Seven Considerations When Building a Data Warehouse Environment in the Cloud

By Wayne W. Eckerson and Stephen J. Smith

Research Sponsored by

About the Authors

Wayne W. Eckerson has been a thought leader in the data and analytics field since the early 1990s. He is a sought-after consultant, noted speaker, and expert educator who thinks critically, writes clearly, and presents persuasively about complex topics. Eckerson has conducted many groundbreaking research studies, chaired numerous conferences, written two widely read books on performance dashboards and analytics, and consulted on BI, analytics, and data management topics for numerous organizations. Eckerson is the founder and principal consultant of Eckerson Group.

Stephen J. Smith is the practice leader for data science at the Eckerson Group. His unique perspective comes from his real-world experience in building the predictive analytics products Darwin, Discovery Server and Optas, which were among the first to deliver machine learning on an MPP computer architecture, implement algorithms directly in SQL, and embed model results in an OLAP tool. He has written the best-selling business technology books: "Data Warehousing, Data Mining and OLAP" and "Building Data Mining Application for CRM" with McGraw-Hill. He received his undergraduate degree in engineering from MIT and his graduate degree from Harvard in machine learning.

About Eckerson Group

Eckerson Group helps organizations get more value from data and analytics. Our experts each have more than 25+ years of experience in the field. Data and analytics is all we do, and we're good at it! Our goal is to provide organizations with a cocoon of support on their data journeys. We do this through online content (thought leadership), expert onsite assistance (full-service consulting), and 30+ courses on data and analytics topics (educational workshops).

Get more value from your data. Put an expert on your side. Learn what Eckerson Group can do for you!

Table of Contents

Introduction

For the past decade, most companies have resisted implementing a data warehouse (DW) in the cloud, largely due to concerns about security. In addition, few had experience using cloud-based applications. But times have changed. Today, the number of companies using the cloud for data warehousing or business intelligence (BI) has increased nearly 50% since 2013, according to a recent survey by Eckerson Group and the Business Application Research Center (BARC).

Benefits

The cloud is attractive for many reasons. Perhaps the biggest is deployment speed. Business units no longer need wait for procurement, legal, the project management office (PMO), and the information technology (IT) department to deploy a data warehouse; they can simply shop at the Amazon AWS marketplace (or other cloud vendors) and spin up a data warehouse, such as Amazon Redshift or Azure SQL Data Warehouse. The cloud vendor installs and administers the hardware and software and manages all upgrades, patches, and interoperability issues. And for an extra fee, some will even design and manage the data warehouse for you.

The best part is that the cloud platform vendor doesn't charge a large up-front fee that requires capital allocation and lots of internal meetings to gain approval. With a monthly, annual, or pay-as-you-consume subscription, organizations can now stand up a cloud data warehouse in a matter of weeks.

A big attraction of the cloud is that it is infinitely scalable and equally elastic.

Another big attraction of the cloud is that it is infinitely scalable and equally elastic. Rather than estimate capacity years in advance, organizations can simply add more capacity by provisioning additional servers online. And they can dynamically add capacity to handle peak loads using built-in scripts or shift capacity to different workloads (e.g., database, ETL, BI) on a schedule. Studies have shown that most data centers use less than 30% of their available computing power (see "Cloud Computing: The Business Perspective"). By maximizing usage, the cloud offers a more efficient and affordable computing platform.

Rather than migrate existing data warehouses to the cloud, many companies build net new applications there. IT departments use the cloud to spawn fully featured development and test environments and conduct proofs of concept; data science teams use it as an analytic sandbox with a complete replica of the data warehouse; and business units use it to spawn departmental data warehouses or data marts. These stepwise experiments give companies valuable knowledge and experience about running data analytics in the cloud.

Considerations

Of course, the cloud is not all roses. BI teams must address significant issues before taking the plunge. We will examine seven challenges that organizations need to consider before implementing a cloud data warehouse:

1. Data Delivery
2. Design and Development
3. Security and Compliance
4. Hybrid cloud
5. Portability
6. Politics
7. Pricing

This checklist report is designed to help you and your team identify key questions to ask before implementing a cloud data warehouse.

