Persistent Myths of Cloud Data Warehouses

WHITE PAPER

early adopter research

Actian™

CONTENTS

early adopter research

INTRODUCTION

It is possible to grow a cloud data warehouse instance to meet the needs of storage or compute-hungry workloads, and then reduce capacity when it is no longer needed.

Even though cloud data warehouses form a crucial part of the IT landscape, they are still misunderstood in important ways. Most notably, not all cloud data warehouses are the same: different cloud data warehouse platforms offer unique features and tools. Companies need to ensure they select the right cloud data warehouse for their current needs and future requirements.

To assist companies in finding the right solution, this white paper first explores the motivations for moving to a cloud data warehouse. It then delves into the most common myths about cloud data warehouses.

Motivations for Moving to a Cloud Data Warehouse

Although cloud adoption is a complex phenomenon, there are generally three main reasons companies move to a cloud data warehouse:

- **Elastic pricing models.** Cloud data warehouses allow users to pay for what they use by the hour rather than having to build an infrastructure to handle peaks.

- **On-demand scalability.** It is possible to grow a cloud data warehouse instance to meet the needs of storage or compute-hungry workloads, and then reduce capacity when it is no longer needed.

- **Fully managed offering.** Cloud data warehouses are assumed to be fully managed offerings in which data center, operations, networking, compute and storage infrastructure is built and managed by the vendor. Many of the administrative and operational details of the cloud data warehouse software may also be handled by the vendor.

These factors dramatically change the economics of using a data warehouse and also increase ease of use and the potential for self service as well.

The cloud can help companies cut costs – but this is highly dependent on the platform and workload composition. In actuality, cloud data warehouse pricing models vary widely.

Underlying all three of these motivations is a desire to cut costs. The right cloud data warehouse choice can lead to significant savings. But while motivations for cloud migration are common across businesses, the features of cloud data warehouse platforms differ significantly – and these differences ultimately contribute to significant variations in whether or not cloud platforms cut costs and also support desired business outcomes. Customers must be informed about these differences to make intelligent choices when selecting a cloud data warehouse vendor.

To help understand these differences and ensure they select the right solution, companies should recognize the following myths about cloud data warehouses (CDW).

Myth 1: All Cloud Data Warehouses Cost the Same

Companies often assume that by moving to a cloud data warehouse, they will immediately see their compute and storage costs decrease. As mentioned above, the cloud can help companies cut costs – but this is highly dependent on the platform and workload composition. In actuality, CDW pricing models vary widely. If a business doesn't understand its usage profile, it can result in expensive mistakes.

Costs of Scaling and Continuity

Scaling usage can be costly and involve steep pricing tiers for higher scalability.

Some platforms, like Snowflake, charge more if the total users exceed a certain number because additional warehouses must be added to provide that capacity. For Snowflake, if more than eight users want to work concurrently, the CDW will need a new warehouse to support each additional group of 1-8 users.

Other CDWs, like Amazon Redshift, pressure businesses to pay for use 24/7, even if they only run analytics from 9 to 5. That's because Redshift is impractical to shut down and restart. It requires a snapshot to be taken, and then if you wish to continue using Redshift, you need to restore the snapshot, and the time required is tied to the size of the data. It's more cost-effective to align usage with need at all times, rather than to pay a blanket fee.

Resource Unit Variation

Ultimate costs are affected by how CDWs charge enterprises. Many CDWs are now being sold in terms of resource units. These compute resource units are not created equal, and this is not obvious at the outset. Pricing may be similar but performance often varies significantly. These types of speed and performance differences matter and must be taken into account when determining the total cost to support a workload in a CDW.

Deep Engineering

Design and engineering of the CDW are also crucial factors that will impact costs. Actian Avalanche uses vector technology to achieve cost-effective scalability. Platforms that do not leverage vectorization typically process data one tuple at a time – which is both time and resource intensive. A tuple is a single record or row in a relational database. Vectorized processing operates on thousands of tuples of data in a single CPU cycle. This leads to better performance so that workloads can be processed faster, with far fewer compute cycles, which can reduce the costs of a CDW.

Myth 2: All Cloud Data Warehouses Are Feature Equivalent

The second common myth is that CDWs are feature equivalent. But again, CDW technology has vastly different pedigrees. Some CDWs are based on older technology that was not designed for a cloud computing paradigm while others are based on newer technology that is not proven at scale. Companies need to carefully inspect the feature offerings of the platform to identify the right solution for their use case.

