Choosing the Right Data Warehouse-as-a-Service for Your Analytical Needs

Prepared by:
William McKnight

MCKNIGHT
CONSULTING GROUP

McKnight Consulting Group
www.mcknightcg.com
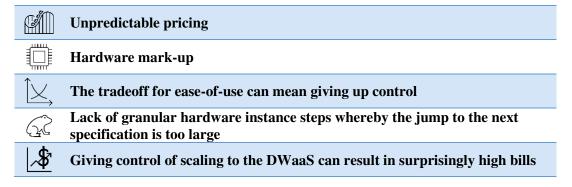September 2021

Contents

Overview

Software-as-a-Service (SaaS) is a licensing model where a customer subscribes to software usage in the cloud. All activity occurs through connection to that cloud. Pioneered by salesforce.com who famously used a strikethough on the word "software" as a logo for many years, SaaS offers unparalleled convenience over Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) models. SaaS alleviates the need to provision or manage the infrastructure and provides a fully configured solution to get started immediately.

In this paper, I'll explore the application of SaaS to a very important enterprise domain – the data warehouse. While many are labeling the data warehouse market as mature, it turns out very few data warehouses in production are truly mature. Most could turn focused attention into profit. One key aspect is in the licensing model where many are considering SaaS.

Many enterprises have multiple databases labeled "data warehouse" (and some have none). I will simply be referencing any large (according to the enterprise-scale factor) analytical databases as the data warehouse(s). Data Warehouse-as-a-Service (DWaaS) is a well-worn model for data warehousing, yet many flavors of DWaaS remain and it's important to get into the correct one for the workload. DWaaS is an automated data warehouse that self-tunes, patches, and upgrades itself without downtime. There is frequently additional functionality in the areas of caching and performance. DWaaS compute and storage shapes can scale up and down without delay.

DWaaS offers fast time-to-value, reduced administration costs, and frequently better performance than a self-managed on-premises or cloud-hosted data warehouse. The cost model is operational, pay-as-you-go without upfront costs. It enables you to leverage modern infrastructure platforms that are up-to-date with the latest performance, security, and operational features. This is, of course, in addition to all the usual benefits of a cloud data warehouse, such as scalability, elasticity, ease of chargeback, and rapid provisioning.

However not all DWaaS are created equal. There are clearly differences in the underlying database as well as differences in making the data warehouse a service in the DWaaS model. Mismatched data warehouse to DWaaS can result in:

	Unpredictable pricing
	Hardware mark-up
	The tradeoff for ease-of-use can mean giving up control
	Lack of granular hardware instance steps whereby the jump to the next specification is too large
	Giving control of scaling to the DWaaS can result in surprisingly high bills

First, I'll examine the different flavors of DWaaS to ensure you get into the right one. Then we'll look at some of the key criteria we should consider when reviewing the cloud database.

Flavors of DWaaS

Behind the covers of the DWaaS term, there are three distinct approaches. While all include most of the benefits I cited above for DWaaS, the differences mean that the benefits will accrue quite differently according to the fit of the model to the enterprise. These are vast enough differences to actually be the deciding factor in the DWaaS selection.

Bring Your Own License (BYOL) [1]

BYOL is a licensing approach that lets enterprises use their existing licenses for a new data warehouse in the clouds. This is a cloud data warehouse, but not technically a DWaaS, since the database is self-managed by the enterprise.

The cloud license belongs to the enterprise. The overhead of managing multiple licenses is removed and upscaling the database across the enterprise is easier. BYOL also enables an enterprise to continue any enterprise reporting around license use. Hybrid use cases are more easily implemented when cloud and on-premises software is identical.

Most enterprises these days have licenses and negotiated discounts with AWS, Azure, and Google Cloud. Part of the negotiation is often a commitment by the enterprise to use at least a certain amount of cloud hardware over a period of time. One advantage of the BYOL model is that it leverages the existing client-negotiated cloud hardware discount, and the data warehouse software uses some of that discounted, but often pre-committed spend.

BYOL Strengths

+	**Enterprise discount on hardware/cloud infrastructure**
+	**Data warehouse software retires committed cloud spend**
+	**More flexibility and control than any other option**

BYOL Weaknesses

−	**Need to ensure license can be utilized for new purpose - some are restrictive to server type, etc.**
−	**Management, scaling, setup, upgrades, etc. all done by customer**
−	**Full control requires administration**
−	**Customer provisioning required**
−	**More work and staff dependence**

---

[1] BYOL can also be used with on-premises.