Considerations

❶ Data Movement

The cloud poses a potential bottleneck to data movement. In a corporate data center, data moves across a high-speed, local-area network, but in the cloud, data moves across a thin internet pipe. Without careful planning, an organization might not be able to complete its nightly loads on time.

It takes more than two days to move a terabyte of data across a relatively speedy T3 line (20 GB/hour).

It takes more than two days to move a terabyte of data across a relatively speedy T3 line (20 GB/hour). And that assumes no service interruptions, which might require a full or partial restart. BI leaders need to work closely with their company's network team to ensure there is sufficient network bandwidth to support data transfers from an on-premises data center to a cloud platform. If data sources already reside in the cloud, they'll need to evaluate the network connections between cloud platforms, if different ones are used for operational applications and the data warehouse.

Initial Loads. It's also imperative to develop strategies for both the initial load of historical data into the data warehouse (assuming the business wants to prepopulate such data) and ongoing loads and updates. For an initial load, some companies ship data on disk to the cloud provider, who loads it manually into the cloud data warehouse. Amazon recently launched its Snowball service for extremely large data sets: it sends a truck with an onboard data center to your premises to establish a secure, direct connection to your system.

Before loading data, administrators need to compare estimated data volumes against network bandwidth to ascertain the time required to transfer data to the cloud. In most cases, it will make sense to use a replication tool with built-in change data capture (CDC) to transfer only deltas to source systems. This reduces the impact on network traffic and minimizes outages or delays. As an enabling technology, CDC supports both batch loads that incrementally update the data warehouse as well as continuous feeds that support real-time analytics as well as higher efficiencies and productivity.

Load Utilities. Even if you can move data quickly and efficiently across a network, load times will suffer unless the cloud data warehouse supports a fast data loading utility and rich interface to a replication or data movement service. Many data warehouses run on high-performance analytic databases that load and process data in parallel. A replication service needs to know how to feed parallel data streams into a database loader to make full use of its processing capacity. Some replication tools, such as those from this report's sponsor, Qlik, have the ability to create and load database tables as well as perform lightweight transformations.

❷ Design and Development

Although deploying and administering a cloud data warehouse is fairly painless, designing one is not. Just because the data warehouse runs in the cloud doesn't obviate the need to model the database, create ETL programs, manage data jobs, and govern data. This process is similar to what is required on-premises, but there is a richer selection of tools in the cloud, many of which are extensions of existing and familiar on-premises tools.

Data Lakes. Many cloud providers have incorporated big data techniques to simplify certain aspects of designing and managing data warehouses. The most common strategy is to use a low-cost, scale-out distributed file system to stage and store data. This so-called data lake eliminates the need to model source data up front and build extensive ETL mappings and jobs to transform and load data into a specific database schema.

In the cloud, customers can basically dump data into the data lake (e.g., Amazon S3 or Hadoop HDFS) as a quick first step with little up-front design or modeling work. It is also very helpful that many cloud data warehouses support nonstandard data types, such as text, logs, JSON, or XML, making the cloud data warehouse a true repository of all data in the company.

From there, the design work begins. Developers must clean, encrypt, tag, and profile incoming data. Then they must parse, integrate, and transform the data to create analytic views or

data marts and conform them to existing dimensions and metrics. Data engineers can use a variety of tools to accomplish this work, including SQL, Hive, Pig, and other data manipulation languages as well as cloud-based ETL tools and new data pipeline software that runs on Hadoop clusters.

Rather than manually running jobs and managing changes, DWA tools automate these processes, supporting fast development cycles commonplace in the cloud.

Data Warehouse Automation (DWA) Tools. Since the hallmark of the cloud is agility and speed of deployment, data warehouse automation (DWA) tools are a natural fit. Rather than manually running ETL jobs and managing changes, DWA tools automate these processes, supporting fast development cycles commonplace in the cloud. These metadata-driven tools, such as Qlik Compose™ (formerly Attunity Compose), support an iterative approach to design, development, and operations that delivers a better outcome in less time.

