LEARNING MADE EASY

Actian
Special 2nd Edition

Operational Data Warehouse Basics

for dummies®
A Wiley Brand

Move beyond traditional reporting

Democratize data and analytics

Drive business outcomes

Brought to you by:

ACTIAN™

Lawrence C. Miller
Emma McGrattan

About Actian

Actian, the hybrid data management, analytics, and integration company, delivers data as a competitive advantage to thousands of customers worldwide. Through the deployment of innovative hybrid data technologies and solutions, Actian ensures that business critical systems can transact and integrate at their very best – on premises, in the cloud, or both. Thousands of forward-thinking organizations around the globe trust Actian to help them solve the toughest data challenges to transform how they run their businesses, today and in the future. For more information, visit www.actian.com.

Operational Data Warehouse Basics

Actian Special 2nd Edition

by Lawrence C. Miller and Emma McGrattan

for dummies®
A Wiley Brand

Operational Data Warehouse Basics For Dummies®, Actian Special 2nd Edition

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

ISBN 978-1-119-87329-7 (pbk); ISBN 978-1-119-87330-3 (ebk)

Publisher's Acknowledgments

Project Manager: Dan Mersey

Senior Managing Editor: Rev Mengle

Managing Editor: Camille Graves

Senior Client Account Manager: Matt Cox

Content Refinement Specialist: Tamilmani Varadharaj

Introduction

In the first edition of this book, we outlined that forward-thinking companies, large and small, need to be more customer-focused — even customer-obsessed — to be successful in today's hypercompetitive market. As we move beyond the COVID-19 pandemic this becomes ever more important, with a more remote and fluid workforce, changes to supply chains, real-time insights to deal with uncertain market conditions, and a changing business climate.

Data drives knowledge about your customers' needs and behaviors, so you can tailor your messaging and offers to rise above the competition and win their business. This knowledge comes from an increasing variety of real-time sources including digital systems and a rapidly growing ecosystem of sensors, devices, and mobile applications that track those activities. However, in the last few years, data privacy and security laws and regulations, cybersecurity threats, and an ever-expanding definition of what constitutes personal data and how much of it is accessible (and by whom) make navigation and use of all customer- and citizen-related data a gold mine with land mines.

Also, the volume of data and variations in the speed at which multiple disparate sources of varying types are coming at us can be overwhelming, and the value of your data can decrease quickly over time. You must have a data analytics infrastructure in place to rapidly consume, curate, and process perishable information and influence when and how you run your operations to improve outcomes. To manage data in the moment, a new approach — an operational data warehouse — is required.

In this book, you learn how an operational data warehouse goes beyond reporting on historic, static data and instead operates with fresh, active data to drive specific business actions and outcomes — in the business moment.

About This Book

Operational Data Warehouse Basics For Dummies, Actian Special 2nd Edition consists of five chapters that explore:

- Important data analytics and data warehousing trends (Chapter 1)
- The limitations of existing data warehouse and data lake solutions (Chapter 2)
- Key operational data warehousing use cases (Chapter 3)
- The technical requirements and capabilities of an operational data warehouse (Chapter 4)
- Key considerations for evaluating whether an operational data warehouse is right for you (Chapter 5)

Foolish Assumptions

It's been said that most assumptions have outlived their uselessness, but this book assumes a few things nonetheless.

The main assumption is that you are someone who understands the value of data to your business. Perhaps you are a senior business or IT executive, such as a chief financial officer (CFO), chief marketing officer (CMO), chief analytics officer (CAO), chief information officer (CIO), or chief data officer (CDO). Perhaps you are a data engineer, architect, or scientist, an application developer, a database developer or administrator, a business or data analyst, or some other IT professional who routinely consumes analytical data. This book assumes that you understand some of the challenges of data analytics, but that you may not necessarily have a technical background. For that reason, technical terms and concepts are explained throughout this book and TLAs — three-letter acronyms — are spelled out!

If any of these assumptions describe you, then this book is for you. If none of these assumptions describe you, keep reading anyway. It's a great book and when you finish reading it, you'll know quite a few things about operational data warehouses.

Icons Used in This Book

Throughout this book, special icons 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 — so let's hope you appreciate these tips. This icon points out useful nuggets of information.

Beyond the Book

There's only so much a 48-page book can cover, so if you find yourself at the end of this book, thinking, "Where can I learn more?" just go to `www.actian.com`.

Where to Go from Here

With my apologies to Lewis Carroll, Alice, and the Cheshire Cat:

"Would you tell me, please, which way I ought to go from here?"

"That depends a good deal on where you want to get to," said the Cat — er, the Dummies Man.

"I don't much care where . . . ," said Alice.

"Then it doesn't matter which way you go!"

That's certainly true of this book which, like *Alice in Wonderland*, is also destined to become a timeless classic.

If you don't know where you're going, any chapter will get you there — but Chapter 1 might be a good place to start. However, if you see a particular topic that piques your interest, feel free to jump ahead to that chapter. Each chapter is written to stand on its own, so you can read this book in any order that suits you.

You won't get lost falling down the rabbit hole.

IN THIS CHAPTER

- Looking at current data analytics trends
- Addressing changing user requirements for data warehouses

Chapter 1
Recognizing Key Trends in Data Analytics and Data Warehousing

In this chapter, you explore some important trends in data analytics and data warehousing that are shaping the business data landscape.

Analyzing Data Analytics Trends

Enterprises today must transform their businesses to compete in the digital economy or risk becoming irrelevant in the same way that Amazon, Uber, and Airbnb have disrupted the retail, transportation, and hospitality industries.

Using data from increasingly digitally active consumers and, in turn, digital businesses, will drive new insights leading to higher customer satisfaction, increased revenues, lower costs, and greater profitability.

Making applications more productive using artificial intelligence and machine learning techniques requires more data — and more current data — to fuel the models that lead to predictive and prescriptive analytics. Here are a few trends to consider:

- **Data volumes are increasing.** There is no shortage of data today. Individual users, businesses, sensors, and devices are creating an exponentially growing amount of data every day.

  From 2020 to 2025, IDC forecasts new data creation to grow at a compound annual growth rate (CAGR) of 23 percent, resulting in approximately 175ZB of data creation by 2025. How much of it should be analyzed and by whom? And when? Businesses recognize the need to analyze data from all sources — whether it's public data or generated from devices, sensors, and business processes — to make informed decisions and take appropriate action.

- **Analytic workloads are moving to the cloud.** Leading analyst firms report that the cloud data warehouse market will be just shy of $15 billion in 2022 and grow by 51 percent in 2023 to reach $22.5 billion, surpassing the entire value of the global data warehouse market of $21.8 billion in 2019.

