LEARNING MADE EASY

Actian Special Edition

# Hybrid Cloud Data Warehouse Basics

for dummies®
A Wiley Brand

Compare data warehouse solutions

Achieve superior price-performance

Enable faster time to value

Brought to you by

ACTIAN™

Lawrence C. Miller

# Hybrid Cloud Data Warehouse Basics

Actian Special Edition

**by Lawrence C. Miller**

for dummies®
A Wiley Brand

# Hybrid Cloud Data Warehouse Basics For Dummies®, Actian Special Edition

Published by
**John Wiley & Sons, Inc.**
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN 978-1-119-70801-8 (pbk); ISBN 978-1-119-70802-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

## Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact info@dummies.biz or visit www.wiley.com/go/custompub. For details on licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Some of the people who helped bring this book to market include the following:

**Development Editor:** Amanda Cross

**Project Editor:** Martin V. Minner

**Executive Editor:** Steve Hayes

**Editorial Manager:** Rev Mengle

**Business Development Representative:** Karen Hattan

**Production Editor:** Umar Saleem

**Actian Contributors:** Pradeep Bhanot, Dahye Huh, Vikas Mathur, Emma McGrattan

# Table of Contents

# Introduction

There is no doubt that companies that leverage more data from a wider variety of sources achieve superior business outcomes and outpace competitors. Data drives knowledge about your customers' needs and behaviors. It can tell you where the hidden risks lie in your business. It helps you find market opportunities. It can also uncover potential landmines.

The insights you can gain come from data from increasing diverse sources that may be dispersed across multiple clouds or even on-premises. The fact is, despite the hype surrounding cloud computing, most organizations still have a large portion of their business-critical data on-premises. The rest is spread across a combination of clouds — storage locations on the Internet.

Therefore, it is essential to have a hybrid data infrastructure in place that can simply and rapidly analyze large sets of data across on-premises and across multiple clouds in real time without the need for data movement. To leverage all an organization's data assets wherever they are, a new approach — a hybrid cloud data warehouse — is required.

In this book, you learn how a hybrid cloud data warehouse enables a 360-degree view of the customer by enabling analysis across all the business data assets regardless of their location, whether cloud or on-premises.

## About This Book

*Hybrid Cloud Data Warehouse Basics For Dummies* consists of nine chapters that explore:

- How business data has changed over the past several years (Chapter 1)
- How the systems that store and manage business data have evolved to keep up with those changes (Chapter 2)
- The important characteristics to look for in a modern data warehouse (Chapter 3)
- The types of data warehouses that are available on the market today (Chapter 4)

- What the data warehouse needs to do to optimize its performance for the price (Chapter 5)
- Approaches to serving disparate purposes with the data warehouse solution (Chapter 6)
- Opportunities to reduce wasted time and money when setting up your data warehouse solution (Chapter 7)
- How players in different industries are using advanced data analytics to differentiate themselves (Chapter 8)
- Things to keep in mind as you shop for a data warehouse solution (Chapter 9)

## Icons Used in This Book

Throughout this book, I occasionally use special icons to call attention to important information. Here's what to expect.

REMEMBER

This icon points out information you should commit to your non-volatile memory, your gray matter, or your noggin — along with anniversaries and birthdays!

TECHNICAL STUFF

You won't find a map of the human genome here, but if you seek to attain the seventh level of NERD-vana, perk up! This icon explains the jargon beneath the jargon.

TIP

Tips are appreciated, never expected — and I sure hope you appreciate these tips. This icon points out useful nuggets of information.

WARNING

These alerts point out the stuff your mother warned you about (well, probably not), but they do offer practical advice to help you avoid potentially costly or frustrating mistakes.

## Beyond the Book

There's only so much I can cover in this short book, so if you find yourself at the end of this book, thinking, "Where can I learn more?" just go to `www.actian.com`.

IN THIS CHAPTER

- » Understanding why digital savvy is essential to your company
- » Exploring the sources of available data
- » Discovering how user demands on data are changing

# Chapter 1 The Evolution of Data in the Modern Enterprise

Today's enterprises recognize that data is a valuable asset. The tools and methodologies to fully realize the value of those data assets are reshaping how businesses use data for strategic advantage. This chapter explores some important trends in data evolution inside an enterprise.

## Using Hybrid Data in the Enterprise

Enterprises today must compete in the digital economy or risk becoming irrelevant. Just as the retail, transportation, and hospitality industries transformed around Amazon, Uber, and Airbnb, every industry can fundamentally change when a major player finds a technological advantage. Your business can't afford to wait until the change happens, or you might be left behind.

Modern enterprises are leveraging data from various of sources to disrupt and drive major shifts in business models. This data comes from a wide variety of sources, some legacy but several so new that companies hadn't even conceived of them a generation ago. The list of possible data sources continues to grow, but some common options include the following:

- **On-premises applications,** meaning business applications that you have installed on servers in your physical offices. The data you store in on-premises applications likely resides on your business's own servers.
- **Software-as-a-service (SaaS) applications on private or public clouds,** which are business systems that you log in to through a browser window. The data you store in a SaaS application resides in a server your company doesn't manage directly.
- **Mobile applications,** such as your company's mobile app, which can collect extremely personalized data about individual users.
- **Social media platforms,** such as Twitter, that enable one-to-one communication with customers and the chance to learn about the people your company serves.
- **Edge devices,** such as sensors on machines or equipment (for example, fleet tracking devices on trucks).

## Exploring the 4Vs

During the "big data era" of the early 2010s, a "3V" model became popular as a way to define the properties of data. The three Vs were *volume*, *variety*, and *velocity*. Since then, technologists have found it useful to describe a fourth aspect of big data, the *venue* in which it resides.

The following sections explore how big data has changed along each of these four dimensions in recent years.

### Volume

Since the advent of everything online — clickstream data, location data, social media data — the world has seen a sudden explosion in the data that modern enterprises can collect and operate on, ranging from the hundreds of terabytes to several petabytes.

### Variety

New sources like clickstream and log data have become common, allowing enterprises to track their customers in ever greater detail. Most of this new data is semi-structured, generated in XML or JSON formats. Unstructured data sources like video are also growing quickly, generating an even greater volume of data.

### Velocity

Velocity is about how quickly your company can collect, analyze, and act on your data. Modern velocity is approaching real time. Data sources like cell phones, wearables, and websites can feed your business data as the events are happening, so of course companies want to leverage that data to influence decisions and fuel campaigns. The capability to analyze streaming data in near-real time is enabling new use cases like cybersecurity, where abnormal behavior can raise flags and trigger an investigation as it happens.

### Venue

As the list of data sources earlier in this chapter indicates, the venue where your data resides is a hybrid. A shift to generating and storing data in the cloud is underway, as more data-producing, as well as data-consuming, apps move to the cloud. However, for most large enterprises, legacy applications are still a reality. In fact, more than 70 percent of all core data still resides on-premises.

REMEMBER

The cloud is just an Internet-enabled place to store data. If the storage is for your company only, it is called a "private cloud." If the storage is for many companies to share, it is called a "public cloud" or just "the cloud."

The need to access all data — from all on-premises and cloud sources — for running analytics means that a hybrid venue is a reality that enterprises must prepare for.

## Changing Data Consumption

Not that long ago, people simply accepted that computer software was difficult to use, even when the program did something that the user understood well. Users have changed their expectations, though, and now they demand intuitive user interfaces even for the most sophisticated of software products. In recent years, data consumption has become more "consumerized" in at least three major ways. The next three sections describe them.