Vendor-Provisioned Cloud

A vendor provisioned cloud DWaaS runs in a vendor-provisioned cloud account on behalf of the enterprise. This is a fully-managed service in which the database vendor essentially serves as the hardware reseller and you run in their provisioned cloud account, usually on AWS, Azure, or Google Cloud.

When you request an account, you choose the region where the account is located. It should be near your users or there may be a cost consideration.

This model may be single or multi-tenant. Be sure to find out when making your selection. Multi-tenant implementations often store metadata for multiple customers in a single database, creating a security issue. There is also the performance challenge that since infrastructure is shared, one customer may use a great deal of resources, reducing the available resources for other clients. This "noisy neighbor effect" means your performance may suffer even when you are not overusing the database.

One thing to be wary of is autoscaling with no safeguards. Autoscaling allows you to take advantage of cloud elasticity, pay only for what you use, and have just as much compute power as needed for workloads. It is one of the greatest advantages of a SaaS model. However, autoscaling means that as your enterprise's users use the data warehouse more, you pay more. Without safeguards, this can lead to unexpectedly high bills that were not in the budget. Since vendors resell hardware, the vendor has no financial incentive to reduce node usage.

Vendor-Provisioned Cloud Strengths

⊕	**No administration needed for licensing through database management**
⊕	**Vendor can have access to regions that an enterprise doesn't**
⊕	**Greatest ease of use of all the options**
⊕	**Autoscaling as workloads increase and decrease**

Vendor-Provisioned Cloud Weaknesses

⊖	**Lack of control – stair-stepped compute options**
⊖	**Only autoscaling of nodes to improve performance - black box so no way to identify chokepoints/optimize queries or reallocate resources**
⊖	**Pay markup prices on hardware; possible vendor lock-in**
⊖	**Greatest ease of use can mean least control (i.e., corporate security standards not utilized)**
⊖	**Possible surprise high bills**

Customer Cloud Account

A customer-provisioned cloud DWaaS is a fully managed service that runs in the customer-provided cloud account, allowing customers to maintain their negotiated hardware cost, as well as their implementation policies. This model is always single tenant, allowing the customer to set security levels, and eliminating any metadata sharing or noisy neighbor performance issues.

Similar to BYOL, this model leverages the existing client-negotiated cloud hardware discount, and the data warehouse software retires some of that often pre-committed spend.

Similar to a vendor-provisioned SaaS, the vendor handles administration – provisioning, autoscaling, upgrades, etc. Also, similar to vendor-provisioned SaaS models, autoscaling provides excellent elasticity, but requires safeguards to prevent unexpectedly high bills.

Customer Cloud Account Strengths

⊕	**Customers retain negotiation power**
⊕	**More control than vendor-provisioned SaaS**
⊕	**Uses corporate standards**
⊕	**Better TCO – no markup on hardware**
⊕	**More ease of use than self-managed BYOL**
⊕	**Autoscaling as workloads increase and decrease**

Customer Cloud Account Weaknesses

⊖	**Requires enterprise to provide their own cloud instance**
⊖	**Possible surprisingly high bills from autoscaling**

Data Warehouse Differences

There could be many databases that fit an enterprise's chosen provisioning model. Whether it comes before or after the provisioning decision, eventually the database itself is a decision. There are important factors of this decision. These are beyond the "table stakes" capabilities that are fairly common across all platforms and don't serve anymore as particularly useful metrics for differentiating solutions for the modern workload. Then, there are the critical capabilities to consider.

Performance is a critical point of interest when it comes to selecting an analytics platform because it ultimately impacts total cost of ownership, value, and user satisfaction. Price/performance for an enterprise's workloads is a chief consideration.

Data warehouses are found in most organizations and some are undergoing vast reengineering. Recent features have been added to many databases that modernize the data warehouse for the advanced analytics like geospatial, event pattern matching, and machine learning workloads required today.

Another factor in the decision is the SQL itself and how machine learning algorithms would be included with the solution. The SQL needs to have extensions in the areas of machine learning and other advanced analytics. Several are partnering currently for AI capabilities. It will be important to look at how integrated those partnerships are with the core product.

The promise of the cloud is immediate access to needed resources. A data warehouse needs to be able to proactively scale and be non-impactful for migration, repartitioning, upgrades, etc. It should be able to analyze many disparate data formats, seamlessly with the rest of the database, with or without ingesting the data first. Finally, a data warehouse should have the capability to combine with a cloud-storage based data lake and extend the reach of queries into the cloud storage and data lake data storage formats, as necessary.

By evaluating these factors, we gain insight into the value each DWaaS can have to an organization.