- **From democratization of data to analytics.** Everyone is talking about the democratization of data. Sure, data scientists have skills and insights that regular business users don't.

  However, what's really needed is the democratization of analytics at the speed of business. For all the talk of artificial intelligence (AI) and machine learning (ML), most business users are still struggling just to get fresh data from all the sources they need, analyzing it and exposing their output in existing visualization and dashboarding tools.

  Data practitioners (engineers, integration specialists, and DBAs) will need to support analysts and other business super users with trusted, flexible, easy-to-use solutions. Cloud data platforms that add more sophisticated features and more AI/ML techniques into operational data analytics tools meant for those less skilled users will enable them to develop new insights that rely more heavily on domain knowledge about the business, markets, and customers rather than technical IT capabilities.

Taking Stock of Data Warehousing Trends

Businesses create data warehouses and data marts in relational databases to store and analyze many terabytes of big data.

The market for relational database solutions for data warehousing or data marts has evolved rapidly over the past few years. Multiple purpose-built products are available for reporting, data analysis, and business intelligence. However, these solutions fail to address current user trends, including:

- **Users want to mine and explore data with ad hoc queries.** In a traditional data warehouse, IT develops and controls the use of standard reports to represent the state of the business, running batch jobs at regular intervals.

  Increasingly, business users need more current data, or they need to explore different combinations of data to identify new trends and correlations. Because the data warehouse is tuned for performance on batch jobs and known reporting patterns, ad hoc queries for exploring can perform poorly and interfere with critical corporate reporting.

  Making data access available on a platform that scales well and does not require tuning in advance can accelerate the data discovery process and lead to new insights faster, without affecting other workloads.

- **Canned reports are being replaced by data exploration.** Data exploration empowers business users to dig deeper for insights from a wider variety of information sources than might be available in the corporate data warehouse.

  The proper self-service tools also free those business users from their dependence on IT resources to develop, allowing them to iterate more quickly and accelerate business value and competitive advantage.

- **Users are demanding more current information.** What good does it do to identify credit card fraud 24 hours after a transaction occurs, or to send a coupon to a consumer who has already left your store or website to buy the item elsewhere? These are missed opportunities!

  Tying the insights from operational data back into the business in near real time can deliver much more value out of data in the moment compared to batch reports and disconnected outcomes. The major implication is that the analytic database must be able to accommodate a steady stream of updates from operational systems, without an impact on query performance delivering those business insights.

IN THIS CHAPTER

- Looking at traditional data warehouse and data lake pitfalls
- Recognizing the need for flexible deployment options
- Leveraging artificial intelligence and machine learning
- Discovering the value of an operational data warehouse

Chapter 2
Understanding the Shortcomings of Existing Solutions

In this chapter, you explore the limitations of existing database solutions and discover how an operational data warehouse addresses these challenges.

Centralized Enterprise Data Warehouses and Data Lakes Can't Meet Today's Needs

Data warehouses have been around for decades and have established themselves as reliable reporting systems. They have also evolved into data marts, specialized appliances, and enterprise data warehouse (EDW) variants to meet emerging needs.

Data lakes have been around a shorter period of time and were developed to address some of the shortcomings of data warehouses, particularly in the areas of large volume and varied raw data reservoirs, data exploration, and data science tasks.

However, these solutions have their drawbacks when it comes to meeting today's business demands for real-time insights from operational workloads.

Some common pitfalls in traditional data warehouses include:

- **Stale data:** As the demand for organizations to operate in real-time or in-the-moment increases, data warehouses need to deliver ever more current and relevant data.

  Traditional data warehouses commonly fail to handle continuous streams of updates and are implemented with two stores — a *write store* where they apply updates, and a separate *read store* against which queries are run. They update the read store at predetermined intervals but there are long periods of time where queries are running against stale data.

  Lack of current data can mean businesses fail to respond to opportunities and threats quickly enough to stay competitive.

REMEMBER

- **Slow query performance:** An analytics query can be slow for many reasons. Possibly the database administrator (DBA) did not anticipate the query and had not defined a specific index, making that database unsuitable for ad hoc queries.

  This problem is compounded by the current and necessary trend toward "citizen" data analysts in which users with a limited understanding of the underlying data structures take a do-it-yourself approach to building queries — potentially bringing a database to its knees.

- **Limited security and privacy protection:** Increasingly stringent data security and privacy regulations and the increasing frequency and severity of data breaches have made security and compliance a top mandate.

  The European Union's (EU) General Data Protection Regulation (GDPR), the Japanese Act on Protection of Personal Information (APPI), the California Consumer Privacy Act (CCPA), and other legislation are the most recent examples of such data security and privacy regulations that

have major implications for organizations interacting with customer and citizen data.

Low-end databases can lack advanced encryption features for data in flight and at rest. For example, many databases lack column-level data masking capabilities and role-based access controls.

- **Growing costs of aging enterprise data warehouses:** Aging enterprise data warehouses based on monolithic or appliance-based proprietary hardware and software are expensive to upgrade and maintain, and they ensure continued vendor lock-in.

  Many of these systems are offered now as services in the cloud which removes the periodic CapEx expense but it is simply replaced by higher than average OpEx cost as the architectures are "lift and shift" and do not fully take advantage of key cloud technologies necessary to optimize for cloud economics.

  As data volumes and data types grow, adding capacity becomes expensive. This is especially true of appliance-based solutions where adding capacity often means buying a bigger appliance. This creates vendor lock-in with customers having little choice other than to buy expensive add-on options or do a complete box-swap upgrade.

  The customer is also dependent on the appliance vendor's engineering investment in the upgrade path, which may not keep pace with advancements in the industry-standard server market or with innovations such as cloud-based solutions for compute and storage.

- **Management complexity and lack of self-service options:** Many databases need a ton of specialized database developer and administration skills to design, deploy, maintain, and optimize. The lack of these skills in many organizations can result in lost opportunities, increased risk, and more downtime, among other potential consequences.

  End-users cannot easily access data or combine it with other external data sources for a more complete dataset and more accurate insights.

Some common pitfalls in traditional data lakes include:

- **Data swamp:** Unlike data warehouses, data lakes store raw data and process that data only at the time of processing and analysis — not at the outset.

  The vast majority of data lakes use inexpensive object storage, including the three major cloud providers. Use by anyone other than developers, engineers, and data scientists can be very difficult and cataloging and governing multiple data types and volumes can be difficult.

- **Optimization for data exploration not real-time operations:** Data lakes have essentially three groupings of components, large-scale inexpensive storage, management for use of that storage, and a collection of tools to perform data processing and analytics on the data in the lake.

