Avoiding Cloud Database Pitfalls and Lock-ins

The critical capabilities of your next cloud analytics solution

By

Steve Sarsfield

About the author: Steve Sarsfield is an independent author and has had thought leadership roles at Cambridge Sematics - AnzoGraph, Talend, Trillium Software and IBM. Steve's writings have produced a popular data governance blog, articles on Medium.com and a book entitled the Data Governance Imperative. Steve continues to offer insight and opinion in the world of data governance and analytics.

Evaluating Solution Completeness

It is a little considered fact, but it takes a talented development team years and many person-hours to build a complete database system. A mature, fully functional DBMS includes a long, difficult-to-build list of features. There are few shortcuts. While users explore early versions of a new SQL database, completeness as an enterprise solution is not formed until many versions later. Shown on the right are some of these long-term development features.

EPIC Features of a Cloud Database

- Scaling and elasticity
- Backup
- Workload management
- Admin and Management
- Security and encryption
- Deployment options
  (SaaS, cloud, containers, on-premises)
- Ecosystem Connections
  - Open Source
  - Data integration
  - BI and Visualization
  - Data file formats
- Query optimization

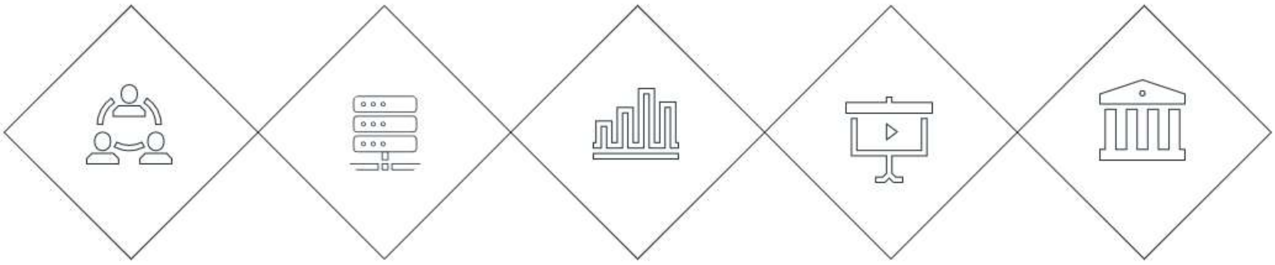
Cloud databases are no different. With so many databases on the market at various development stages, IT must consider the completeness of the database system when evaluating how well the solution meets the needs of the organization.

In this white paper, I'll explore a framework you can use to evaluate databases. Whether you are planning to move all your analytical workload to a single public cloud vendor, multiple cloud vendors, on-premises using cloud technologies like object store, or plan to have a combination of on-premises and cloud, the analytical engine you choose can have a significant impact on costs and productivity.

A framework for cloud database capability

In many corporations, there is no clear framework for how databases are selected. Business analysts often pick which solutions they want to adopt, and as usage increases, that drives adoption of a preferred platform. Data scientists and business analysts bring knowledge about languages, visualization tools, and analytics engines when they join a company already preferring the tools they know. On the other hand, IT and data governance teams tend to consider other crucial factors like data governance, operational efficiency, and cost control on an equal footing to deployment simplicity.

As you evaluate solutions for capabilities and long-term cost to the organization, it's essential to consider the following main categories of how a cloud-based analytical platform behaves:



Deployment

Options available for analytics deployment

Storage and Workload

Able to handle alternative storage and analyzing data in-place

Optimization

Options available to improve slow queries

Analytics

Able to provide analytics to serve a wide audience

Data Governance

No surprises in costs or security features

Deployment

Cloud database deployment varies widely, but the decisions made during implementation usually have a lasting long-term impact. Consider that the deployment options you choose must be simple for all. Still, your cloud platform must meet the corporate objectives of limiting unnecessary copies of data, security concerns, regulation and compliance, and overages on cloud computing costs.

Some vendors market a “cloud-first” or “cloud-native” strategy for their database technology stack. You can usually interpret that to mean a *cloud-only* development strategy. You would lose with such a vendor the ability to handle workloads that need to run on-premises and other options outlined below.

Deployment methods include the following:

Vendor’s Cloud – Many Software-as-a-Service (SaaS) database vendors deploy in a vendor-managed cloud, often using public cloud technology. The database vendors, not your corporation, negotiate with public cloud vendors on your behalf to govern, support, and secure your data. Your company pays one invoice to the cloud database vendor, and they pay the cloud vendor. Data is now under the database vendor’s control. Convenience is an advantage, but lock-in is a disadvantage.

Cloud Vendor Analytics – A minor variant to Vendor’s Cloud is when you use a SaaS service from a public cloud vendor. Companies like Amazon, Google, and Microsoft offer both the infrastructure and the database. They govern it and control security. Again, billing convenience is an advantage, but vendor lock-in is even greater. You aren’t just limited to cloud deployments, you’re limited to only one cloud deployment.