Vectorized Processing and Columnar Stores

Companies should look for platforms that are architected for today's technology, for example with vectorized processing and in-chip data processing like that offered by Actian with its Actian Avalanche technology. This technology enables users to maximize their use of computing resources.

Columnar store capabilities are not the same across CDWs. All columnar stores read data in columns, but some, like Redshift, then take the data read from the columns and store that data as a series of rows for further processing. This creates additional overhead and slows down performance. Actian Avalanche keeps the data blocks for each column read completely separate, reducing overhead which leads to maximum efficiency.

Companies should look for platforms that are architected for today's technology, for example with vectorized processing and in-chip data processing like that offered by Actian with its Actian Avalanche technology.

With platforms that use EBS such as Actian Avalanche, companies do not have to throttle up to a new cluster until they reach a far larger number of concurrent users.

With more advanced columnar store implementations like Avalanche, companies experience better performance and a simpler schema enabled by the auto-indexing of data as it is loaded into the database. This is another crucial difference compared to Redshift which requires that the DBA create and maintain indexes.

Legacy technology migrated to cloud

There are a number of legacy data warehouses that can run in the cloud but have gone through limited engineering changes. Such forklift migrations of on-premise architectures fail to exploit the power of the cloud.

Concurrency

Concurrency varies greatly among vendors as well. Companies should not assume that an unlimited number of users can be on the CDW without encountering any issues. Actian Avalanche's architecture is designed to handle a much higher volume of concurrent users compared to other solutions. This helps to keep costs down when deploying at production scale. For example, using the same compute resources, Snowflake handles eight concurrent users while Actian Avalanche allows up to 64 concurrent users out of the box. If more than eight users need to access Snowflake concurrently, it will either queue up those additional users until a concurrency slot become available, or it will spin up a new compute warehouse to facilitate those additional users. This auto-scaling in Snowflake can drive up costs rapidly.

Use of Storage

The way cloud storage is used by the warehouse also varies. Many platforms, such as Snowflake, store their indexed data in Amazon S3 because it is much cheaper than high-performing Amazon Elastic Block Store (EBS). However, with S3, the CDW has to rely on an expectant cache, which results in massive problems when users attempt to retrieve data when the cache size is exceeded or overrun. This is why some platforms have to throttle up to a new cluster when a certain number of users try to work concurrently. With platforms that use EBS such as Actian Avalanche, companies do not have to throttle up to a new cluster until they reach a far larger number of concurrent users.

Storage Speed

When it comes to scalability, storage speed also matters. It comes down to the format in which the CDW stores data. In many CDWs, data is stored and processed in different formats. Actian Avalanche stores data in the same vector format from CPU to storage. This makes it easier to support scalability.

Myth 3: All Cloud Data Warehouses Can Support a Hybrid Model

Companies often have a misconception that they must move all workloads completely to the cloud. Or once they go to the cloud, they can never go back.

This isn't true. Moving data warehouse workloads to the cloud is not an all-or-nothing proposition. For the foreseeable future, both for vendors and users of CDWs, the ability to support a hybrid model in which some CDW workloads run on-premise and others run in the cloud will provide a variety of advantages.

But there are many concerns that must be addressed when supporting a hybrid deployment, as outlined below.

Data Locality

When determining whether workloads should exist in the cloud or on-premise, consider data locality: namely, analyzing the data where it naturally resides. If a marketing campaign is being conducted in the cloud, for example, it makes sense to do the analysis in the cloud. Analyzing hundreds of on-premise ERP application logs is naturally an on-premise task.

In general, it is optimal to minimize data movement as ingress and egress from the cloud is expensive and slow. Processing and querying data where applications are producing it, whether on-premise or the cloud, often produces the best results in terms of cost and performance.

For the foreseeable future, both for vendors and users of CDWs, the ability to support a hybrid model in which some CDW workloads run on-premise and others run in the cloud will provide a variety of advantages.

Actian Avalanche is unique in its ability to have the same data warehouse service delivered on-premise or in multiple clouds such as AWS and Azure.

Compliance and Security

Hybrid models can also support situations where a company seeks to retain complete control over sensitive datasets, particularly those that are subject to regulatory compliance or that have stringent security requirements. Highly secure workloads can stay in the known environment while the cloud is used for other workloads. When security and compliance in the cloud has been certified, the workloads can then move.

Amortizing On-Premise Investments

Many companies have made a significant investment in on-premise infrastructure and want to make use of those resources and shift to the cloud gradually over time. The hybrid model allows workloads that are optimal for the cloud to move while those that can be handled on-premise run on infrastructure that has already been paid for and may be cheaper.