  Historically, the toolsets for data lakes are geared to data exploration and analysis by data scientists, engineers, and developers, most notably for model development, training, and tuning for machine learning (ML) where all the necessary data is local to the data lake and manipulated out of the data lake on-premises (Hadoop, for example) or cloud storage (ADLS, AWS Blob, GCS). On the other hand, real-time operational workloads often need to pull data from multiple sources into the data warehouse and run real-time queries on that data out of computer memory.

TECHNICAL STUFF

REMEMBER

Engineers and scientists strive to preserve raw data to ensure there are always known, uncorrupted baseline datasets. Analysts and other power users prefer curated data to avoid the time and effort of repeated starts with unprocessed data. This difference is reflected in the core principles around reading and writing for data warehouses and data lakes, manifesting itself as *schema-on-write* (where raw data is formatted against a predefined schema prior to being placed in tables in the data warehouse) versus *schema-on-read* for a data lake (where data is manipulated only after it is taken out of or leveraged by a tool sitting on top of the data lake).

Enterprises Require On-Premises and Multi-Cloud Options

The center of gravity for operational data is moving in the direction of cloud solutions. However, the reality is that most enterprises will need both cloud and on-premises solutions for the foreseeable future. Additionally, hybrid cloud and multi-cloud solutions are becoming increasingly important to enterprises.

There's a general consensus that the vast majority of companies and government organizations have a cloud-first strategy, and a recent survey by Flexera states that the top three workloads being migrated/deployed in the cloud are:

- Data warehouses
- Database as a service (relational)
- Database as a service (NoSQL)


TIP

The Actian DataCast 2020: Hybrid Data Trends Snapshot survey found that a hybrid landscape is unavoidable with 85 percent of respondents having an average of 52 percent of their data in the cloud, citing security, cost predictability, and regulatory compliance as the top three reasons the expectation that some data will remain on-premise.

In today's data environment, data will be sourced from any number of places both inside and outside of the enterprise, and it may be too costly or violate data sovereignty rules to create a separate copy of the data in a centralized location. Increasingly, enterprises are looking for ways to access data as if it were in a virtual data warehouse or lake, going to the data where it naturally resides rather than copying it everywhere in anticipation of it being needed later.

However, many enterprise data warehouse solutions have limited deployment options. Appliance-based solutions, for example, can't be deployed to the public cloud. The cloud data warehouses associated with the three major cloud service providers are focused only on their own environment and ecosystem. Virtualized versions of these appliance-based and single-cloud solutions may not perform optimally for data pipelines supporting operational workloads that span on-premises and multiple clouds — increasingly the norm and not the exception.

Databases Must Be Self-Tuning with Dynamic-Elasticity

As database systems become increasingly sophisticated and more complex, the ability of IT staff to maintain and optimize these systems has become tenuous. The global shortage of IT professionals — particularly those with specialized IT skills such as data scientists, data architects, database administrators, and security analysts — further exacerbates the challenge of database maintenance and optimization.

Also, as mentioned in Chapter 1, these database systems support mission-critical operations that require agility, the ability to respond to changing business and market stimuli with real-time insights that may require large sets of data from more diverse sources to be analyzed at the speed of business.

Representative use cases are explored in Chapter 3, but the point is these systems must be scaled up and down at the drop of a hat for more data, more concurrent users, faster response times — or all of the above.

REMEMBER

To effectively address these gaps, modern databases must have automated self-tuning capabilities including maintenance, repair, security, and performance optimization. They must also be able to scale up and scale down without starting and stopping the workloads they're running. But all of this must be done in a cost-conscious environment, if the system isn't being used, it should automatically stop and then automatically restarted when required.

Artificial intelligence (AI) and machine learning (ML) technologies enable modern databases to go beyond automated alerts and scripted actions to intelligently act on system events without manual human intervention.

Introducing the Operational Data Warehouse

Data warehouses are great — at reporting yesterday's data. Data lakes are also great — as a low-cost historical archive or in support of science projects (because their query performance is so poor and the tools used are not easily approached by non-IT power users). However, businesses in today's fast-paced hypercompetitive environment need to harness all of their data in-the-moment through a more inclusive and broader team.

An operational data warehouse (ODW) combines the power of a traditional data warehouse, the scale of a data lake, and the economics of the cloud in a real-time solution that runs on commodity hardware or in the cloud. Table 2-1 compares key capabilities of enterprise data warehouses, operational data stores, data lakes, cloud-only data warehouses, and operational data warehouses.

Key characteristics of an ODW include:

- **Fast:** Built on an underlying architecture optimized for advanced analytics and query performance, requiring little or no tuning in anticipation of certain workloads (like indexing or aggregations) and maximizing the variety of workloads it can support
- **Scalable:** Scales to large data capacities with an economical and flexible storage layer, connecting to a variety of existing legacy and new sources of data
- **Trusted:** Offers multiple data protection mechanisms to meet enterprise security requirements and comply with tough regulatory environments
- **Flexible:** Offers flexible deployment options, including both on-premises and multi-cloud options
- **Easy-to-Use:** Directly useable by analysts and other non-IT power-users (business technologists) through simple menus to ingest and prepare data, run analytics, and leverage visualization and other analytics tools they're already familiar with
- **Robust:** Delivers enterprise-level resiliency and manageability

TABLE 2-1 Comparing Data Solutions

	EDW	Operational Data Store	Data Lake	Cloud- Only DW	ODW
Description	Single source of truth for archiving and reporting	Domain- specific data mart for operational workloads	Comprehensive, economical repository	Rapid deployment, flexibility, and scale	Designed for data- driven in-the- moment business
Performance	Best	Best	Poor	Good	Best
Economic scale	Poor	Poor	Best	Good	Best
Data integrity	Best	Best	Poor	Best	Best
Real-time insights	Good	Best	Poor	Good	Best
Security	Best	Best	Poor	Best	Best
Deployment flexibility	Good	Poor	Good	Poor	Best

Key benefits of an ODW include:

- Run complex, ad hoc queries against billions of records in seconds
- Process hundreds of records in a single CPU instruction cycle with vector processing
- Execute updates without any performance penalty
- Get consistent query results even if the data changes
- Exploit dedicated CPU core and caches running 100 times faster than random-access memory (RAM)
- Scan data faster using self-indexed blocks
- Leverage the latest innovations and economics of the cloud

REMEMBER

What differentiates an operational data warehouse from an enterprise data warehouse is the capability to be deployed on a variety of infrastructure components (either on-premises or in the cloud), scale from gigabytes (GB) to petabytes (PB), run standard SQL, and handle operational updates to data with atomicity, consistency, isolation, durability (ACID) compliance — all at an economical cost without sacrificing enterprise levels of reliability, availability, and security.