About Vendor Lock-in

Your requirements for a cloud database are not likely to stay the same forever. Therefore, understanding just how much it will take to change vendors or cloud platforms is imperative. Ask your cloud vendor about egress policies and procedures before finalizing your choice.

Company-Controlled Public cloud(s) – Many companies who buy databases have set up virtual private clouds within the public cloud infrastructure of prominent vendors like AWS, Azure, Alibaba, and Google. Admins implement VPNs, access control, security groups, centralized logging of data access, and other security features that specific industries need. This is often called Platform-as-a-Service (PaaS) and some database vendors can deploy this way. With PaaS, IT has more control over the management of data and costs. IT can negotiate with the major cloud vendors for the best price and even take advantage of spot pricing when the workloads are temporary. Your team has better control over the type of instances (vCPU/Memory) you need to complete the analytical workload.

Hybrid On-Premises, Cloud – There are reasons to keep some of your data on-premises or leverage public online data stores for analytics. When loading all your data into the clouds is inconvenient, or impossible due to compliance with some law or regulation, or simply too costly, having the capability to analyze data anywhere it sits is a promising solution. The sheer amount of existing on-premises data, may make it impractical to move it to cloud. Admins can simplify an on-premises/on-cloud hybrid approach even more by making local storage look like cloud object storage with applications like Minio or specialized high-performance object storage compatible hardware like Pure Storage FlashBlade technology, or HPE Scality RING. These give data center hardware similar characteristics to cloud object-stores, making the two environments similar. This fills a data governance need when certain data must remain on-premises due to regulations or corporate mandates.

Why Is This Important?

Considering your deployment options allows you to decide between the simplicity of SaaS, the flexibility of PaaS, or the necessity of hybrid. Saas vendors require you to load all data into a specific place in one particular cloud and one format, theirs. This locks the customer into one solution, not allowing the freedom to move to a different cloud easily, or take advantage of lower-cost computing when available. Some vendors offer no solution for on-premises analytics. if you need to maintain on-premises analytics, or pull workloads back on-premises due to changing regulations, SaaS deployment won't work well.

Workloads may call for a low latency environment. For example, if you have big data spread across 10s or 100s of nodes, you might consider non-standard cloud compute nodes with faster connections. Amazon EC2 Instances can vary from "Up to 10 Gbps" (and they do not guarantee that speed) up to a guaranteed 100 Gbps of networking throughput. As you might imagine, there could be considerable benchmark differences in analytical performance when selecting the same vCPU/Memory but low network bandwidth. Choosing the right type of network is crucial for maintaining performance. Some vendors offer no such choice.

Finally, you should carefully consider *the next big thing* for deployment infrastructures. If a new technology comes along to make analytics even more efficient, consider the costs to egress data from a locked cloud to the new solution. The deployment methods you choose now impact time to value, scalability, flexibility, and security for years to come.

Storage and Workload

Cloud databases have variations in the way that they manage data storage. Before understanding your storage and workload choice, it's essential to understand the interaction between a data warehouse and data lake in many large corporations.

Traditionally, data warehouses are used when ACID compliance, predictability, and reproducibility are required. Data warehouses offer carefully managed metadata, which are held to standards set by the data governance team. They are often augmented by data lakes, which give up some of those ACID characteristics for scalability and low cost. Data lakes offer the opportunity to use low-cost storage, like Hadoop or object stores like Amazon S3, to perform analytics at very large scale. Metadata is more loosely managed in a data lake, and the data within might be less structured than data in the data warehouse.

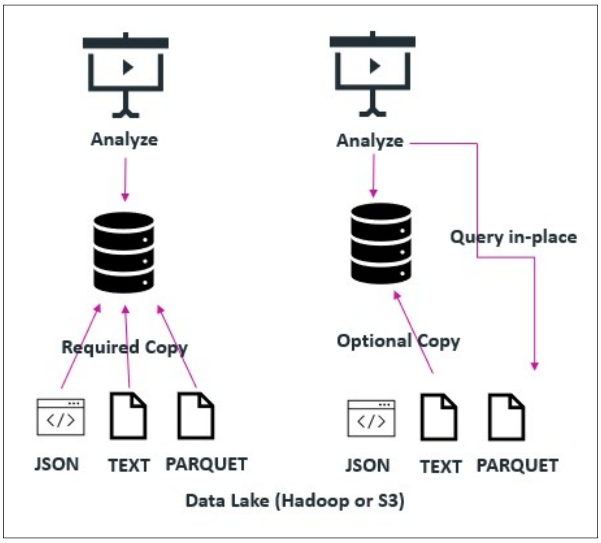


Figure 1: Does your cloud database require all data to be loaded into the databases, or can it access data directly in object store, Hadoop and other data lake repositories?