❸ Security and Compliance

Data is vulnerable to hackers in transit to the cloud and once stored there. The cost of a security breach can be enormous in terms of government fines and employee or customer legal settlements. But the cost to intangibles, such as the company's brand and reputation, can be even more substantial.

Security Certifications. Fortunately, cloud providers stake their reputation on their ability to provide high levels of security. Most cloud platform vendors now boast a variety of security certifications, such as SOC 2 (Service Organization Control) or Type II SAS 70 (audits of service organizations for control of processes) and ISO 9000 (for validation of processes to assure quality). Most large organizations have put cloud providers through grueling security audits of their own. The good news is that many of these providers offer better data security than corporations provide in their own data centers.

Many companies now believe their data is safer in the cloud than in their own data centers.

Views on cloud-based security and privacy protection are changing. Many companies now believe their data is safer in the cloud than in their own data centers. Highly public and damaging security breaches at TJX, Target, and Sony have reinforced this notion. This is especially true for mid-sized companies. According to the Eckerson/BARC study, only 39% of mid-sized companies consider security a major challenge compared to 50% of large companies with more than 2,500 employees.

Nevertheless, companies need to do their due diligence about security protections before committing to a cloud data warehouse. Some questions to ask include:

- **Data Location.** Where will the data reside? Can the cloud provider pinpoint the servers and data center in which your data will be held? This is critical to conform to certain privacy regulations, especially those in Europe.
- **Data Co-Mingling**. Does the cloud vendor interweave your data with another company's data in the same database? For most companies, this is a non-starter. Most require (at additional cost) a virtual private server and database to isolate their data.
- **Backup/Restore.** What is the vendor's backup and restore service? Where are the backup servers? Is there geographic separation? How quickly can data be restored in case of a failure?
- **Encryption.** Is network and database encryption performed automatically, or are you responsible? If the cloud vendor holds the encryption key, what controls are placed to prevent the vendor (or the host nation government) from looking at your data?
- **Authentication.** Is multi-factor authentication available?
- **Alerts.** What facilities are there for intrusion or breach detection? How will you be notified?
- **Certifications.** What security certifications does the cloud provider support? How do they deal with moving personal information outside of the country? Do they have on-the-ground coverage in countries that have data movement restrictions?

4 Hybrid Cloud

In a hybrid cloud environment, data and computing resources are split between on-premises and cloud systems or between multiple cloud providers. This can increase complexity and costs, but it may be necessary to comply with industry regulations. Or it may be a fact of life as you transition a data warehouse from on-premises to the cloud. Some companies also run a hybrid environment to support elastic computing or maintain a data warehouse replica (i.e., analytics sandboxes) in the cloud.

A hybrid cloud environment requires high-speed network connections to ensure sufficient bandwidth to avoid performance issues.

A hybrid cloud environment requires high-speed network connections to ensure sufficient bandwidth to avoid performance issues. (See the data movement discussion above.) It also requires a secure connection, especially if BI users are on-premises and the data warehouse

is in the cloud or vice versa. Many companies establish VPN connections to safeguard communications and data and use HTTP tunneling approaches to punch through firewalls in a safe way.

BI Tools. The way a BI tool interacts with data can significantly impact performance and costs, especially if the BI tool runs on-premises and the data warehouse runs in the cloud. Some BI tools extract data from a data warehouse and load it into a local, high-performance database or cache, which fields user queries. If the extract happens dynamically in response to a query, this slows performance and racks up data transfer charges. Unless data volumes are small, most BI vendors recommend creating extracts or caches at night.

Other BI tools dynamically query a data warehouse, leveraging the built-in processing power of the database. This gives users access to all the data—not an extract—and the freshest data. It also minimizes over-the-wire data transport and charges. This approach works well for simple queries, but can suffer performance issues with long-running, complex queries.