## Consumerization of the enterprise

With SaaS and mobile apps, the expectations for how an application should behave in terms of ease of use, speed of access, and comprehensiveness of feature set have seen a paradigm shift. Today's enterprise decision-makers and analysts need:

- Tools to surface insights, rather than raw data
- Access to relevant data wherever they are, such as on their laptops, mobile phones, and even wearables
- The ability to ask questions of the data as needed to assist making particular decisions
- Speed — it isn't okay for a dashboard to take more than ten seconds to load

## The rise of AI and ML

Petabytes of data coming in real time from a variety of new sources make it an inhuman task to comprehensively find insights. That's where non-humans can help. Artificial intelligence (AI) and machine learning (ML) enable businesses to consume and act on huge amounts of data.

The recent growth of compute power, combined with mountains of data to feed and train the machine learning and deep learning models, has made it possible to leverage AI and ML as inference engines. *Compute power* means computer processing power to crunch numbers or perform any other operations.

TECHNICAL STUFF

*Inference engines* are tools to apply rules of logic to the data to try to infer new implications automatically.

## Data analytics for all

In modern enterprises, developing new insights takes less technical skill and relies more heavily on domain knowledge about the business, markets, and customers. Thanks to sophisticated visualization tools and more AI/ML techniques being incorporated into new data access tools, business users are catching up to data scientists and are able to make data-driven decisions.

IN THIS CHAPTER

- » Discovering the available data warehouse technology solutions
- » Exploring the strengths and weaknesses of each kind of solution
- » Using a hybrid solution to maximize strengths

# Chapter 2 Changing Technology to Serve Changing Data

As the data in the modern enterprise has exploded, the technologies to store, manage, and use that data have advanced to meet the needs. A *data warehouse* is a central repository of information aggregated from various business sources like those discussed in Chapter 1. Data warehouses are not new, but they are evolving, as shown in Figure 2-1. This chapter explores some important trends in data warehouse evolution.

## Uncovering the Drawbacks of Engineered Appliances

Appliance-based data warehouses evolved to provide performance for a large amount of data and have established themselves as reliable reporting systems. However, these solutions have their drawbacks when it comes to meeting today's business demands. The following sections discuss some common pitfalls in traditional data warehouses.

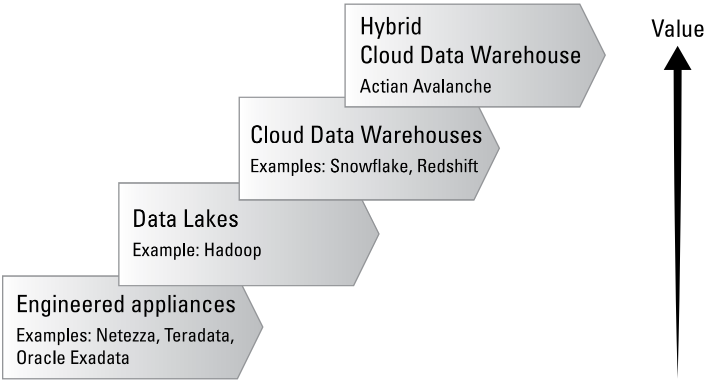


**FIGURE 2-1:** Data warehouse evolution.

## High cost to acquire, maintain, and upgrade

Appliance-based data warehouses rely on proprietary hardware to get their performance. The particular CPUs, memory, disks, and network interconnections that make up the appliance allow it to deliver consistent performance and reliability. Should any individual component fail, a certified technician typically must replace it with an exact replica. Because components fail over time, upgrades usually involve replacing the entire appliance and migrating the data to the new environment.

TECHNICAL STUFF

CPU stands for "central processing unit." A CPU is the part of a computer that executes the instructions that make up a software program. For general purposes, you can think of a CPU as a processor.

## Stale data

As the demand for organizations to operate in near-real time increases, data warehouses must be able to deliver fresh data. Traditional enterprise data warehouses work optimally with infrequent batch updates usually done at night when very few users are on the system. Lack of current data for operational reporting could lead to sub-optimal responses and hence reduce the ability of businesses to stay competitive.

## Data silos

As enterprises adopt new applications and services that generate new data, “puddles” of data are created that are disconnected from the main data flow. IT backlogs delay connecting these new systems to the enterprise data warehouse, which leads to data silos.

## Limited security and privacy protection

Stringent data security and privacy regulations, as well as the increasing frequency and severity of data breaches, have made security and compliance a top mandate for companies. The European Union’s (EU) General Data Protection Regulation (GDPR) is a recent example of a regulation that has major implications for organizations interacting with EU citizens.

TIP

Effective since May 2018, the GDPR puts restrictions on companies that do business in the EU and European Economic Area with the intent to protect the privacy of the EU citizens.

Some traditional data warehouses can lack advanced encryption features for data in flight (data traveling across a network) and data at rest (data waiting in storage). In addition, many data warehouses lack column-level dynamic data-masking capabilities to hide portions of sensitive data from users who don’t have adequate privileges, such as when you see the first several digits of a credit card number masked as Xs.

## Lack of elasticity

Appliances are designed for peak load with no elasticity to adjust to load variation during the day or the year, implying that a lot of capacity is underutilized most of the time. When the business must add capacity to accommodate growing data volume and increasing users, adding capacity becomes expensive because it means buying additional appliances.

Depending on the modularity and scalability designed into the product, an appliance-based solution may be highly proprietary. This creates vendor lock-in with customers having little choice other than to buy expensive add-on options or completely replace the appliance with an upgraded one.

## Exploring Data Lakes

Data lakes provide a low-cost environment to store massive amounts of raw data in its native format. Unlike a data warehouse, which is typically populated with structured and filtered data, the data in a data lake can be semi-structured or unstructured and isn't filtered until it is needed.

The following sections cover some of the pros and cons of data lakes.

### Open source but not free

To keep costs low, data lakes typically utilize open source software utilities. Apache Hadoop-based infrastructure or cloud-based storage solutions are common choices because they utilize commodity servers, disks, and operating systems. However, these solutions have other hidden costs.

Just like data warehouses, data lakes may require specialized skills and vendor or cloud-provider lock-in. Hadoop introduces the complexity of implementing a solution with many components, including clustered hardware and software, which requires integration skills and significant administration.

### Long development cycles

Before you can incorporate data from the data lake into your prepared data ecosystem for analysis, you must discover, prepare, and cleanse the data. Each of these steps requires specialized tools and skills. An organization that depends on a data lake but lacks people with these skills may find it has lost opportunities and is at increased risk of missing valuable insights.

### Low performance

While data lakes became cheap storage for raw data, they could not live up to the high performance expectations of typical data warehouse users. Hence, data lakes could not replace data warehouses and ended up creating a new silo.

# Examining Cloud Data Warehouses

Cloud data warehouses came into being with the intent to deliver analytics without requiring investment in expensive infrastructure or the skills required to install, configure, and maintain a complex technology. Some cloud data warehouses can even scale dynamically to meet the evolving needs of the business.

The following sections cover some of their drawbacks.

## Data silos

As mentioned previously, upwards of 70 percent of enterprise data is generated and remains on-premises. Cloud-only data warehouses don't have a solution today to incorporate on-premises data sources, so cloud and on-premises data end up siloed away from each other. Organizations have no choice but to select a separate analytics solution for their on-premises data, which means a bifurcation of skills, applications, and processes.

## Concurrency costs

The democratization of business intelligence analytics requires that all business decision-makers have access to the data warehouse. Cloud data warehouses typically use one of two possible approaches to deal with this high demand:

- Put the users in a queue to await their turn.
- Spin up additional compute clusters to handle the load.

The first approach can frustrate users and cost your business lost opportunities, while the second approach affects your budget.

## Cost containment challenge

Because cloud data warehouse environments may be scaled dynamically to meet business needs, predicting costs with any degree of certainty is impossible. Couple this with an architecture that launches additional compute instances to meet performance or user demands, and the budget required often can be two or three times the original estimate.

# Introducing Hybrid Cloud Data Warehouses

Hybrid cloud data warehouses deliver analytics across cloud and on-premises environments without the need for extensive data movement. Hybrid solutions enable the enterprise to make a journey to the cloud at a pace that makes sense for the business. Data that naturally belongs on-premises can remain there, dynamic data that is best suited to the cloud can live there, and the query capability executes queries across these warehouses as though they were a single data warehouse.