Vertica and Vertica Accelerator

Vertica is one software code base offered in several convenient models, on-premises, on Hadoop, on all major clouds as BYOL, and now Vertica Accelerator, which is a customer cloud DWaaS currently offered on AWS. Vertica has the same capabilities, regardless of deployment model: SaaS, cloud BYOL, Hadoop, VM, containerized on Kubernetes, bare metal, edge, or laptop.

Vertica is built on Massively Parallel Processing (MPP) and columnar-based architecture for excellent performance even at extremely high scale. Vertica SQL has more than 650 built-in analytics functions including geospatial, text analytics, time series, event pattern matching, and machine learning. Open source interfaces such as VerticaPy offer the ability to access these capabilities with Python and Jupyter notebooks.

Vertica can analyze ORC and Parquet complex types like arrays, maps, and structs, and export Vertica data to Parquet. It also supports streaming message data, and semi-structured data formats like JSON, Avro, CSV, and log or sensor messages for IoT. Vertica can also fully encrypt data, and has security authorization down to the cell level. It integrates with Kafka, Spark, Python, R, and a wide variety of visualization tools.

Vertica is available as a BYOL data warehouse on all three major clouds (AWS, MS Azure, and Google Cloud) as well as Alibaba Cloud, and it runs on-premises. Vertica in Enterprise Mode configuration works with compute and storage linked on the clouds, on commodity hardware, and on Hadoop. Vertica in Eon Mode configuration separates compute and storage for elastic compute both on the clouds and on premises on HDFS or MinIO, and on Pure Storage FlashBlade, Dell EMC ECS, Scality Ring with HPE Apollo, and other object storage compatible hardware.

You can run a Vertica license on any of the major clouds by installing the software on the procured instances.

Vertica Accelerator offers the mature functionality of Vertica in Eon Mode in a managed service on the customer's cloud account. Currently, Vertica Accelerator is only available on AWS, so enterprises would need an AWS account to use it.

There is no metadata sharing with Vertica Accelerator as it is single-tenant. All your data stays on your instance. Since your own cloud account is used, there is no markup on hardware (compute or storage).

Vertica's data compression rates are very high, so it uses less storage space for the same data. Since Vertica has on-premises and BYOL in its history, optimized compute hardware utilization is also a fundamental functionality. This means that Vertica Accelerator accomplishes the goals of performance at scale often with fewer nodes, helping to control cloud spending. Since Vertica does not resell the cloud hardware, there is no financial incentive to use more.

While Vertica Accelerator allows enterprises to let all their data scientists and analysts just use the software, not manage it, Vertica Accelerator is not a black box. Query chokepoint optimization, user-defined caching, workload isolation for better concurrency, and other optimizations are available options for users who want to have ways to improve performance beyond paying for more nodes. Vertica Accelerator also includes autoscaling for elastic compute, but gives the enterprise the option of setting user-defined guardrails to avoid unexpectedly high cloud bills.

Summary

Data warehouses are a prominent focus for investment in the enterprise today. DWaaS is emerging as having the best distribution of responsibilities for most data warehouse workloads.

By running in a customer-provisioned cloud account, all data and compute resources are on your cloud account. The vendor can directly communicate with the database, providing tuning, monitoring, and upgrades. The customer keeps control over data and can fully apply a security and compliance model.

Vertica is a high-performing columnar data warehouse with a large number of built-in functions, strong compression, and integration with data lakes. Vertica Accelerator is currently available as a DWaaS in AWS.

About McKnight Consulting Group

With a client list that is the "A list" of complex, sustainable and successful information management, McKnight Consulting Group (MCG) has broad information management market touchpoints. Our advice is an infusion of the latest best practices culled from recent, personal experience. It is practical, not theoretical.

We anticipate our customer's needs well into the future with our full lifecycle approach.

Our focused, experienced teams generate efficient, economic, timely and politically sustainable results for our clients.

- We take a keen focus on business justification.

- We take a program, not a project, approach.

- We believe in a model of blending with client staff and we take a focus on knowledge transfer.

- We engineer client workforces and processes to carry forward.

- We're vendor neutral so you can rest assured that our advice is completely client oriented.

- We know, define, judge and promote best practices.

- We have encountered and overcome most conceivable information management challenges.

- We ensure business results are delivered early and often.

MCG services span strategy, implementation, and training for turning information into the asset it needs to be for your organization. We strategize, design and deploy in the disciplines of Big Data, Data Warehousing, Analytic Databases, Master Data Management, Artificial Intelligence, APIs, MLOPs, Cognitive Search and Business Intelligence.

www.mcknightcg.com