IN THIS CHAPTER

- Calculating the benefits of data analytics in financial services
- Recognizing the value of analytics in retail
- Getting to know your data in social media
- Distributing data insights in transportation and distribution
- Taking care of data analytics in healthcare and clinical research

Chapter 3
Exploring Use Cases Driving Changes in Demand

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

Financial Services

Market participants today face a costly and complex labyrinth of challenges that have become the new normal. Global capital markets data, currently fragmented by asset type and class, flows in at unprecedented speeds and volume. New market regulations come with increased scrutiny from regulators. Effective risk management is not only a challenge but also a necessity for business survival.

In this new world of big data, many financial services firms face an uphill battle because they still carry the burden of legacy systems, outdated analytical methodologies, and old databases. To overcome today's challenges and stay competitive, financial organizations must modernize their information management systems to overcome big data issues, slow response times, or a lack of data scientists to decipher the data and identify an appropriate level of risk with which to maximize profits and minimize loss. Overall, organizations need better decision making via big data analytics.

Big data analytics is changing the world of capital markets and global banking. Firms deploying new analytics platforms are calculating enterprise credit and market risk in minutes versus 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 in the ongoing search for alpha (the active return on investment).

TIP

Big data is a winner's game. Leaders who embrace it will 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 will set them apart from the competition.

By transforming data into real-world business value with speed, efficiency, and transformational analytics, an operational data warehouse can help global capital market firms

- Exploit parallel processing for extreme computational performance to yield better insights
- Consolidate silos of data across front, middle, and back offices to one central view, regardless of source or format, structured or unstructured
- Present one consolidated view of all market and product data, merge with additional contextual data, and analyze every granular fluctuation
- 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 or prevent business-compromising events that violate regulations or corporate ethics before they happen
- 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 (ms) response time requirement for its Eikon and Elektron data and trading applications. The company implemented an Actian Vector farm consisting of 80 servers with each server hosting multiple client accounts. Vector provides a hub for all the data used by the Eikon and Elektron applications and meets the service-level agreement (SLA) requirements for complex queries with sub-20ms response times. While the solution remains on-premises today, there are plans afoot to move these workloads to the cloud.

Retail

Retail use cases for data analytics leveraging an operational data warehouse are as numerous as bargains and sales on Black Friday and Cyber Monday. Some examples include:

- **Customer profile:** Granular, multi-channel, near real-time customer profile analytics can tell you about your customers, the best means to connect, the targeted offers that will resonate, their predilection to churn, and the best ways to personalize the entire customer experience to win more business and drive up loyalty levels. To gain a more complete and accurate profile, you need to mine all available information, in any format, from any location or channel, whether structured or unstructured. 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. You can use micro-segmentation models to find and classify small clusters of similar customers, and customer value models 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. With an operational data warehouse, you can
  - Connect to all of your data, from account histories and demographics to mobile and social media interactions, and blend these disparate sources with speed and accuracy
  - Uncover key purchase drivers to understand why someone purchases or rejects your products
  - Assign customer value scores by correlating which characteristics and behaviors lead to value at various points of time in the future
  - Optimize outbound marketing to give prominence to your high-value customers
  - Customize inbound customer touch centers by arming call centers with highly personalized customer scores
  - Increase customer lifetime values cost effectively with individual precision, improving both loyalty and profitability

- **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 data available to you, structured or unstructured, from traditional enterprise sources, social networks, customer service interactions, web click streams, and any other touch points that may occur.
- **Campaign optimization:** Traditional campaign optimization models use limited samples of transactional data, which can lead to incomplete customer views. An operational 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. An operational data warehouse increases the accuracy of churn predictions by combining and analyzing traditional transactional and account datasets with call center text logs, past marketing and campaign response data, competitive offers, social media, and a host of other data sources.

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, this knowledge became more of a challenge.

By deploying Actian Vector, 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 Vector enabled Sheetz to double its access to historical data and be ready for expected growth over the next few years.

- **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. An operational data warehouse brings in additional sources, in varying formats, enabling discovery of critical patterns, at any product level, to create a competitive advantage.

Social Media

Social media has emerged as one of the largest sources of data for organizations in virtually every industry. In fact, a new specialized form of data science known as *media mining* is focused on the mining of data from sources such as Facebook, YouTube, Instagram, Twitter, LinkedIn, and others. The motivation behind media mining is to glean better insights regarding opinions and preferences from a broad range of demographic groups. These insights are then used to conduct targeted marketing campaigns, tailored to each demographic group that has shown a propensity to buy a product or service. In addition, marketing groups can study the reaction to a general marketing campaign. The goal is to determine which content most favorably resonates with a given demographic of interest.

SOCIAL MEDIA GIANT NK LIKES ACTIAN VECTOR

Social media companies often use advertising to monetize user behavior. NK (`http://nk.pl`) is a large social media site in Poland, twice as large as Facebook in that country. The product managers at NK were answering business requests for data about user behavior with workaround solutions built on MySQL databases with huge queries. They frequently had to wait days or weeks for their queries to complete against the vast amounts of data about their users.

NK implemented a new solution that collects data from various sources, including Hadoop, and imports it into Actian Vector for its fast processing performance. Actian Vector has enabled faster report generation, improved user experience via dynamic dashboards, simplified queries, and better access to harmonized data.

Consider a scenario where a retailer contracts the use of a set of large video screens at airports, shopping centers, or other areas of high foot traffic. They run a set of different ads on the screens in different locations over a short period of time — a few days to weeks. The marketing group collects tweets via the Twitter application programming interface (API), filtering on time (the days that the ad was visible) and geo-location (where the ad was visible). Next, a natural language processing (NLP) system is used to score each tweet, which identifies the tweet as "favorable," "neutral," or "unfavorable." If the retailer has a loyalty program that includes phone numbers, the tweets can be mapped to known customers and correlated with previous buying habits. The retailer can now track opinions as well as transactions to determine which marketing strategy produces the best "lift" for a particular product and demographic combination.

Transportation and Distribution

The transportation industry is combing Internet of Things (IoT) technologies with near real-time data analytics to improve operational efficiencies and increase customer satisfaction. For example, commercial airlines have thousands of flights departing every day of the week. Many of the passengers have flights that are composed of one or more connections before arriving at their final destination. Consider a group of ten people on a flight from Los Angeles to New York with a connection in Denver. If the flight departs from Los Angeles late, these ten people will potentially miss the connecting flight from Denver to New York. Now, if another flight is leaving Denver heading to New York and has five seats available, the airline can accommodate five of the affected customers.