The following sections delve into the specifics of how a hybrid solution addresses problems.

## Break data silos

While appliances have trouble incorporating cloud data and cloud solutions have trouble processing on-premises data, the hybrid cloud data warehouse enables you to process both using the same technology as well as join that on-premises data with cloud data when needed. This solution treats all the data as though it is part of the same warehouse, with minimal data movement.

## Data gravity

Over 50 percent of companies have their cloud business systems operating in separate clouds — some by choice, many by necessity. Although the cloud data warehouse providers are building solutions to address this reality, only hybrid solutions can currently incorporate data from a multi-cloud infrastructure. No matter where data gravity dictates the data lands, the hybrid cloud data warehouse can analyze it where it naturally lives.

## Honor regulatory requirements

Some government regulations require that data is subject to very specific requirements as to how it is collected, stored, and processed. A hybrid solution provides the flexibility to comply with those requirements while seamlessly querying across on-premises and cloud data sets.

IN THIS CHAPTER

- **Reviewing the characteristics of a good hybrid data warehouse**
- **Querying disparate data sources and data types**
- **Taking advantage of the flexibility of the cloud**

# Chapter 3 Understanding the Characteristics of a Modern Hybrid Cloud Data Warehouse

Chapter 2 establishes some of the strengths of using a hybrid data warehouse rather than one that is purely appliance-based or cloud-based. This chapter illustrates how industry best practices are defining the technical requirements for hybrid data warehouses and lists the capabilities you can expect to find in a modern, high-performance hybrid cloud data warehouse.

## Hybrid Deployment Options

Almost all modern businesses have some data on-premises and some in the cloud, often in more than one cloud (AWS, Azure, or GCP are the most common choices). In general, you want to manage these siloes using the same technology and query across the environments when needed.

You want to process and query data wherever the applications are generating it so you can get the best results in terms of cost and performance. If your marketing team is conducting a campaign in the cloud, for example, doing the analysis in the cloud makes sense. By the same token, analyzing hundreds of on-premises logs from your enterprise resource planning (ERP) application is naturally an on-premises task.

## Federated data access

Unfortunately, queries that span cloud and on-premises data warehouses often have significant performance issues. Moving raw data from an on-premises system to a cloud system or across multiple clouds prior to querying inherently introduces latency and cost.

But users have a low tolerance for data-processing delays.

Since multiple data warehouses are likely to be a part of companies' IT ecosystems for many years, these problems aren't going away on their own. Fortunately, several options can address the performance issues of queries across these siloes:

- **Merge multiple data warehouses into a single instance in the cloud.** This solution calls for exporting all your data out of your on-premises data warehouses and importing it into your cloud data warehouse. Although this seems like the most straightforward solution, it often is not cost-effective. Legacy infrastructure investments, migration costs, and business disruption when migrating to the cloud can add up.
- **Separate the queries and perform aggregation processing in the application layer.** Many small companies have used this approach, relying on either web services or client applications to combine data from distributed sources. Application infrastructure typically has less processing capacity and speed than database infrastructure. Hence, this approach rarely results in the performance gains you want.
- **Perform intelligent federated query.** Intelligent federated query lets you create a query in one place, then the system distributes and executes the query across on-premises and cloud instances of the data warehouses. This approach does not require data movement across sources. Intelligent federated query automatically runs computes in the right place and delivers results to wherever you initiated the query.

## Scaling Storage Independent of Compute

Historically companies have kept compute power and storage in data warehouse systems tightly coupled to derive maximum performance. However, new storage techniques like column-based storage have made queries faster while processor caches and main memory (RAM) have increased in size. As a result, storage no longer limits the execution of queries. Therefore, decoupling storage and compute now makes sense to control costs while allowing each to scale up and down as business needs change.

TECHNICAL STUFF

A column-based database stores data columns together, as opposed to the traditional approach of storing rows together. Most analytical queries access few columns, hence columnar storage of data improves performance substantially.

## Scaling Compute Dynamically

One of the primary reasons companies select cloud infrastructure is its ability to scale up and down on demand to meet ever-changing business needs. Hybrid data warehouses are typically massively parallel processing (MPP) systems, which distribute the workload across several processors or nodes. The "elasticity" that cloud solutions provide offers the ability to add more nodes to the system to meet workload requirements and to scale those resources back when the work no longer requires them.

## High Performance at Scale

Although cloud elasticity provides significant economic benefits, it should not come at the cost of performance. Data warehouse users expect high performance and sub-second results of their queries to stay productive. A modern data warehouse needs to leverage software and hardware efficiently to provide the best experience to users at the lowest cost. In addition, performance must be demonstrated at scale, which comes in two forms:

- Volume of data being queried
- Number of users or applications concurrently querying the system

## Making Real-Time Updates without Penalty

One of the biggest challenges to delivering real-time insights is keeping the data up to date while the system is still serving the needs of the business. This challenge can be addressed with high-performance, in-memory positional delta trees (PDTs).

TECHNICAL STUFF

A *PDT* lets you update data warehouses with the latest data without slowing the performance of querying the data.

## Seamlessly Connecting to SaaS Applications

Integration with the business systems (for example, ERP applications) that generate the data is a critical component of the data warehouse workflow.

You must consider the cost, time, and effort to update and manage the web of connections that bring the data from operational applications to the data warehouse.

Modern data warehouses should make it quick and easy to move the data from diverse data sources, including software-as-a-service (SaaS) and on-premises applications, into the data warehouse with pre-built templates. The integration capability should have the intelligence to match, merge, enrich, and deduplicate data as it enters the system, as well as apply governance policies.

The best way to lower risk and contain overhead cost is to select a hybrid data warehouse platform that natively offers integration, so that the two technologies are always compatible and no extra personnel is required to make the function work.

IN THIS CHAPTER

- » **Reviewing the characteristics of an excellent hybrid data warehouse**
- » **Understanding how the characteristics translate to business results**
- » **Exploring how Actian Avalanche meets the criteria for excellence**

# Chapter 4 Comparing Data Warehouse Solutions

Chapters 1–3 establish the characteristics that make a hybrid cloud data warehouse powerful and essential to a modern business. This chapter brings that theory into reality with a concrete example.

In this chapter, you discover how Actian Avalanche provides increased value compared to modern alternatives as well as earlier incarnations in the data warehouse space.

## Comparing a Solution to the Checklist

This section demonstrates how to evaluate a solution provider against each of the characteristics of excellence by using Actian Avalanche, as an example. Figure 4-1 summarizes the architecture of Avalanche, which is a modern hybrid cloud warehouse.