Cross-Platform Automation. A final consideration is when an organization wants to apply data warehouse automation across platform boundaries. For instance, an organization might add Amazon Redshift to its portfolio and need its DWA tool to support both its existing on premises data warehouse and its new cloud data warehouse. Or it might split its existing Teradata data warehouse across on premises and cloud channels and use its AWS tool to populate and support both. A hybrid AWS product that works across platform boundaries eases migration challenges and improves total cost of ownership.

❺ Portability

Although the cloud promises liberation from heavy-footed data centers and IT departments, it doesn't guarantee freedom from vendor lock-in. Richard Stallman, creator of the Free Software Foundation, warned that cloud providers "trap" customers in proprietary systems that are difficult to escape. Vendors may then increase prices with impunity, because most customers won't expend the time and money to migrate off the platform.

To avoid vendor lock-in, take the following steps:

- **Legal contracts.** Review legal contracts carefully. Make sure you aren't locked into a long-term contract and that you don't pay penalties for an early exit. Also, make sure your contract contains a protection clause if the cloud provider goes out of business or is acquired.
- **Open standards.** Make sure the cloud provider uses open, standard products so porting data is easier. Many now use open source software, such as Hadoop, Spark, and Kafka, and others build on existing cloud platforms, such as Amazon S3 and other AWS services.

- **Customizations.** Avoid complex customizations to the cloud environment that are difficult to replicate and hard to upgrade when the cloud provider ships new releases of its underlying software or infrastructure.
- **Spread the Wealth.** Consider not putting all your data on one platform. Create a data warehouse backup or replica on another cloud platform or on-premises as a safeguard in case your strategy or plans change.
- **Universal Connectivity.** Make sure your connectivity and ETL tools run against multiple cloud platforms in case you need to change providers.

6 Politics

In the Eckerson/BARC research, "politics" was considered the second major challenge to deploying a cloud data warehouse. It was second only to security in terms of survey respondents' level of concern. Often, IT professionals are those who feel most threatened by the cloud. If the company outsources its computing and data infrastructure to a cloud provider, many IT professionals fear they will lose their jobs. And this is a distinct possibility.

However, IT professionals shouldn't worry too much. There is a lot of design and management work required to maintain a cloud data warehouse. (See "Design and Development" above.) Also, outsourcing hardware and database administration frees up IT professionals to focus on more value-added activities, such as architecture, design, and optimization.

Because many cloud data warehouses are initiated by business units, IT can use the opportunity to align itself more closely with the business.

Because many cloud data warehouses are initiated by business units, IT can use the opportunity to align itself more closely with the business, something that has eluded IT for decades. It can help business units modernize legacy data and analytics environments, improve data governance, establish data controls, evaluate cloud BI tools, and train and support business users, among other things. In fact, IT just might discover that a cloud data warehouse is its ticket to a successful future.

7 Pricing

Pricing in the cloud is complex and sometimes requires a specialist to untangle it. However, as more companies implement data warehouses in the cloud, providers are under pressure to simplify pricing so mere mortals can understand it.

The cloud is not necessarily cheaper than a comparable on-premises system. [In fact,] most customers hit a breakeven point within two to five years….

What is clear, however, is that the cloud is not necessarily cheaper than a comparable on-premises system. Because cloud vendors use subscription pricing, most customers hit a

breakeven point within two to five years, after which the cloud platform in aggregate becomes more costly than an on-premises system. However, there are many variables, so it can be difficult to fully compare costs.

Also, each cloud vendor prices its offerings differently. Some charge by physical resources, such as memory, CPU, disk, solid-state drives, and network bandwidth, while others charge by gigabytes of storage per month, number of queries, or data transfers between cloud servers or outside the cloud platform. Most also charge additional fees for services, such as auto-scaling, encryption, and managing software updates.

As customers discover these pricing bumps, we suspect cloud providers will smooth them out to avoid alienating prospects.