The question becomes: Which five customers from the group of ten will be seated on the new flight? The answer is not as straightforward as you might think. A simple approach might lead you to believe you should give the seats to your "frequent fliers," but what if you have a young couple who just got married and paid first-class to go to New York? Maybe they would be a better choice. By tracking details regarding each customer's special circumstances, the operational data warehouse can make a more informed decision based on insights that have been gleaned from analysis of years of historical data. This concept is known as *situational awareness* and is used in automated decision making across many industries.

LUFTHANSA SYSTEMS RELIES ON ACTIAN TO MAKE THOUSANDS OF FLIGHTS RUN SMOOTHLY

As one of the world's leading full-service IT providers and industry specialists to the airline and aviation sector, Lufthansa Systems offers LidoFlight Planning as well as other airline solutions to some 300 national and international commercial airlines across the world.

Lufthansa's Lido flight and route planning software is built on Actian's Actian X database technology and the Actian OpenROAD application development platform, delivering real-time availability of data, high levels of uptime, reliability, and stability. "We have been using Actian X for some time in our business," explains Rudi Koffer, senior database software architect at the Lufthansa Systems Airlines Operations Solutions division in Frankurt Raunheim, Germany.

Healthcare and Clinical Research

Healthcare innovators are paving the way to outcome-based healthcare instead of fee-for-service healthcare, 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 (see Figure 3-1). Using analytics, they are transforming data into actionable insights that improve healthcare delivery, reduce re-admissions, promote preventive care and accelerate research. Analytics also holds the key to reducing fraud, waste, and abuse in healthcare systems in order to stop revenue leakage.

TIP

Analytics-based insights can meet many healthcare priorities, including:

- **Better patient outcomes**
  - Decrease re-admissions
  - Improve preventive care

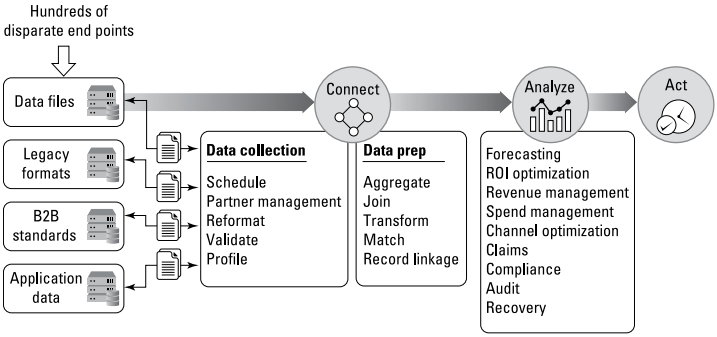


FIGURE 3-1: Data analytics for healthcare providers and payors.

- Enable patient self-service
- Enhance patient safety

» Improved claims/cost management

- Identify eligibility fraud
- Analyze member cost
- Process claims more efficiently
- Avoid improper payments

» Big gains in operational effectiveness

- Prevent revenue leakage
- Capture real-time insights
- Optimize staffing levels
- Eliminate operational redundancies

» Health informatics

- Leverage the value of EMRs
- Gain member profile insight
- Analyze population and behavioral health studies
- Conduct trend and risk analysis
- Improve marketing effectiveness
- Increase customer satisfaction

» Advanced research and development

- Use predictive insight to improve disease prevention
- Accelerate drug and clinical research and development
- Analyze complex biotech research

CTSU ANALYZES MASSIVE DATA SETS IN MINUTES

CTSU (the MRC/Cancer Research UK/BHF Clinical Trial Service Unit and Epidemiological Studies Unit of Oxford University) primarily studies the causes and treatment of chronic 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 Vector to perform analyses of these massive data sets. Alan Young, director of information science at CTSU, says: "Without Actian Vector, we simply would not be able to process this information, without having to wait days or weeks for each output."

PEDIATRIX FINDS REPORTING IS 80 PERCENT FASTER WITH ACTIAN DATA INTEGRATION

Peadiatrix, a leading provider of newborn, material-fetal, and pediatric physician subspecialty services, was unable to integrate data from multiple sources which caused bottlenecks and delays in patient insurance claims processing.

With Actian DataConnect Data Integration they were able to create a single source of truth for improved claims processing. "With DataConnect, we can create mapping for each report and form them into a single database table that users can process. It takes five minutes to run the report and there is only one report to process; it is uniform and much easier to manage," says Jennifer Arriza, director of applications for Pediatrix Medical Group.

IN THIS CHAPTER

- Understanding advances in processor technology
- Using best practice techniques to optimize data warehouse performance

Chapter 4
Defining Requirements for An Operational Data Warehouse

In this chapter, you learn how advances in central processing unit (CPU, or "processor") technology and industry best practices are defining the technical requirements and capabilities of a modern operational data warehouse.

Exploiting the CPU

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 increases in clock speed and the number of transistors on the chip. CPU manufacturers have introduced additional performance features, such as multi-core CPUs and multi-threading, which are transparently leveraged by most database software.

REMEMBER

Moore's Law describes a long-term trend in which the number of transistors that can be placed on an integrated circuit (IC) 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
- Other CPU performance features (super-scalar functions, out-of-order execution, and hardware-accelerated string-based operations)

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 process data one tuple at a time (scalar) using a single instruction single data (SISD) model.

A *tuple* is a single record (row) in a relational database. Most of the CPU time is spent managing tuples rather than on the actual processing. In contrast, SIMD enables a single operation to be performed on a set of data at once (vector), as shown in Figure 4-1.

REMEMBER

In traditional scalar processing, operations are performed on one data element at a time. In vector processing, operations can be performed on multiple sets of data at a time.

A modern operational data warehouse uses vectorized processing to dramatically increase its processing speed — using CPU resources to perform actual work rather than to manage overhead — in the following ways:

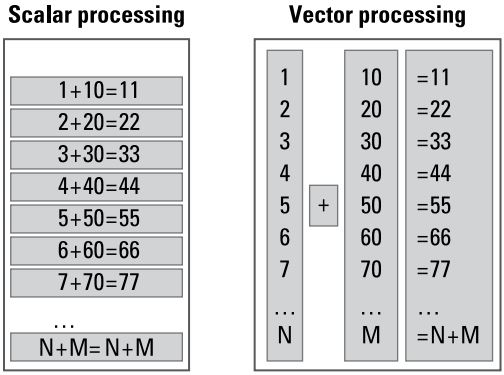


FIGURE 4-1: Scalar and vector processing operations.