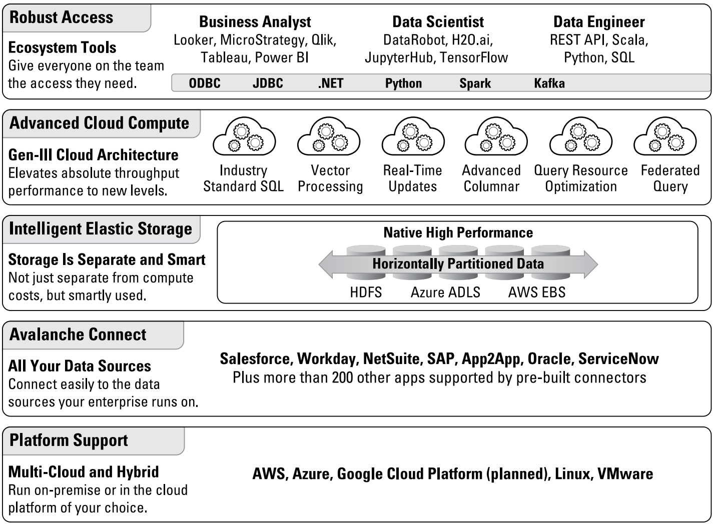

Robust Access
Ecosystem Tools
Give everyone on the team the access they need.
Business Analyst
Looker, MicroStrategy, Qlik, Tableau, Power BI
Data Scientist
DataRobot, H2O.ai, JupyterHub, TensorFlow
Data Engineer
REST API, Scala, Python, SQL
ODBC
JDBC
.NET
Python
Spark
Kafka
Advanced Cloud Compute
Gen-III Cloud Architecture
Elevates absolute throughput performance to new levels.
Industry Standard SQL
Vector Processing
Real-Time Updates
Advanced Columnar
Query Resource Optimization
Federated Query
Intelligent Elastic Storage
Storage Is Separate and Smart
Not just separate from compute costs, but smartly used.
Native High Performance
Horizontally Partitioned Data
HDFS
Azure ADLS
AWS EBS
Avalanche Connect
All Your Data Sources
Connect easily to the data sources your enterprise runs on.
Salesforce, Workday, NetSuite, SAP, App2App, Oracle, ServiceNow
Plus more than 200 other apps supported by pre-built connectors
Platform Support
Multi-Cloud and Hybrid
Run on-premise or in the cloud platform of your choice.
AWS, Azure, Google Cloud Platform (planned), Linux, VMware

**FIGURE 4-1:** Actian Avalanche brings a flexible architecture approach to create value for business.

## Modern and flexible deployment options

An excellent solution helps you future-proof your investment by avoiding vendor lock-in. The Avalanche hybrid cloud data warehouse is built for multi-cloud by design.

The Avalanche platform is available on Amazon Web Services (AWS), Microsoft Azure, and on-premises. All deployments, regardless of location, are compatible so you don't have to relearn the technology or alter your queries.

## Intelligent federated querying

Avalanche is implementing an intelligent federated query mechanism that will enable queries to run across multiple Avalanche instances (across regions, across clouds, or across cloud and on-premises). This mechanism enables businesses to retain their data where it needs to reside (for example, for privacy) and query all the data when needed.

Avalanche is building an intelligent query optimizer that will split the query processing across the data warehouses (see Chapter 7 for more details).

## Separation of storage and compute

Avalanche separates compute from storage so that you can scale these two components independently of each other. Pay only for what you use by switching off compute resources altogether when they are not needed.

Avalanche's automated storage indexes and advanced compression algorithms lead to the most optimized use of storage. This helps reduce your total cost of ownership.

## Dynamic scaling of compute

Avalanche has the built-in ability to scale compute up and down on demand by the user. You no longer have to buy for peak load. Instead, you can scale up when needed and scale down after the peak. This is a great capability for businesses in the retail industry who experience peaks during holidays or the back-to-school season.

## Scalable performance at low price

Avalanche takes advantage of performance features in today's CPUs that other cloud data warehouses and relational databases do not use. As a result, Avalanche can process data much faster and deliver analytic workload results quicker than most other relational databases.

Much faster data processing performance opens opportunities for more iterations during model tuning by data scientists and more "what if" scenario runs for teams of business analysts trying to determine the best business decisions in real time. Actian Avalanche delivers support for large data sets (hundreds of terabytes), enterprise wide users, and complex workloads. It also provides high performance out of the box without the need for extensive tuning by automatically creating min-max storage indexes.

TECHNICAL STUFF

Think of *min-max indexes* as catalog numbers in a library. They are markers that tell the database the minimum and maximum values stored in a column file. This allows the system to directly read a small subset of file blocks to process a query, thus increasing performance significantly. This matters a lot when you are dealing with billions of rows of data in a table.

Avalanche uses vector processing to operate on hundreds of pieces of data at a time and keeps the data compressed even in cache, delivering higher CPU utilization and significantly faster performance.

TECHNICAL STUFF

Whereas traditional data warehouses operate on a single data item, vector processing operates on an array of data.

Avalanche's efficient query processing enables it to handle a large number of concurrent users and high data volumes without needing to spin up significant additional compute infrastructure. This results in significant cost savings when compared to other solutions in the market, which are resource hungry and spin up new clusters to handle concurrency.

## Direct application integration with only a few clicks

Avalanche is the only hybrid cloud data warehouse to provide a comprehensive set of connectors to software-as-a-service (SaaS) and on-premises applications built into the product. It provides more than 200 pre-built enterprise connectors to popular data sources such as ServiceNow, Salesforce, and SAP. You can set up the connection with a few clicks and, more importantly, the integrations run on a schedule or whenever a triggering event occurs.

## Democratized data analysis

Avalanche works with all analytical applications to support the needs of users with a wide range of technical skills. Each person, including business analysts, data scientists, and data engineers, can interact with Avalanche using familiar tools and languages that your organization has already invested in. Hundreds of users can access the data and perform their own queries at the same time without experiencing performance degradation.

**IN THIS CHAPTER**

- **Understanding advances in processor technology**
- **Using best practice techniques to maximize the use of data warehouse resources and lower costs**

# Chapter 5
# Achieving Superior Price-Performance

Cloud-only data warehouses are designed for the abundant, boundless resources offered by the cloud. So often, they do not make the most efficient use of resources. Why make the effort when you can simply spin up a new cluster when in need of extra compute power? One reason is because each new cluster costs more money to use!

Hybrid cloud data warehouses, however, must also consider the scarce resources of an on-premises deployment where there will be a fixed and limited amount of CPU, RAM, and disk available so they must maximize their utilization of these resources. This efficiency directly translates into 10x better performance and significant cost savings since one only pays for the resources one uses in the cloud. Customers can choose how they want to balance performance and budget to meet the organization's needs.

In this chapter, you learn how advances in central processing unit (CPU, or "processor") technology and industry best practices enable a superior price-performance ratio.

## Exploiting the CPU

The central processing unit (CPU) is the computer that does the work in your data warehouse — it's what provides your compute power. Over the past three decades, CPU capacity has roughly followed Moore's Law. However, improvements in CPU data processing performance are not only the result of the number of transistors on the chip. CPU manufacturers have introduced additional performance features, such as multi-core CPUs and multithreading, which most database software leverages transparently to improve the system's performance as if it had several CPUs working at the same time.

REMEMBER

*Moore's Law* describes a long-term trend in which the number of transistors that an integrated circuit (IC) can accommodate doubles approximately every two years. Although Moore's Law specifically refers to the number of transistors, it is casually used to describe technology improvements in general, which double performance every two years.

However, other CPU optimizations that have been introduced in the last decade are not typically leveraged transparently by most database software today. Some examples, discussed in the following sections, include:

- Vectorized processing and single instruction, multiple data (SIMD) instructions
- CPU cache as execution memory

WARNING

Most database software today is based upon technology developed in the 1970s and 1980s. This database software has become so complex that a complete rewrite of the software would be required to take advantage of modern performance features.

### Vectorized processing and SIMD instructions

At the CPU level, traditional databases are scalar, meaning they process data one tuple at a time using a single instruction, single data (SISD) model. A *tuple* is a single row in a relational database. The CPU spends most of its time managing tuples rather than on processing.

In contrast, single instruction, multiple data (SIMD) enables a single operation to be performed on a set of data, called a *vector*, at once. A database that can do this uses "vectorized" processing, as shown in Figure 5-1.

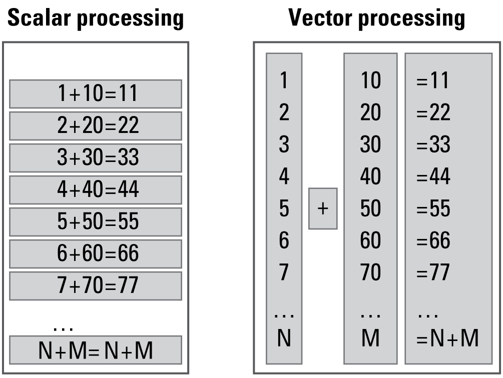


**FIGURE 5-1:** Scalar and vector processing operations.

REMEMBER

In traditional scalar processing, operations are performed on one data element at a time. In vector processing, operations can act on multiple sets of data at a time. In the example in Figure 5-1, the values from the two columns are being added. In the scalar processing case, each of these individual additions is a separate operation. In the vector processing case, the two sets of values are added together in a single operation.