Prices also change frequently. Vendors continuously seek a balance: They must keep costs low to attract new customers, yet generate enough revenue to cover the costs of running complex data centers and multi-tenant software. Although initial costs are usually attractive, many customers have discovered a hockey-stick ramp-up in fees after their environment reaches a certain level of data processing. As customers discover these pricing bumps, we suspect cloud providers will smooth them out to avoid alienating prospects.

Ways to Lower Costs. Before consulting a vendor price sheet, it is important to know what services you need. You may need to conduct a proof of concept with one or more vendors. Most also provide price calculators to give you a ballpark estimate, but it's best not to rely on these.

You'll also need to manage your environment astutely to avoid costs. For instance, it's important to de-provision CPU resources when they are not needed. You can also move your data from expensive analytic resources (e.g., analytic databases such as Amazon Redshift or Google Big Query) to less expensive storage (e.g., Amazon S3 or Google Nearline) when it is not being used.

Most providers do not charge for moving data into the cloud, but most charge when exporting it.

Storing and Retrieving Data. Finally, make sure you understand the costs of moving and storing your data. Most providers do not charge for moving data into the cloud, but most charge when exporting it. And fees change as data volumes grow. Currently, it is about as expensive to retrieve a gigabyte of data as it is to store it for a month, but it is more expensive to retrieve a petabyte than store it for a month.

Finally, customers should know that most cloud platform vendors use standard hardware and databases to run customer workloads. But customers who want to use a specialized analytic database with MPP processing will likely pay a premium. And they'll pay more to reserve a dedicated instance of the database.

Is a Cloud Data Warehouse Right for You?

The cloud offers a raft of benefits: speed, agility, elasticity, scalability, and subscription-based pricing. And although more organizations are embracing the cloud for their data analytics workloads, there are several challenges they need to consider. Chief among them are data movement, security, pricing, and politics.

Before embarking on a cloud data warehousing initiative, consider whether the cloud is a good fit for your organization. Some companies now have a "cloud-first" policy that encourages business and technical managers to evaluate the cloud against other alternatives when planning a new data-driven project.

Key Factors. There are four key factors that tend to determine whether organizations will implement a cloud data warehouse:

1. The degree of business users' dissatisfaction with their current data environment.
2. A business culture encourages innovation and risk.
3. The degree of autonomy that a business unit or department has in making IT-related purchases.
4. The percentage of operational applications that already run in the cloud, which is a good indicator of the company's openness to cloud computing.

If your organization scores high on these four factors, there's a good chance it has already evaluated or implemented a cloud data warehouse or is planning to do so in the near future. Before embarking on a cloud data warehouse, take into account the seven considerations described in this report to ensure a successful project.

About Eckerson Group

Wayne Eckerson, a globally known author, speaker, and advisor, formed Eckerson Group to help organizations get more value from data and analytics. His goal is to provide organizations with a cocoon of support during every step of their data journeys.

Today, Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps you abreast of the latest trends, techniques, and tools in the data analytics field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate your business requirements into compelling strategies and solutions.
- **Our educators** share best practices in more than **30 onsite workshops** that align your team around industry frameworks.

Unlike other firms, Eckerson Group focuses solely on data analytics. Our experts each have more than 25+ years of experience in the field. They specialize in every facet of data analytics—from data architecture and data governance to business intelligence and artificial intelligence. Their primary mission is to help you get more value from data and analytics by sharing their hard-won lessons with you.

Our clients say we are hard-working, insightful, and humble. We take the compliment! It all stems from our love of data and desire to help you get more value from analytics—we see ourselves as a family of continuous learners, interpreting the world of data and analytics for you and others.

Get more value from your data. Put an expert on your side.
Learn what Eckerson Group can do for you!

About Qlik

Qlik's vision is a data-literate world, one where everyone can use data to improve decision-making and solve their most challenging problems. Only Qlik offers end-to-end, real-time data integration and analytics solutions that help organizations access and transform all their data into value. Qlik helps companies lead with data to see more deeply into customer behavior, reinvent business processes, discover new revenue streams, and balance risk and reward. Qlik does business in more than 100 countries and serves over 50,000 customers around the world. For more information, visit www.qlik.com.