The above description of data lakes and data warehouses is accurate, except when your cloud-only database supports only *one* of these methods. Some cloud analytics solutions require that all data be loaded into a specific structured format in the data warehouse and offer no data lakes, sometimes with the exception of a completely vendor-specific technology stack.

Since all corporate analytical workloads are unique, you may find that the process of storing all data in one database type is flawed. The data science team likely has different expectations for performance than, say, users of an executive dashboard. IT teams often want to set up a tiering architecture, which means paying a premium when the service level agreement (SLA) requires fast response, and getting a discount when the workload has a less critical SLA.

Data can exist in many file systems and in many formats. Commonly, they include:

- File Systems: Apache Hadoop HDFS, Cloud or On-premises Object Store, Unix NFS
- File Formats: Text, CSV, JSON, ORC, Parquet, and so on.

Why Is This important?

When choosing your cloud database vendor, explore how it can bring together a data lake's scale and economics with the predictability and reproducibility of a data warehouse. Consider how well your solution understands the difference between workloads. Can it keep your hot data hot in optimized database-native formats and your cold data cold in other formats like JSON and Parquet? Or is the only option to store all data in the database format?

Optimizations

Cloud database capabilities diverge significantly when users are trying to resolve slow-running queries. It is common for cloud-only databases to offer only node-based optimization. The theory is that if your queries are running too slow to meet the organization's needs, simply add more nodes. The technology trend favors the availability of cheap computing, handling performance by spinning up additional compute power.

However, in most companies, analytical workloads are not universal. Performance of your database may be impacted by quarterly reports, a particularly successful marketing campaign that causes more data to be generated, experimental or poorly written queries that involve a lot of JOINS, to name just a few. As you make your cloud database selection, understand what options are available for speeding up individual queries. Some examples might include:

- **Massively parallel architecture** – Administrators should understand how the massively parallel architecture works in your database. Some databases are built on a transactional backbone and may have poor scaling for analytics. Some of these architectures may require manual sharding and special query modification to leverage the cluster.
- **Node scaling** – When under stress, can administrators scale nodes at will to handle unusually large workloads? In modern databases, scaling of nodes is commonplace. Does your IT team control node size or configuration?
- **Workload Management** – In a highly concurrent workload environment where hundreds or thousands of queries are running at once, does your software allow you to map query resources like memory and CPU to power specific query types (or sets of users)? For example, can you ensure that the CEO's dashboard won't be disrupted by rogue queries, or ad-hoc data science? Does the database solve this challenge with workload isolation features that can assign users or workloads a certain amount of resources, thus preventing interference with other workloads?
- **Separation of Compute/Storage** – Another technique for shared data in a concurrent environment is to use the separation of compute and storage. Data is kept in object storage while groups of compute nodes spin up to serve concurrency, backup, dashboards, and data science. You might use a lot at end of quarter, and none over a holiday. Separating compute from storage is the only way you can get cloud elasticity of scaling compute up and down as needed, while storage stays constant.
- **Query optimization** – Most SQL databases have query planners that parse SQL queries and attempt to answer them with the lowest stress on the system. Query planners figure out the best way to limit the data reads and memory needed to answer the query, but they don't always get it right. You'll find that

access to query optimization varies significantly, with some solutions offering virtually no access. You can only add nodes to improve query speeds. Other solutions provide more in-depth options.

- **Projections/materialized views** – Query planners may look for copies of the data sorted and optimally stored for certain queries. For example, suppose a report is looking for a certain region to answer a query. In that case, presorting the database by region eliminates the need to read the entire database when answering a query and therefore greatly speeds analytics.

Why Is This important?

All analytics are not the same, nor should every analysis be considered the same. Make sure that the database you select has options to properly manage all types of workloads and service level expectations. If only node-based optimization is offered, you may be missing out on methods to keep your cloud costs low while improving query performance.

Depth of Analytics

Cloud databases vary a great deal when it comes to the depth of analytics. In the past, IT teams and a relatively small number of business analysts might have been content with traditional SQL, but today's data-centric companies have questions that reach beyond the capabilities of standard SQL. Consider all groups in your organization who may be leveraging data in your cloud database, and how deep their different questions may go.

	Business Users	**Analysts**	**Data Scientists**
Analysis	Business metrics, pipeline	In-depth reports, patterns	Hidden opportunities, predictions, recommendations
Tools	Tableau and other visualization tools	SQL, Tableau and other visualization tools	R, Python, Jupyter/Zeppelin notebooks
Expectations	< 3 seconds response Dashboard always works	BI tools integration Sub-Minute Response Expanded SQL capability Geospatial analysis Time-series analysis Slow query remediation	Data Wrangling Raw data Access Scalability ML tools integration or ML algorithms built in

To serve this wide range of users, you must consider the depth of analytical functions that is offered in your cloud database solution.