Actian Avalanche uses vectorized processing to dramatically increase its processing speed — in the following ways:

- It processes vectors consisting of multiple tuples (hundreds or thousands) of data elements all at once.
- It executes basic computation operations, such as the addition in Figure 4-1, as branch-free loops over arrays of column values stored in a cache.
- Operations leverage SIMD to perform a single instruction on multiple column elements at a time.

TIP

Because typical data analysis queries process large volumes of data, the use of vectorized processing maximizes the efficiency of the CPU and greatly speeds computations.

## CPU cache as execution memory

When most traditional databases entered the market in the 1980s, high-end computers had only roughly 8 MB of memory (RAM, also known as "main memory"). Databases were designed to optimize the movement of data between disk and memory.

Most of the improvements to database server memory (RAM) over the past several years have resulted in larger memory pools, but not necessarily faster access to memory. As a result, relative to the CPU's ever-increasing speed, access to memory has become slower over time. In addition, with more CPU cores requiring access to the shared memory pool, contention can be a bottleneck for data processing performance.

Today, CPUs typically have 8 MB or more of memory built into the processor itself as level-1 (L1) and level-2 (L2) cache, which processes, such as a modern database server, can access much faster than shared memory (RAM).

TECHNICAL STUFF

L1 cache memory is usually built directly onto the CPU itself, whereas L2 cache memory is usually built onto a separate chip, such as an expansion card.

To achieve maximum data processing performance, a high performing hybrid cloud data warehouse avoids using shared RAM as execution memory. Instead, it uses the CPU core and CPU caches as execution memory, thereby delivering significantly faster data processing throughput.

Figure 5-2 shows the relative performance characteristics of disk, memory, and chip cache:

- **Disk:** 40MB to 1 gigabyte (GB) per second, depending on whether the system uses hard disk drives (HDDs, or "spinning disks") or solid-state drives (SSDs)
- **Memory:** 20GB to 50GB per second
- **Chip Cache:** 100GB to 200GB per second

The cycle numbers show the access latency for each layer (disk, memory, and cache) in clock cycles per second.

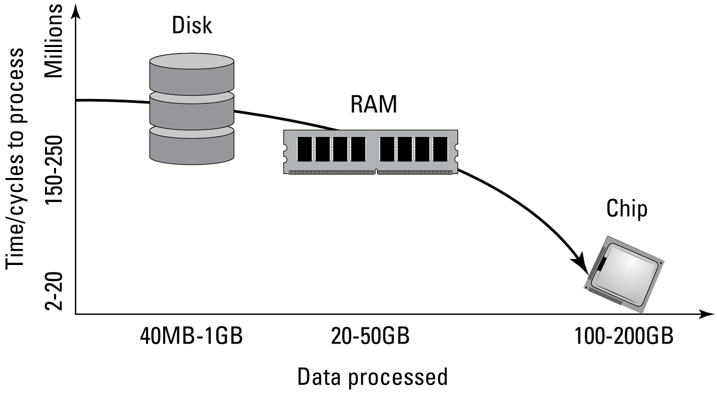


**FIGURE 5-2:** Data processed "on chip" is exponentially faster than data processed in memory and on disk.

TECHNICAL STUFF

"Access latency" is the time delay that occurs between sending a data request (for example, to disk) and fully receiving the requested data.

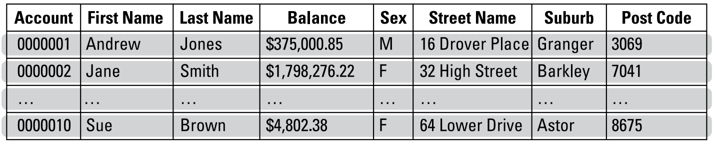
CPU speed is measured in clock cycles. A CPU with a speed of 4Ghz can execute 4 billion operations in a second where an operation could be fetching an instruction, accessing memory, or writing data.

# Leveraging Industry Best Practices

Specialized data warehouse products use many well-known techniques to achieve fast performance. In general, because of the data-intensive nature of a data warehousing workload, most techniques focus on limiting and optimizing input/output (I/O).

## Column-based storage

Early relational database software implemented "row-based" storage, in which all data values for a row are stored together in a data block (or "page"). Data is always retrieved row-by-row, even if a query needs only a subset of the columns (see Figure 5-3).



| Account | First Name | Last Name | Balance | Sex | Street Name | Suburb | Post Code |
|---|---|---|---|---|---|---|---|
| 0000001 | Andrew | Jones | $375,000.85 | M | 16 Drover Place | Granger | 3069 |
| 0000002 | Jane | Smith | $1,798,276.22 | F | 32 High Street | Barkley | 7041 |
| … | … | … | … | … | … | … | … |
| 0000010 | Sue | Brown | $4,802.38 | F | 64 Lower Drive | Astor | 8675 |

**FIGURE 5-3:** Row-based storage.

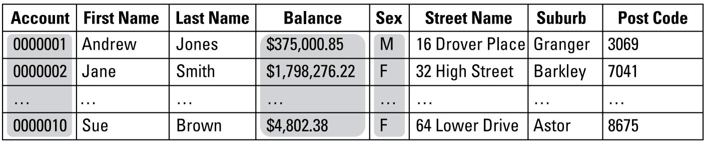
Row-based storage works well for online transaction processing (OLTP) systems, such as order entry systems or retail sales systems, in which

- Stored data is highly normalized, so tables are relatively narrow (they have a small number of columns).
- Queries typically retrieve relatively few rows
- A small number of queries involving large volumes of data occur

In contrast, data warehouses have different characteristics:

- Tables are often partially denormalized, meaning data might be grouped inconsistently or repeated, resulting in many more columns per table, not all of which are accessed by most operations
- Most queries retrieve many rows
- Periodic or ongoing, controlled streams of data, rather than ad hoc processes, add large data sets to the data warehouse

As a result of these differences, a row-based storage model typically generates a lot of unnecessary I/O for a data warehouse workload. A column-based storage model, in which data is stored together in data blocks on a column-by-column basis, is generally a better storage model for data analysis queries (see Figure 5-4).



| Account | First Name | Last Name | Balance | Sex | Street Name | Suburb | Post Code |
|---|---|---|---|---|---|---|---|
| 0000001 | Andrew | Jones | $375,000.85 | M | 16 Drover Place | Granger | 3069 |
| 0000002 | Jane | Smith | $1,798,276.22 | F | 32 High Street | Barkley | 7041 |
| … | … | … | … | … | … | … | … |
| 0000010 | Sue | Brown | $4,802.38 | F | 64 Lower Drive | Astor | 8675 |

**FIGURE 5-4:** Column-based storage.

In column-based storage, database queries read only the columns that are needed to solve the query. This saves I/O bandwidth, fits more relevant data into memory, and results in less "cache pollution" because only the relevant data is being written to cache.

TIP

In addition to the benefit of focusing on the required data when accessing fewer than all table columns in a query, column-based storage provides better data compression than row-based storage because all of the data in a column will be of the same type. For example, it will all be text, a date, or a numeric. With row storage, you have a mix of data types in a row, which makes finding an optimal compression algorithm difficult.

## Data compression

Most databases compress data before storing it so that it moves through the system more quickly and occupies less space on disk. Compression in column-oriented storage is more efficient than compression in row-oriented storage. In row-oriented storage compression, the choice of a single compression algorithm is difficult because a row typically has a mix of datatypes. An algorithm that works well for text, for example, may not work well for numeric data.

