MORE

Click here to get more information about how to modernize your data platform and to get instructions specific to migrating from your existing, on-premises data warehouse or data lake to Snowflake Cloud Data Platform.

ABOUT SNOWFLAKE

Snowflake Cloud Data Platform shatters the barriers that prevent organizations from unleashing the true value from their data. Thousands of customers deploy Snowflake to advance their businesses beyond what was once possible by deriving all the insights from all their data by all their business users. Snowflake equips organizations with a single, integrated platform that offers the only data warehouse built for any cloud; instant, secure, and governed access to their entire network of data; and a core architecture to enable many other types of data workloads, including a single platform for developing modern data applications. Snowflake: Data without limits. Find out more at **snowflake.com**.