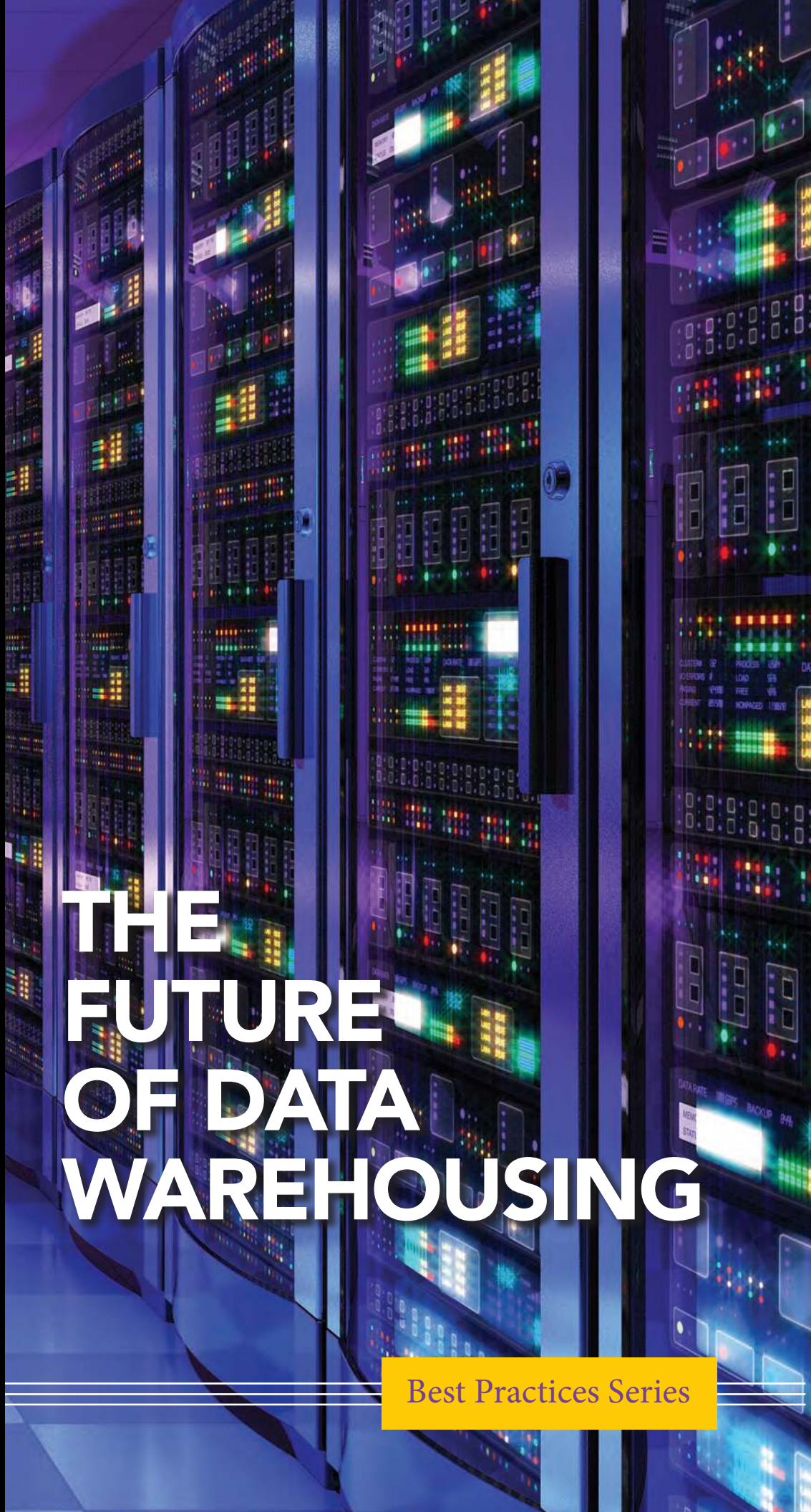


# THE FUTURE OF DATA WAREHOUSING

Best Practices Series

database
TRENDS AND APPLICATIONS

# RETHINKING THE FUTURE OF DATA WAREHOUSING

Best Practices Series

**Does the data warehouse** still have a place in today's fast-moving, real-time digital enterprise? Many pundits, analysts, and vendors have proclaimed the impending demise of data warehousing, suggesting that it has become too slow, isolated, and cumbersome to deliver insights at the push of a button. However, the data warehouse has proven the doomsayers wrong, with evidence that it is evolving into an integral and essential piece of the big data landscape.