Column-oriented storage compression allows the algorithm to fit the data type of the column, as well as the data domain and range (even where the domain and range are not explicitly declared). Data compression happens on a column-by-column, page-by-page basis using an algorithm best suited to the data being compressed.

## Storage indexes

The data in database columns are written to disk as a series of files, and each file is made up of a number of blocks, known as *database blocks.* A high-performance database makes every effort to read only the database blocks that contain relevant data, since reading data from disk is slow. A storage index keeps track of the contents of every database block. For example, it might keep track of the minimum and maximum value in that block. If a query required data that was outside of the range of values in that particular database block, that database block would not be read when solving the query.

## Parallel execution

Almost all relational databases support some means for a single operation to take advantage of multiple CPU core resources. For some databases, particularly pure massively parallel processing (MPP) databases, the use of multiple CPU cores is mandatory and virtually every operation uses all CPU cores in the system. Other databases use some form of a shared architecture and therefore support a wider range of possible degrees of parallelism.

The ancient proverb that "many hands make light work" applies to query execution as well; the more helpers helping, the quicker they complete the task.

For example, imagine that a large financial services institution wants to calculate the average balance in its customers' savings accounts. Using a machine with eight processor cores, the company calculates the sum of all the values and then divides that result by the number of customer accounts. With parallel query execution, the company can divide the task across all eight cores and assign each core the task of summing one-eighth of the values. The coordinator then adds those eight results and divides the answer by the number of accounts. This process should finish in approximately one-eighth the time a single core would take to complete the same work serially.

A modern hybrid cloud data warehouse typically answers queries for many users simultaneously. For maximum performance, the data warehouse parallelizes each of these queries. The data warehouse balances the level of parallelization that makes sense when trying to serve a large number of users simultaneously; too many parallel tasks may overwhelm the system, and too few will mean the system isn't performing to its full potential.

IN THIS CHAPTER

- » Considering different kinds of processes during data warehouse design
- » Uncovering the costs and performance of splitting workloads
- » Smoothing out peaks in demand with resource pools

# Chapter 6 Handling Concurrent Workloads

Your data warehouse can manage data from many different sources, for many different departments, and with many different purposes. Some cloud data warehouses now suggest that the best way to handle the concurrency of these diverse factors is to split the individual workloads into their own separate compute warehouses. However, this approach brings its own drawbacks and expenses.

This chapter explores the pros and cons of splitting workloads across separate clusters to determine whether it is a true innovation or just an expensive workaround.

## Architecting for Concurrency

Most organizations aspire to provide all their business decision-makers with access to all the data required to make the best decisions for the organization. But the cost of providing everyone with access to a cloud data warehouse can be expensive and depends upon the architecture employed to achieve that concurrency.

REMEMBER

Hybrid cloud data warehouses, which were designed for resource efficiency, can expand to support growing numbers of users by adding new nodes into the resource pool. In a cloud environment, these nodes can be added and removed dynamically as the business needs change. This is also possible in an on-premises deployment, but typically once nodes are added to a cluster they are never surrendered back to the pool.

Cloud data warehouse solutions, designed for an abundance of resources, expand by adding whole new compute clusters to satisfy the needs of new groups of users — typically each group of up to eight users requires a separate compute cluster so the costs and the architecture can quickly spiral out of control.

## Pooling Resources

Resource pooling is, just like it sounds, the process of serving multiple purposes with the same compute resources. You might have already learned the reasoning behind the resource pooling concept in micro-economics or operations research class. Any system with variable resource utilization, like a data warehouse, can pool the resources together to balance out the peaks with troughs in the pattern of how it uses the resources.

TIP

Retail uses resource pooling extensively. Demand for various products at a particular store vary by location and by day. Having a distribution center to supply inventory to multiple stores in an area balances out spikes in demand for a product in one store versus lower demand for the product in another store. Resource pooling helps modern retailers avoid stockouts when demand for a product spikes in one store, while also reducing overall inventory, since each store doesn't have to plan for peaks individually.

## Demonstrating the Cost

You can see how resource pooling saves costs more clearly by looking at an example. The scenario in Table 6-1 considers the cost of a popular cloud data warehouse solution that splits the workload across separate clusters. (In the table, the total number of elapsed hours does not equal the sum of the workload hours because some of the workload hours overlap.)

TABLE 6-1 **Cost with Cloud Data Warehouse**

| Workload | Size and Number of Cloud Virtual Warehouses | Duration Active | Cost per Day |
|---|---|---|---|
| Nightly ETL | M (4) | 8 hours | $64 |
| Finance Canned Reports | L (8) | 2 hours | $32 |
| Finance Analysts (ad-hoc) | XL (16) | 8 hours | $256 |
| Marketing Canned Reports | L (8) | 4 hours | $64 |
| Marketing Analysts (ad-hoc) | XL (16) | 12 hours | $384 |
| Total | | 16 hours (actual elapsed) | $800 |

The same enterprise scenario can be accomplished with a single cluster in a hybrid cloud data warehouse utilizing resource pooling, as shown in Table 6-2. Though the direct users performing ad-hoc analysis do not have dedicated virtual warehouses, the cluster they share with the other functions is more powerful, so they see better performance on their queries. Query performance matters a lot to these users. They don't want to spend more than a few seconds waiting for a dashboard to render.

TABLE 6-2 **Cost with Avalanche**

| Workload | Avalanche AU | Duration Active | Cost per Day |
|---|---|---|---|
| Nightly ETL | 8 AU | 16 hours | $256 |
| Finance Canned Reports | | | |
| Finance Analysts (ad-hoc) | | | |
| Marketing Canned Reports | | | |
| Marketing Analysts (ad-hoc) | | | |
| Total | | 16 hours | $256 |

## Justifying the Workload Split

If pooling resources leads to better performance for less cost in most situations, then why do some cloud data warehouses encourage workload splitting?

Many cloud data warehouses on the market today are unable to handle concurrency. When your system works best when users use it for just one thing at a time, you always encourage your users to split the workload across clusters. Of course, you don't want to tell prospects that any time concurrency increases (which in enterprise scenarios can go as high as hundreds or thousands of users during peak time), your performance degrades significantly! However, if you look at the results from the 2018 MCG benchmark report comparing Actian with Competitor S, you can see that is exactly what happens. Competitor S's only way out to guarantee performance is to create separate clusters by workload.

Plus, it doesn't hurt that splitting workloads makes life easier for IT as it can easily bill different departments based on their compute consumption. Although this is a genuine benefit, most companies would rather save money than simplify billing. Besides, this feature is available on a modern cloud data warehouse like Avalanche as well. Without any special set up, you can group users from a department, compute each group's utilization cost, and report it to IT.

IN THIS CHAPTER

- » **Discussing the unexpected costs of copying all data to the cloud**
- » **Finding better ways to reduce expenses and effort**

# Chapter 7 Enabling Faster Time to Value With a Hybrid Approach

Historically, data warehouse and analytics technology vendors have designed solutions that push enterprises to move all their relevant data into the data warehouse solution. This chapter explores how you can save time, money, and hassle with a data warehouse solution that reflects how data is generated and consumed within your enterprise.

## Making the Data Do the Work

In on-premises data warehouse scenarios, various on-premises business systems generate transactional data and store it in OLTP databases. Traditionally, the data warehouse solutions required moving the data from the multiple OLTP systems to a central data warehouse on a nightly or weekly basis. Then, the analytics could run on that central data warehouse.

TECHNICAL STUFF

OLTP stands for "online transactional processing." It is the general term to describe a database that focuses on inserting, updating, and deleting data from a large number of transactions.

At the time, this approach was the only way to get analytical performance through use of engineered appliances tuned for analytics. Unfortunately, this approach also wasted significant time and energy in creating extensive extract, transfer, load (ETL) pipelines to move and transform data. Over the years, the pipelines have become extremely hard to maintain. As a result, businesses try not to let their requirements shift, which significantly hinders advances by business.