- Vectors, consisting of multiple tuples (hundreds or thousands) of data elements, can be processed all at once.
- Operations ("primitives") are implemented as branch-free loops over arrays of cache-resident column values.
- Interpretation (function call) overhead is amortized.
- Operations leverage SIMD to perform a single instruction on multiple column elements at a time.
- Maximum utilization of superscalar CPU pipelines is achieved through fewer branches and function calls, as well as a higher instruction cache hit ratio, which minimizes data dependencies.

TIP

Because typical data analysis queries process large volumes of data, the average computation against a single data value can take less than a single CPU cycle when leveraging SIMD.

CPU cache as execution memory

When most traditional databases entered the market in the 1980s, high-end computers only had roughly 8 megabytes (MB) of random-access memory (RAM, also known as "main memory" or "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 ever-increasing clock speed of the CPU, 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, computers typically have 8MB or more of memory built directly into the processor itself as level-1 (L1) and level-2 (L2) cache, which can be accessed much faster than shared memory (RAM).

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 modern operational data warehouse avoids using shared RAM as execution memory. Instead, the CPU core and CPU caches are used as execution memory, thereby delivering significantly faster data processing throughput.

Figure 4-2 shows the relative performance characteristics of disk, memory, and 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
- **Cache:** 100GB to 200GB per second

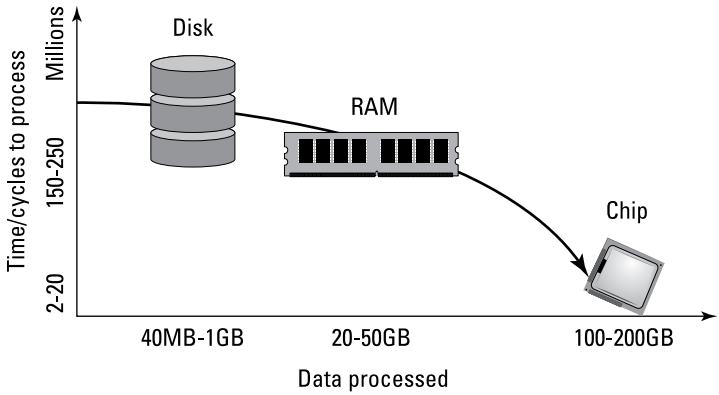


FIGURE 4-2: Data processed "on chip" is exponentially faster than data processed in memory and on disk.

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

Access latency is the time delay that occurs between sending a data request (for example, to disk) and fully receiving the requested 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 accesses only a subset of the columns (see Figure 4-3).

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 4-3: Row-based storage.

Row-based storage works well for online transaction processing (OLTP) systems in which

- Stored data is highly normalized, so tables are relatively narrow
- Queries typically retrieve relatively few rows
- A small number of queries involving large volumes of data are processed

In contrast, data warehouses have different characteristics:

- Tables are often partially denormalized, resulting in many more columns per table, not all of which are accessed by most operations
- Most queries retrieve many rows
- Large data sets are often added at once or through an ongoing, controlled stream of data, rather than through an ad hoc process

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 4-4).

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 4-4: Column-based storage.

In column-based storage, database queries read only the attributes that are needed. This saves I/O bandwidth, fits more relevant data into the buffer pool (RAM), and results in less "cache pollution" and misalignment.

TIP

In addition to the benefit of data elimination when accessing fewer than all table columns in a query, column-based storage provides better data compression than row-based storage.

Hybrid column store

In a hybrid column store, data is stored using a pure column-by-column approach by default. For tables that are indexed on more than one column, indexed columns are typically accessed together, and the indexed columns are stored together in a single data block. However, within the block, data is still stored column-by-column to optimize compression.

TIP

The user may choose to store data row-by-row if data allocation for column-by-column storage requires too much upfront data allocation. The choice for row-based storage can make sense for extremely wide tables or those with relatively few rows.

Positional Delta Trees (PDTs)

One of the biggest challenges with most column-based databases is incremental small insertions, updates, or deletions (as opposed to large bulk data load operations).

This challenge can be addressed with high-performance in-memory positional delta trees (PDTs). PDTs enable online updates without affecting read performance in VMs and containers and on Linux, Windows, and Hadoop. Conceptually, a PDT is an in-memory structure that tracks the tuple position and the change (delta) — insertions, modifications, and deletions — at that position. PDTs are designed to merge changes quickly by providing the tuple positions where differences must be applied during a scan.

Queries efficiently merge the changes in PDTs with data stored on disk (see Figure 4-5). Because of the in-memory nature of PDTs, small data manipulation language (DML) statements can be processed efficiently. A background process writes the in-memory changes to disk once a memory threshold has been exceeded.

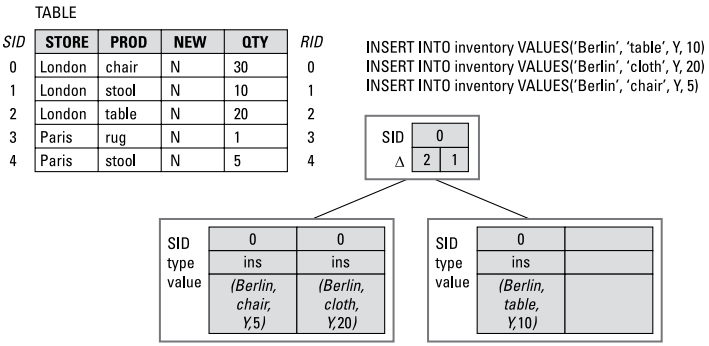


FIGURE 4-5: PDTs enable real-time updates in a column store.

REMEMBER

PDTs store small incremental changes (inserts that are not appends), as well as updates and deletions, with the exception of truncations.

Data compression

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 an algorithm that works well for text, for example, may not work well for numeric data.

Column-oriented storage compression allows the algorithm to be chosen according to the data type, as well as the data domain and range (even where the domain and range are not explicitly declared). Data is compressed on a column-by-column, page-by-page basis using any one of the following algorithms, or a combination of them:

- **Run-length encoding (RLE):** A data value is stored, as well as the number of subsequent values that are the same. This compression algorithm is efficient on ordered data with relatively few unique values.
- **Patched frame of reference (PFOR):** A base value is determined per data block, and other values in the same block are encoded by storing the difference with the stored value using as few bits as possible.

  This approach is helpful because the range of the data is typically much smaller than the range of a used data type.