For decades, companies have invested millions of dollars designing, implementing, and updating enterprise data warehouses as the foundation of their business intelligence systems. And they are still just about everywhere. Most larger enterprises still maintain data warehouses, and small to medium-businesses are also finding data warehouses a cost-effective option, thanks to the cloud.

Now, data warehouses are poised to play a leading role in next-generation initiatives, from AI to machine learning to the Internet of Things. While data warehouses do not appear frequently in marketing literature or analyst reports on these emerging technologies, data warehousing will remain a critical cornerstone of the foundation of the digital era ahead.

"If you can't build a data warehouse, you shouldn't do AI," Andrew Ng, noted computer scientist and co-creator of Google Brain, has said. This requirement runs deep through every digital engagement.

"Every single company I've worked with and talked to has the same problem without a single exception so far—poor data quality, especially tracking data," according to Ruslan Belkin, vice president of engineering for Salesforce.

For further proof of the continuing importance of data warehousing, look to skills demand in the IT workforce. Tellingly, data warehouse engineer has ranked sixth of the top 10 jobs in demand, a recent analysis by Indeed found.

There has been pressure on today's legacy data warehouses to evolve—both architecturally and technologically—to deliver

the agility, scalability, and flexibility that business need to thrive in today's data-driven economy. Alongside new architectural approaches, a variety of technologies have emerged as key ingredients of modern data warehousing, from data virtualization and cloud services to Hadoop and Spark and machine learning and automation.

Here's the shape of the future of data warehousing:

**Data warehousing is going to be cloud-based.** What was unimaginable just a decade ago is no longer the working reality today—enterprises are turning to cloud to power and store their data warehouses. It will be versatile, providing both real-time and historical insight. The data warehouse will work in unison with other components of the environment. Information from data warehouses will increasingly be the source of insights for both real-time and analytical actions to provide customer service at the time it's needed, while also serving as a repository for historical data. There has been rapid growth and excitement in recent months and years in cloud data warehouses hosted by leading internet companies such as Google and Amazon, which is essentially putting a stamp of approval on the concept of data warehouses in the cloud. In addition, traditional cloud providers also offer their capabilities as a cloud service, along with their traditional on-premise products.

**Data warehousing is being extended into modern analytics ecosystems through the use of data virtualization.** By federating multiple data warehouses, data virtualization can augment traditional ETL and data replication processes by acting as a virtual data source while also isolating applications from the complexity of disparate and changing underlying data sources.

**Data warehousing is going to be analytical.** The data warehouse world has blended with the analytics world to the point where they are one and the same. Data warehouses, for all intents and purposes, are data analytics platforms. Companies recognize that data analytical power is crucial to every aspect of their operations and products, and data warehouse technology is already delivering this power.

*Data warehouses are poised to play a leading role in next-generation initiatives, from AI to machine learning to the Internet of Things.*

**Data warehousing is going to empower users like never before.** The key advantage to data warehouse environments is the emphasis on self-service. Business end users have long had the capability to build queries or ask questions of their data that had never been asked before, due to the limitations of data silos. Data environments are only growing more diverse and complex, and budgets for IT staffing are getting tighter. The platform data warehouses provide for building queries is proving invaluable at a time when decision makers can't afford to wait on their IT or data management departments for answers.

**Data warehousing is going to feed into data lakes, Hadoop, and Spark—as well as the other way around.** There has been a great deal of discussion about the future of data warehouses in a world increasingly served by data lakes and about how traditional that extract, transform, and load environments are encumbrances when data needs to tapped on-the-fly for any and all applications.

**Data warehousing is going to require fewer people to populate and operate.** As with many other elements of the data environment, data warehouses have increasingly become autonomous. These environments were originally designed to be run with as little DBA time as possible.

**Data warehousing is going to support AI and machine learning to deliver results.** Not only will data warehouses be the foundation of datasets for AI, but AI will also enhance the operations and capabilities of data warehouses. For example, Google has incorporated machine learning into its BigQuery data warehouse.

**Data warehousing is still going to occupy a central place in delivering the customer experience.** The heritage of the data warehouse is built on understanding the customer in new and profound ways. No other environment maintains data that is so vital to CX. Data warehouses have long been the established repositories for not only historical customer data and demographics, but also can be blended with real-time data streams to provide on-the-spot services and responses to customers.