A similar scenario played out when data sizes started growing and data variety increased with data arriving in semi-structured formats like JSON and XML. The appliance data warehouses could not scale to handle these changes.

Hadoop with NoSQL came up to fill the vacuum with data lakes. Once again, businesses built extensive ETL pipelines to move data into these lakes, sometimes from data warehouses and at other times directly from source OLTP systems. Analytical performance on these systems never lived up to what was advertised, though, and enterprises now have hundreds of millions of dollars and hours sunk into the data lakes and ETL pipelines that are now aging.

More recently, cloud data warehouses have emerged. While cloud data warehouses provide benefits like elastic scalability, they are again forcing enterprises to move data into the cloud without much consideration of where the data producing applications reside or what it takes to create new ETL pipelines to the cloud.

## Forcing Everything to the Cloud

Although much attention has focused on cloud data warehouse solutions over the past several years, the reality is that most enterprises will need both cloud and on-premises solutions for the foreseeable future. Regulatory compliance requirements and data sovereignty requirements may prevent some data assets from *ever* moving to the cloud.

The "RightScale 2018 State of the Cloud Report" found that:

- Eighty-one percent of enterprises surveyed have a multi-cloud strategy today.
- Fifty-one percent of enterprises have a hybrid cloud strategy.
- Ninety-six percent of enterprises now use public (92 percent) and private (75 percent) cloud solutions.
- Organizations leverage five clouds on average.

Increasingly, enterprises are looking for ways to access data as if it were in a virtual data warehouse: requiring the compute power to go to the data where it naturally resides rather than making a copy of the data for the analysis to use.

However, many enterprise data warehouse solutions have limited deployment options. Appliance-based solutions, for example, can't be deployed to the public cloud. Virtualized versions of these appliance-based solutions may not perform optimally on commodity hardware in the cloud, and the database software may not be designed for the cloud.

## Optimizing Time to Value

Ultimately what matters is how quickly the enterprise can pick new and existing data sets, merge them, and start extracting insights from them. A best-fit approach molds the solution to the enterprise needs rather than requiring the enterprise to change its processes to fit what the solution can do.

TIP

Actian's hybrid approach lets your enterprise leave the data on-premises or in the appropriate cloud and works on it there. It allows an enterprise to move to the cloud at its own pace while getting insights from the data no matter where the data resides.

## Achieving a True Hybrid via Federated Query

Actian has taken a three-pronged approach to align its solution to the reality of where the data resides for enterprise customers:

- **Avalanche on Microsoft Azure or AWS:** The first prong of Actian's approach brings into Avalanche the new data sets that software-as-a-service (SaaS) applications (such as marketing data from Marketo or CRM data from Salesforce) are generating. It also addresses getting data from custom applications that have already been migrated to the cloud into Avalanche. This service is available across AWS and Microsoft Azure, with Google Cloud Platform (GCP) coming soon.

- **Avalanche on-premises:** The second prong deals with data sets that reside in OLTP systems or on-premises enterprise data warehouses. These data sets are easy to pull into Avalanche on-premises through Actian's connector technology or through the customer's favorite ETL/ELT or replication tool. Because the data that's already on-premises can stay on-premises, you do not have to worry about cloud security.
- **Hybrid federated query across all data:** Actian's approach will use federated query to virtually bring the different Avalanche instances together. This approach pushes query processing to wherever the data resides, thus eliminating unnecessary data movement.

WARNING

Federated query is not recommended in cases where large volumes of data from different platforms (for example, on-premises and cloud) are being joined in a query. The data movement required to satisfy the query will be slow and potentially expensive if the cloud provider charges for ingress and egress.

## Benefits of a Hybrid Solution

The hybrid solution is an innovation that saves you time and money while giving you superior performance:

- It doesn't force you to move all your data to the cloud, so you can migrate to the cloud at your own pace.
- Compute moves to where the data lives, so you don't have to build and support ETL pipelines, or on the security concerns that always come with moving data.
- It integrates with your existing tools so you can keep using the skill sets your team already has.
- All the benefits of cloud deployment, such as elastic and independent scaling of compute and storage, are available for the cloud components of the solution.
- You retain legacy on-premises applications as long as needed.

**IN THIS CHAPTER**

- » **Calculating the benefits of data analytics in financial services**
- » **Recognizing the value of analytics in retail**
- » **Taking care of data analytics in healthcare and clinical research**

# Chapter 8 Exploring Advanced Analytics in Different Fields

In this chapter, you explore several industry use cases for hybrid cloud data warehouses and learn about real-world Actian customer success stories.

## Empowering Quick Trades in Financial Services

Big data analytics is changing the world of capital markets and global banking. In the constant search for alpha (the active return on investment), firms deploy new analytics platforms. They are calculating enterprise credit and market risk in minutes instead of hours, achieving close to real-time transaction cost analysis (TCA), observing and anticipating fraud patterns in near-real time, and introducing data sets and techniques previously not possible.

TIP

*Alpha*, or active return on investment, measures how well an investment performs as compared to the market as a whole.

Big data is a winner's game. Leaders who embrace it learn to analyze massive data sets and leverage the insights to drive enterprise-wide revenue and efficiencies. Doing so with extreme speed, accuracy, compliance, security, and scalability sets them apart from the competition.

By transforming data into real-world business value with speed, efficiency, and transformational analytics, a hybrid cloud data warehouse can help global capital market firms:

- Consolidate silos of data across front, middle, and back offices to one central view
- Apply established, as well as new, scientific techniques to enable a new perspective on the markets to discover alpha or potential signals of risks to avoid
- Predict and prevent business-compromising events that violate regulations or corporate ethics
- Manage and control risk by identifying and auditing firm-wide anomalous behavior

## REFINITIV ACCELERATES FINANCIAL ANALYTICS WITH ACTIAN

Refinitiv financial and risk solutions deliver critical news, information, and analytics to the global financial community, enabling transactions and connecting communities of trading, investing, financial, and corporate professionals.

Refinitiv needed to meet a 20-millisecond response time requirement for its Eikon and Elektron data and trading applications. The company implemented an Actian Avalanche server farm consisting of more than 100 servers, each hosting multiple client accounts. Avalanche provides a hub for all the data used by the Eikon and Elektron applications and meets the service-level agreement requirements for complex queries with sub-20 ms response times.

## Knowing Your Customers in Retail

Retail use cases for data analytics leveraging a hybrid cloud data warehouse are as numerous as bargains and sales on Black Friday and Cyber Monday. This section covers just a few of the many examples.

### Customer profile

Granular, multi-channel, near-real-time customer profile analytics can tell you all about your customers, with characteristics like:

- The best means to connect
- The targeted offers that are most likely to resonate
- Their predilection to churn
- The best ways to personalize the entire customer experience to win more business and drive up loyalty levels

TIP

*Customer churn* is another way of saying "customer attrition" or losing a customer.

Valuable information comes from a growing number of sources, such as sales transactions, web usage, social media, mobile devices, purchase history, and service history.

### Micro-segmentation

Most companies doing segmentation use basic account information and demographics to find groups of customers based on high-level account and behavior metrics. However, you can go further and use micro-segmentation models to find and classify small clusters of similar customers, and you can use customer value models to predict the value of each customer to the business at various intervals.

Combining the output of these two models into a personalized recommendation engine gives you the information you need to take action that gives you a distinct competitive advantage. You can optimize your supply chain, customize campaigns with confidence, and ultimately drive meaningful, personalized engagements.

## Customer lifetime value

It is generally easier to sell to existing customers than to acquire new ones. You need to measure and maximize current and forecasted customer value across products, segments, and time periods to design new programs that accentuate your best customers and provide you with a distinct business advantage.