  What makes PFOR special is the treatment of outliers. For example, if 99 percent of values are in a range from 0 to 255 and one percent of values are large (for example, around a million), with PFOR, most of the data is stored using only one byte.
- **Delta encoding on top of PFOR:** To reduce the values of the integers with PFOR, it is sometimes more efficient to store the delta from the previous value. This method can be efficient on ordered data.
- **Dictionary encoding:** This algorithm stores pointers to a dictionary of unique values (PDICT). Dictionary encoding is efficient for a limited number of frequently occurring values.
- **LZ4:** This Lempel-Ziv lossless data compression algorithm detects and encodes common fragments of different string values. LZ4 is efficient for medium and long strings.

Storage indexes

A storage index simplifies database schema by negating the need for a complex indexing scheme. A storage index is efficient in determining whether a database block is a candidate block for a query, either explicitly because of filter criteria or implicitly as a result of processing table joins.

In extreme cases, a storage index provides the same benefit as data partitioning does for other databases without the overhead of multiple database objects or having to design and maintain a partitioning strategy.

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 there are, the quicker the task will be completed.

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.

REMEMBER

Many analytic queries are compute-intensive. Parallelizing those computational tasks is extremely important for satisfactory performance.

For any sequential plan, you can find the optimal parallel plan using the following process:

1. Estimate the rough cost of all the operators based strongly on hints provided by the database management system (DBMS) server. Up-to-date statistics and heuristics are critical.
2. Try possible parallelization rewrites at different operators using special Xchange operators.
3. Estimate the cost of the parallel plan.
4. Select the cheapest plan.

A modern operational data warehouse typically answers queries for many users simultaneously. For maximum performance, each of these queries will be parallelized. The data warehouse will balance the level of parallelization that makes sense when trying to service 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

- Maximizing flexibility, scalability, and performance
- Protecting data security and privacy
- Ensuring query integrity and continuous updates
- Providing interoperability and extensibility
- Delivering resiliency and lowering total cost of ownership
- Automating the flow of data

Chapter 5
Ten Things to Consider When Evaluating an Operational Data Warehouse

Here are ten important considerations to help you evaluate which operational data warehouse is right for your organization.

Maximizing Deployment and Operating System Flexibility

An operational data warehouse should provide flexible platform deployment options, both on-premises, and the cloud.

When considering on-premises options, look for a solution that runs on commodity hardware to avoid vendor lock-in issues associated with proprietary hardware appliances or that only runs on and is optimized for a single cloud environment.

For maximum flexibility, an operational data warehouse should support public, private, and hybrid cloud deployment models across Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) offerings.


TIP

Database Platform as a Service (dbPaaS) and Database as a Service (DBaaS) are specialized PaaS offerings and include traditional relational and NoSQL database offerings.

Popular public cloud solution offerings typically include

- **Amazon Web Services (AWS)**
  - Elastic Compute (EC2)
  - Elastic Block Store (EBS)
  - Simple Storage Service (S3)
- **Microsoft Azure**
  - Virtual Machines
  - HDInsights
  - Azure Data Lake
  - Azure Blob Storage
- **Google Cloud Platform (GCP)**
  - Google Compute Engine
  - Google Kubernetes Engine
  - Google Cloud Storage

Cloud software licensing options should include

- **Bring your own license (BYOL):** Allows you to transfer your existing on-premises licenses to the cloud
- **Pay as you go (PAYG):** Typically structured as hourly consumption charges with monthly pricing plans
- **Prepaid plans:** Typically offered as annual pricing at a discounted PAYG rate

- **Managed services:** Covers monitoring, management, and sometimes operations

Look for an operational data warehouse that supports popular operating systems including Windows, Linux (such as Red Hat Enterprise Linux, CentOS, SUSE, Debian, and Ubuntu), and Kubernetes container deployment on Google, Amazon Web Services, and Microsoft Azure.

Leveraging Scale-out Performance

Scalability must be provided by an operational data warehouse in three dimensions:

- **Vertical scalability (scale-up)** enables workloads to take advantage of more processor and storage capacity on a single system.
- **Horizontal scalability (scale-out)** provides the ability to grow the operational data warehouse to handle larger databases and more users when you have saturated the hardware capacity of a single system by adding a cluster of systems.
- **Cloud scalability (elastic scale-out)** enables workloads to utilize more capacity across underlying cloud resources (compute, storage, and networking) without any management burden.

REMEMBER

The ability to exploit multiple resources in a clustered or cloud environment is the key to supporting larger datasets. A single server or its cloud processing equivalent supports a finite amount of processor cores, memory, storage, and data bandwidth. A database scale of ten terabytes or its use by multiple concurrent users will most likely saturate a single system.

TIP

An additional advantage of Kubernetes-orchestrated workloads is that you can scale up in one cloud, across multiple clouds, and even to hybrid environments on-premises and the edge.

Optimizing Performance and Maintenance

Changes to ODW data should be made with the lowest performance penalty. Columnar data blocks that maintain their min-max value metadata eliminate the overhead of creating indexes that must be updated with every change, as traditional row-based databases do.

Ensuring Security and Privacy

When asked why he robbed banks, the notorious bank robber Willie Sutton is reputed to have answered, "Because that's where the money is."

Data has become the most valuable asset for businesses and organizations everywhere, and the database has become a modern vault to protect valuable data from cybercriminals. To protect your organization's data, an operational data warehouse must have robust data security and privacy protection capabilities, including:

- **Data masking:** Database administrators (DBAs) or security administrators can mask individual columns within a table so that only authorized users can see the underlying data. Unauthorized users will see an obfuscated version of the data.

  For example, when dealing with credit card numbers, unauthorized users may see only the last four digits of the card number and an X representing each preceding digit, though authorized users see the full credit card number.
- **Encryption:** Both column-level data-at-rest encryption and data-in-motion encryption are critical data security and privacy capabilities.

  Function-based data encryption further enables applications to encrypt and decrypt sensitive data stored in tables through the AES_ENCRYPT_VARCHAR and AES_DECRYPT_VARCHAR SQL functions.

- **Stored procedures:** Stored procedures give database administrators greater control over database access.

  This means DBAs can grant permission to execute a stored procedure even if the user has no direct access to the underlying tables referenced in the procedure.

  Stored procedures can be used to guard against SQL injection attacks because their input parameters are treated as literal values and not as executable code.

REMEMBER

The growth of cybercrime and ever-increasing number of data privacy regulations means that even "internal" systems must be secured. A good operational data warehouse needs to offer built-in support for advanced encryption, auditing, role-based security, and data masking.

Maintaining Query Consistency

Some databases sacrifice query integrity for speed. Even as the underlying data changes, an excellent operational data warehouse provides row-level locking and full read consistency for running queries.

Providing Fresh Data Efficiently

Having the data continuously updated by micro batches or streamed singleton updates throughout the day provides the most current information for analytics-based decision making.