These types of analysis include:

- **Time Series** – Features where SQL functions are built into the database to handle the Internet of Things (IoT) and log data recorded over set intervals of time. For example, consider a use case where a device is scheduled to transmit an update every 5 minutes. Due to a power outage, the device stops sending updates for several hours and later resumes. Does your database have gap-filling and interpolation features to handle the interruption, or will you have to deal with the gaps manually? Time-series capabilities vary a lot among cloud databases.

- **Geospatial** – Features where SQL functions are based on lat/long and elevation. Can you import shapefiles? Can you use your cloud database to calculate whether a home is inside or outside a flood zone? Can you build fair sales territories with your data? Can you perform optimal route analysis when managing a fleet of trucks? Geospatial features vary between cloud databases. Some charge extra licensing costs to use geospatial.
- **Machine Learning** - These days, analysts need predictive analytics, too. Some cloud databases offer the capability to train, manage, and deploy machine learning models. Some license machine learning separately, or charge per algorithm. You'll need these capabilities to perform pattern matching, regression, clustering, and others. Look carefully at the cloud database's capabilities if you want to leverage deep learning, too.
- **Alternate Frameworks** – Data scientist often bring their own experiences and tools, so it is crucial to understand if you can offer data access while supporting languages like R and Python, or interfaces like Jupyter notebooks. Look carefully at the world of data science outside of SQL and how this aligns with your cloud database's capabilities.

Why Is This Important?

Be ready to support a wide range of analytical use cases and a wider team of professionals as your cloud database grows increasingly popular in your organization.

Data Governance

Business and IT Control

While the appeal of cloud-only databases is that anyone, anywhere, at any time, can spin up a database and perform analytics, the simplicity of this process has also found many corporations in a bind over data regulations and data analysis cost. If business users set up an instance against a corporate account, they must be responsible enough to manage the following:

- **Access Control** – Traditionally, the IT team manages access to data and is responsible for it. Controlled access to personal information – Tax ID numbers and credit card numbers, for example – is an essential requirement under GDPR. Companies should be cautious that a SaaS analytics system does not remove centralized access control under IT.
- **Encryption and Security** – To help mitigate the risks associated with sensitive data processing, many regulations require encrypting your data. Understand whether your solution can encrypt data and analyze it in its encrypted form. Often referred to as Format-Preserving Encryption (FPE), this feature protects companies from regulatory penalties, even if data breaches occur. It doesn't require decryption for analytics to run. Encryption in motion is also essential to cloud operation. If your data is in a public cloud, then it needs to be encrypted even when it moves from one place to another, to avoid data breaches.
- **Cost Control** – Ensure that your cloud solution allows users to spin down compute when not in use. There are countless stories of budget overruns in costs on the cloud. In a report titled "The AWS bill heard around the world," one AWS customer described how an object store used for storing weblogs, which typically cost about $30 per month, suddenly went to $2700 one month. It took weeks to sort out what happened, as the crew considered denial of service attacks and other possibilities.

- **Managing Copies of Data** - Business users and/or IT must be responsible enough to limit multiple copies of the same data to reduce costs and protect data security. Those in charge of budgets for cloud services may not be aware of waste until the bill arrives, and it can be a shocker. Shared storage allows multiple teams to use the same data without making copies.
- **Auto-Scaling Costs** – Teams must understand how the solution will auto-scale on long-running or complicated queries, or when many concurrent workloads hit the system at once. If the database spins up extra nodes automatically, those additional nodes are commonly billed in monthly segments. Even if you use additional nodes for one hour, your solution provider may bill you for the entire month.
- **Egress Cost Control** – Understand the costs of moving data out of your SaaS-provider's cloud. Many providers will levy an egress fee, charged per megabyte, on data taken out of their platform and put into another one. Be wary of putting your data in any platform that's going to charge you to get it back out.

Why Is This Important?

Data governance features in a cloud database should help you adhere to industry and government rules around managing your data. By following the proper data management rules, you should be able to close the door to excessive fines and embarrassing data breaches. Ensure that you aren't making unnecessary copies of your data, as duplicates open you up to security, GDPR, and other regulatory risks. When your data warehouse is encrypted, hackers have a more challenging time taking advantage of its contents, and your company is exposed to far less risk. Finally, consider the choices you're making now, and how they will impact the future when you'll need to move data out of a cloud onto the next big thing.

What Should You Do

This white paper has covered a wide range of cloud database capabilities and what factors to weigh when considering a cloud database solution to make your company successful. Your decision must account for many important aspects, including managing and controlling costs, meeting SLA's, and the costs of moving data in and out of the database. The decision is also about ensuring that the cloud database offers the most people the widest range of analytical features and therefore offers flexibility and competitive advantage to the organization.