## Next best action

You can maximize long-term customer value not only by predicting what a customer will do next but influencing that action as well. If you want specifics about customer behavior and spending, you need all the data available to you, structured or unstructured, from traditional enterprise sources, social networks, customer service interactions, web clickstreams, and any other touchpoints.

REMEMBER

*Clickstream data* is a log of where a user clicked while using a website. Your own corporate site can track this kind of behavioral information.

## Campaign optimization

Traditional campaign optimization models use limited samples of transactional data, which can lead to incomplete customer views. A hybrid cloud data warehouse allows you to connect to social media and competitor websites in real time to learn which competitive offerings are gaining traction in the marketplace.

## Churn analysis

Churn prediction models have traditionally been limited to account information and transactional history, which represents a tiny fraction of the available data. A hybrid cloud data warehouse increases the accuracy of churn predictions by combining and analyzing traditional transactional and account data sets with call center text logs, past marketing and campaign response data, competitive offers, social media, and a host of other data sources.

## Market basket analysis

Market basket analysis models are typically limited to a small sample of historical receipt data aggregated to a level where potential impact and insights are lost. A hybrid cloud data warehouse brings in additional sources, in varying formats, enabling discovery of critical patterns, at any product level, to create a competitive advantage.

### ACTIAN MAKES RETAIL ANALYTICS FAST AND CONVENIENT FOR SHEETZ

Sheetz is a $5 billion convenience store business with a reputation for progressive marketing and fierce competitiveness in the marketplace. From day one, company executives recognized the value of having a finger on the pulse of what consumers want from a convenience store. As the business grew, advancing this knowledge became more of a challenge.

By deploying Actian Avalanche, Sheetz gained the ability to analyze a more comprehensive set of data (more than three billion rows), returning query results in seconds. It offered performance improvements of as much as 70 times over conventional technology by utilizing the latent processing power in the company's existing hardware infrastructure, with the added benefit of reduced operational costs. In addition, Actian Avalanche enabled Sheetz to double its access to historical data and be ready for expected growth over the next few years.

## Improving Patient Outcomes in Healthcare and Research

Healthcare innovators are paving the way to connected health, and analytics provide the catalyst. For providers and payors, healthcare technology infrastructures must evolve. They need the ability to connect seamlessly and instantly to electronic medical records (EMRs), electronic health records (EHRs), and healthcare information exchanges, as well as hundreds of disparate endpoints and data sources. Figure 8-1 shows an example of the flow of data and data analytics in a healthcare environment.

Using analytics, the healthcare industry is transforming data into actionable insights that improve healthcare delivery, reduce readmissions, promote preventive care, and accelerate research. Analytics also holds the key to reducing fraud, waste, and abuse in healthcare systems.

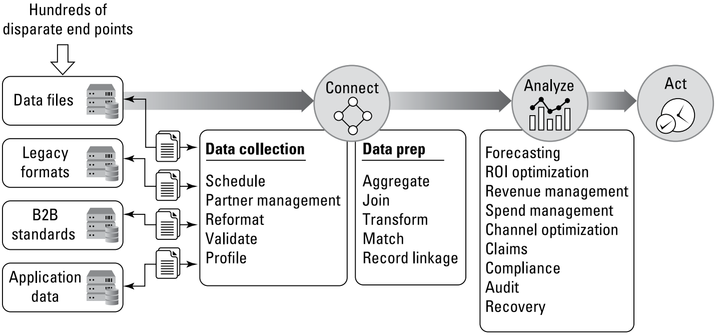


**FIGURE 8-1:** Data analytics for healthcare providers and payors.

## OXFORD UNIVERSITY ANALYZES MASSIVE DATA SETS IN MINUTES

Oxford University's CTSU (the MRC/Cancer Research UK/BHF Clinical Trial Service Unit and Epidemiological Studies Unit) primarily studies the causes and treatment of diseases such as cancer, heart attack, and stroke, which collectively account for most adult deaths worldwide. Researchers analyze vast data volumes to look for a needle in a haystack.

CTSU selected Actian to perform analyses of these massive data sets. Alan Young, Director of Information Science at CTSU, says: "Without Actian, we simply would not be able to process this information, without having to wait days or weeks for each output."

IN THIS CHAPTER

- Looking for speed and performance
- Integrating with tools you already use
- Protecting the security of your data

# Chapter 9
# Nine Things to Consider When Evaluating a Hybrid Cloud Data Warehouse

Keep these nine considerations in mind as you evaluate which hybrid cloud data warehouse is right for you:

- **Performing on commodity hardware:** A superior hybrid cloud data warehouse takes advantage of the performance features in standard servers to maximize performance.
- **Scaling storage independently from compute:** As data continues to grow, hybrid cloud data warehouses need to grow to cheaply ingest and store large data volumes while not requiring compute (and its cost) to scale at the same rate.
- **Lowering total cost of ownership (TCO):** One of the main reasons companies consider moving their data warehouses to the cloud is to reduce their TCO. However, cloud data

warehouses differ in cost savings due to architectural differences.

- **Working flexibly through hybrid deployment:** A hybrid cloud data warehouse should provide flexible platform deployment options — on-premises, in public clouds, and on private clouds. It should provide technical and commercial parity across these environments to make it easy for customers to move data and consumption across the platforms.
- **Unifying through federated query:** A true hybrid cloud data warehouse should be able to query data sitting on-premises as well as across multiple clouds through a single query mechanism.
- **Integrating with the tools you use:** A good hybrid cloud data warehouse provides open application programming interfaces (APIs) such as Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), and American National Standards Institute (ANSI) SQL to work with the query tools an organization might use.
- **Providing built-in connectors to ingest data:** The ability to ingest data at high speed is a critical hybrid cloud data warehouse requirement. If you cannot load your data in a reasonable time, the result is having to work with summary data or, worse, stale data. Ideally, your hybrid cloud data warehouse should come with pre-built connectors to on-premises and SaaS applications, databases, and ETL tools.
- **Ingesting data in real time:** If data updates happen continuously throughout the day in micro batches or streamed updates, then you can be sure you are working with the most current information for analytics-based decision making. Having fresh data is becoming a critical requirement to enable newer data science-driven use cases.
- **Offering full data security and privacy:** An effective hybrid cloud data warehouse needs to offer built-in support for enterprise firewalls, intrusion detection, SIEM logging, SOC-II compliance, authentication integration, key management, and security patching. In addition, at the data level, it should provide row- and column-level access controls, encryption for data at rest, as well as encryption for data in motion.

ACTIAN™

# Actian Avalanche Hybrid Cloud Data Warehouse

## Built for High-Performance Analytics

### Do More With Less

- Up to 20x price-performance advantage over alternative solutions
- Deploy and manage on-premises, AWS, Azure and Virtual Private Cloud
- Native integration to popular SaaS and legacy applications

**Visit actian.com/avalanche to learn more and start a free trial**

**JOIN OUR WEEKLY LIVE DEMO AND SEE THE BLAZING FAST ANALYTICS IN ACTION**

Avalanche
Hybrid Cloud Data Warehouse

# Enable analysis across your data assets

Valuable business insights emerge from an increasingly diverse range of data sources. It is essential to have a hybrid data infrastructure that can simply and rapidly analyze large sets of data located on-premises and across multiple clouds in real time. To leverage all your organization's data assets, a new approach — a hybrid cloud data warehouse — is required. This book shows you how this solution enables a 360-degree view of the customer by enabling analysis across all your data assets, regardless of their location.

## Inside...

- Explore the modern hybrid data warehouse
- Achieve high performance at scale
- Design for concurrent workloads
- Hybridize through federated query
- Apply industry best practices
- Explore advanced analytics use cases

ACTIAN™

**Lawrence C. Miller** has worked in information technology for more than 25 years. He has written almost 200 For Dummies books.

ISBN: 978-1-119-70801-8
Not For Resale

9 781119 708018

for dummies®
A Wiley Brand