Phased, Non-Disruptive Migrations

Migrating away from legacy or on-premise applications is often a multi-year process. In some cases, certain groups or divisions may continue to need custom on-premise solutions for the foreseeable future. The hybrid model allows certain data warehouse workloads to stay on-premise until the workload is ready to move.

On-Prem and Multi-Cloud Scenarios

Actian Avalanche is unique in its ability to have the same data warehouse service delivered on-premise or in multiple clouds such as AWS and Azure. Actian Avalanche offers technical and commercial equivalency across these platforms, which makes decisions about deployments or migrations focused on business considerations and not on accommodating technology differences.

Complete Support for Hybrid Operation

Many companies assume their CDW will support hybrid operation and deployment. But this isn't the reality. For instance, some of the most notable CDWs, like Snowflake and Redshift, do not have options to deploy on-premise. They don't support federated queries. And for companies that want or need to continue to integrate legacy infrastructure, this poses a huge problem and can lead to substantial cost overruns.

With Actian Avalanche, the ability to handle hybrid deployment is built in. All that is additional in the cloud is cloud economics, centered on elasticity and dynamic scalability. Companies don't have to pay more for hybrid deployments.

Hybrid Deployments Need External Tables

One key feature required by hybrid deployments is support for data coming from external tables. In a hybrid deployment by definition, data is stored in multiple places. The goal, of course, is to have all data needed by workloads running at a certain location to be located there. But that goal is never reached, and it can be very convenient to be able to access a table external to the local data. For example, a query in the cloud may access a table in the on-premise instance and join that data with cloud-resident data.

Federated Queries

A platform that offers federated queries across on-premise and across multiple clouds is essential for optimal support of hybrid environments.

A platform that offers federated queries across on-premise and across multiple clouds is essential for optimal support of hybrid environments.

Myth 4: All Cloud Data Warehouses Reduce Risk of Lock-in

The final common myth about CDWs is that because they exist in the cloud, there's far less risk of lock-in than with on-premise infrastructure. But lock-in can be just as big a risk in the cloud if companies are not careful about the platform they select. For instance, Amazon Redshift is only available on AWS and Microsoft Azure SQL Data Warehouse is only available on Azure. Microsoft and Amazon lock companies into their clouds. Snowflake and Actian Avalanche are true multi-cloud offerings that help companies avoid the risk of lock-in.

Legacy vendors are replicating their proprietary codes, scripts, administration tools and other mechanisms to ensure installed base retention, limiting the ability of their customer base to take advantage of the economy of scale offered by the open ecosystem of the cloud. A side effect of this withholding is an increase in cost, decrease in flexibility and agility and, over time, a reduction in the ability to hire vibrant new staff concerned with working on open standard platforms to maintain skill set marketability.

Before you choose a cloud data warehouse, it is well worth considering the very real differences in cost, engineering, and architecture.

REALITY CHECKS

Now that we've explained the myths, here are some reality checks.

All cloud data warehouses do not cost the same. Because of its deep engineering, Actian Avalanche offers more high performance computing power per resource unit, processing thousands of tuples of data in a single CPU cycle. Avalanche supports a significantly higher number of concurrent users in a single warehouse. Together this means more work done at lower cost.

All cloud data warehouses are not feature equivalent. A deeper look shows that cloud data warehouses differ in their implementation of key features. Actian Avalanche offers vectorized processing and in-chip data processing, enabling you to maximize your use of computing resources. Avalanche is an advanced columnar store, offering all the benefits and speed of columnar processing for aggregations and calculations. Data is auto-indexed as it is loaded into Avalanche, with no effort from DBAs.

By using Amazon EBS instead of cheaper S3 buckets, Avalanche delivers higher performance and concurrency. Avalanche stores data in the same vector format from CPU to storage, with storage speed that supports scalability.

All cloud data warehouses cannot support a hybrid model. Hybrid architectures are needed for many important business reasons, including data locality, compliance and security, and non-disruptive migration. Avalanche supports all the requirements of a hybrid architecture, including external tables and federated queries.

All cloud data warehouses do not reduce risk of vendor lock-in. If a cloud data warehouse is tied to a cloud provider's platform, it cannot support a hybrid model that embraces multiple clouds as well as on-premise deployments. Only Actian Avalanche delivers the same data warehouse service on-premise or in multiple clouds such as AWS and Azure.

Before you choose a cloud data warehouse, it is well worth considering the very real differences in cost, engineering, and architecture.

This paper was written by Early Adopter Research and sponsored by Actian. Learn more about **Actian Avalanche** and **get started for free with a 30 day trial**.

Connect with us