Broad Interoperability and Connectivity

A good operational data warehouse provides open application programming interfaces (APIs) such as Open Database Connectivity (ODBC) and American National Standards Institute (ANSI) SQL to enable it to work with the multitude of query tools an organization might use. Many organizations use more than 20 different visualization and query tools.

The ability to consume or ingest data at high speed is a critical operational data warehouse requirement. If you cannot load your data in a reasonable time, the result is having to work with summary data or, worse, using stale data.

The usefulness of an analytic database is closely tied to its ability to ingest, profile, store, and process vast quantities of data. Data is typically ingested from disparate sources with diverse data types such as operational databases, JSON files, comma-separated values (CSV) files, and continuous data streams.

Extensibility is another critical requirement for an operational data warehouse. Modern data science requires access to the rich machine learning (ML), and artificial intelligence (AI) functions provided by integration with Apache Spark.

Enterprise-class Backup and Recovery

Delivering enterprise-level resiliency and manageability means having solid backup, recovery, failover, and replication capabilities for the operational data warehouse.

Economy

The total cost of ownership for a database technology being used to support a particular business case can be affected by several factors.

One is running standard servers to avoid esoteric appliances when on-premises. Others include leveraging cloud economics with "pay for what you use as you go" models.

Cloud elasticity allows you to quickly and affordably scale-up and down according to business requirements, data volumes, and concurrent users, optimizing operating efficiencies and budget outlays.

Integrated ETL and Fast Data Loading

Being able to populate your operational data warehouse with masses of data from disparate data sources (both internal and external to your organization) is essential.

The primary data source for the operational data warehouse in a traditional enterprise environment may be dozens of systems of record and engagement applications and transactional databases with long-standing, structured data that continually populate the data warehouse on an ongoing basis.

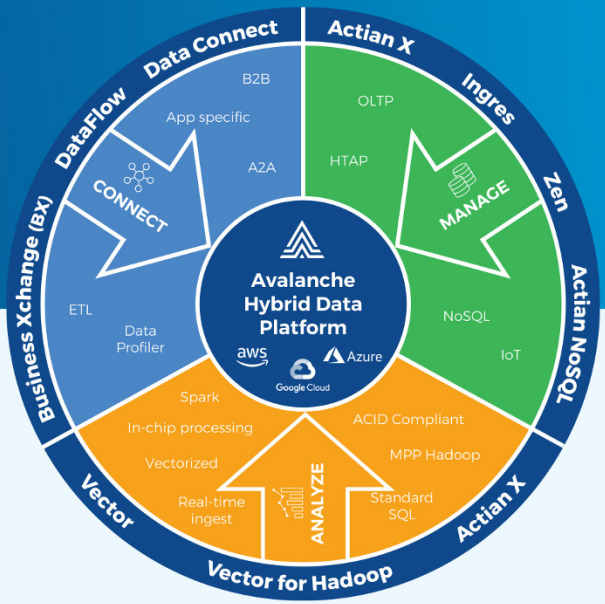
Over the last ten years, additional sources such as SaaS applications (such as Salesforce, Workday, NetSuite, ServiceNow) are also being deposited into the operational data warehouse. In all cases, the operational data warehouse must be able to automate the flow of data extracted from many sources, clean it, augment it, transform it, and load it fast.

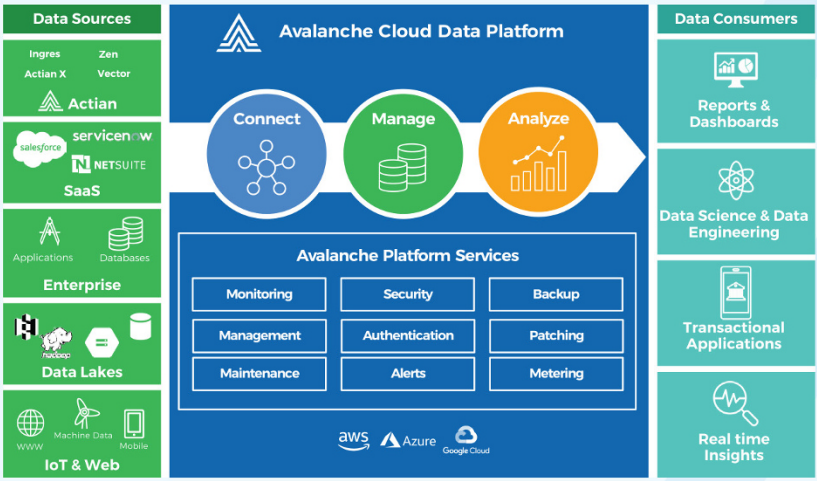
REMEMBER

It's essential that your operational data warehouse support data ingestion and preparation by a broad set of users, including data engineers and scientists, developers, DBAs, analysts, and other power users in various lines of business. Supporting such a wide range of users means providing low- or no-code options for data consumption — preferably through simple menu-driven interfaces and drag-and-drop connections between various data fields to manipulate source-to-destination relationships for data exploration, starting with easy-to-use templates and pre-built connectors for standard applications and data repositories.

Unlike a data lake or a data hub, an operational data warehouse is focused on production environments. This means that ad hoc data exploration and initial preparation must be replaced by loading or scheduling, orchestration, and management of data integrations across multiple data sources with automated query and advanced analytics processing that delivers results to populate various business and operational dashboards, visualizations, and embedded analytics products across several parallel data pipelines.

Delivering trusted, flexible, and easy-to-use data analytics





For more on Actian's Data Integration, Management, and Analytics Products, **visit Actian.com**

ACTIAN™

Operate in the business moment

Business uncertainty is at an all-time high. Competition in this ever-changing landscape requires your IT and business technologists to be able to extract value from disparate, diverse, and often real-time data sources to quickly deliver accurate insights that drive your business forward. This data-driven approach requires a new type of data platform: an operational data warehouse that is flexible, easy to use, that can start small and scale fast. This book shows you how an operational data warehouse goes beyond reporting on historic, static data and instead operates with fresh, active data to drive business actions and outcomes.

Inside...

- Understand key trends in data analytics
- Understand the limitations of existing data warehouse solutions
- Explore data warehousing use cases
- Define requirements and capabilities
- Evaluate your options

Lawrence C. Miller has worked in information technology for more than 25 years. He has written more than 60 For Dummies books. **Emma McGrattan** leads R&D at Actian and has over two decades of experience in high-performance analytics, data management, integration, and application development technologies.

Cover images: binary code © ratios / Shutterstock, cloud courtesy of Actian

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-87329-7
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.