The data warehouse—as a system, as a concept, and as a way to delivery insights about customers, markets, and operations—isn't going away anytime soon. Data warehouses are increasingly becoming an even more critical part of the digital world. ■

*—Joe McKendrick*

# Operational Analytics at the Speed of Business

**Accelerating analytics to** operate in-the-moment. From strategic decision-making to low-level operations and customer experience, your entire company must have up-to-date information and insights to keep pace with the speed of business. It isn't okay for your business to be waiting on daily batch updates.

## LEADERS NEED REAL-TIME INSIGHTS TO MAKE INFORMED DECISIONS

Technology innovations, customer preference, global economics and market changes are causing the environments in which companies operate to change quickly and dramatically. Business agility is a necessity to survive and thrive in modern commerce. Market opportunities are short-lived, and threats are more impactful than ever. For leaders to be effective in recognizing changes in the environment and make informed decisions that lead to favorable outcomes, they need not only complete and accurate data, but also current data, so they can respond to changes in the moment.

## MANAGEMENT NEEDS REAL-TIME INSIGHTS TO ACHIEVE PRODUCTIVITY, PROFITABILITY AND QUALITY GOALS

Sales, customer service, HR, finance, manufacturing and logistics—almost every business process in modern companies is technology-enabled. This can be good if the systems and people involved in operations are working smoothly together and everything is going well. Managers depend on data-driven insights about these business processes to understand operational performance, process quality and cost drivers, enabling them to see where problems exist that require attention.

## EMPLOYEES NEED REAL-TIME INSIGHTS TO DO THEIR JOBS EFFECTIVELY

Modern businesses are complex, with operations spread across teams, IT systems and often geographic locations. For employees to be effective in their individual roles, they must understand what is occurring in the other parts of the company with which they interact. Manufacturing employees and planners need visibility of the sales-and-order-management pipeline. Sales teams need visibility to delivery schedules and logistics. Customer-service agents need visibility of customers' orders. To manage this complexity and make informed, tactical decisions, these employees need accurate and real-time data insights.

## CUSTOMERS EXPECT REAL-TIME INSIGHTS AS A PART OF THE MODERN CUSTOMER EXPERIENCE

Employees and company leaders aren't the only people who have a need for real-time data insights. Modern customer experiences are highly automated, and customers expect the data they view on the company's Website to be current. Product availability, order status, shipping data and returns processing are where real-time operational data drive digital customer experiences. If there is a change, then customers expect to see the change reflected immediately—they have little tolerance for waiting until the next day for data to be refreshed.

Businesses evolve quickly, in big strategic ways and in small tactical ways. Real-time data and information insights are what enable all parts of your business to identify, understand and respond to changes quickly and decisively. Actian Avalanche—Cloud Data Warehouse Service that enables you to collect and harvest data insights in near real-time and at enterprise scale. This can help you accelerate your business-process execution, monitor and better respond to opportunities and threats and provide employees and customers with the data they need to be informed and effective. ■

ACTIAN
www.actian.com

denodo

# The Logical Data Warehouse as the New Standard for Data Analytics

**Data warehouses are** a great tool to consolidate data from a variety of operational systems to become the reference for corporate reporting. They are specifically shaped for analytics and run on specialized hardware.

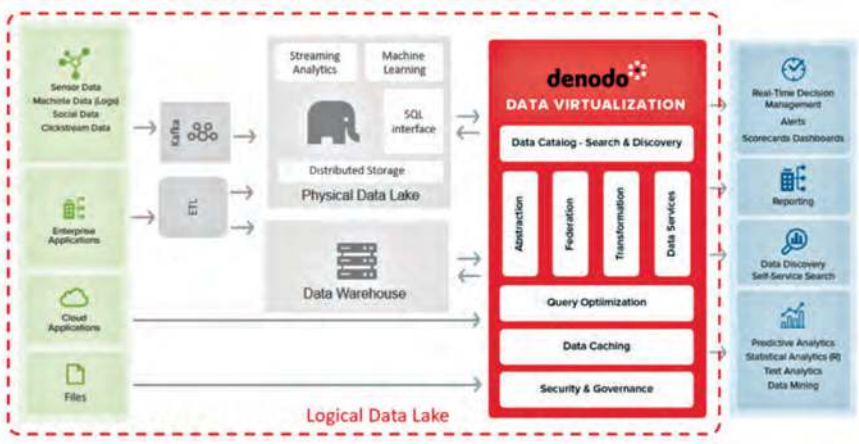
However, especially in the last few years, some of its core principles have been challenged:

- The rise of data driven decision making required storage of vast amounts of raw data. Traditional EDW appliances, with an elevated cost per stored byte, were too expensive. Cheaper distributed storage solutions (HDFS, S3, etc.) took the lead.
- The star/snowflake schema of an EDW is not the best way to store data for certain problems. Key-Value pair, graphs, and other NoSQL systems are designed to address specific challenges.
- Cloud vendors dominate the market. Specialized Software as a Service applications are the reference in many sectors and cloud mega-vendors are driving infrastructure to the cloud.

Although these factors provided huge advantages they also broke the premises of the data warehouse. The data landscape is fragmented, not just in location, but in shape and processing paradigms.

Physical re-consolidation, although possible, is less attractive than before. Volumes are too high, and replication to multiple systems creates brittle point-to-point connections. Out-of-synch data and uncontrolled replication leads to "data swamp" scenarios. End users pay the cost of a fragmented landscape in the form of extended time to market (or, more accurately, "time to data").

Thus, it seems that a logical approach is more feasible: a logical layer that connects different systems and exposes them as one. The complexity of the back-end systems is hidden from the end user. Security, governance and auditing are again centralized.




Data virtualization software like Denodo follows the ideas of relational databases. It provides a metadata catalog and an execution engine. It allows for the definition of derived views and data models. But unlike a database, it does not provide storage. Instead, connections to different systems will feed the data models in execution time. A virtual layer is focused on data delivery, not on storage.

How does execution work in a system like this? Underlying databases usually provide an execution engine, therefore, the virtualization engine takes advantage of them . This is called query push-down. It serves a double purpose: reduces processing in the virtual layer and network traffic. If all the data required for a query is in a single system, the virtual layer does the SQL dialect conversions and completely delegates the query to the source.

However, when data comes from multiple sources, the optimizer needs to come up with a multi-source execution plan. The plan is split into multiple branches that bring partial results from each source, and combines and aggregates them together in the virtual layer. Optimization techniques, although similar to those in relational engines, have evolved differently to deal with the nature of this problem. Techniques like complex query rewriting, on-the-fly data movement between sources, and MPP capabilities provide the processing muscle to perform efficiently.

The value propositions for these logical architectures is simple:

- There is one place to get data. Data exploration and "time to data" are greatly simplified
- Replication needs are significantly reduced, which reduces HW and operation costs
- Data governance is improved. Data is logically consolidated, traced to the source, and secured

As you can imagine, the benefits of a logical data layer go beyond warehousing and reporting, and can be applied to other scenarios like Logical Data Lakes to feed data scientists. ■

DENODO
www.denodo.com

SLAMDATA

REFORM

# The Missing Link in ETL

**ETL (extract, transform, load)** has been around for decades. Its primary purpose is moving data from source locations to data warehouses so analytics and data science teams can perform analysis across a range of critical data sources with standard tools. More recently, with the rise of low-cost cloud object storage, like AWS S3, Azure Blob storage, and others, this process has morphed into ELT (extract, load, transform). In this process, the data transformations are pushed further down the pipeline which somewhat streamlines the problems and also lowers overall costs. ETL/ELT tools have flourished in the last decade as the volume and variety of data sources that enterprises need to handle has exploded. However, there is one obvious gap in the solution space, complex JSON data, which coincidentally is also one of the most popular and rapidly growing kinds of data we see in the market. Unlike traditional relational or tabular data, JSON does not have a one-size-fits-all data model. In fact, it can range from very simple to unbelievably complex depending on the whims of the developers building the applications that create the JSON data. When existing ETL/ELT vendors say that they support JSON they mean they support VERY simple flat JSON. As soon as complexity goes up, they go down, and revert to the familiar approach: Start writing custom code! Some vendors don't even try to avoid code; they actually build a coding engine in their platform to handle complex JSON. So, the harsh reality is that complex JSON is the last ETL problem to be solved.

## BUT IS JSON DATA DIFFERENT?

JSON data is some of the most common data created today. Virtually all SaaS applications, Mobile applications, and IoT have JSON as the default data model. And the majority of Web APIs provide a JSON payload. JSON data is unlike traditional relational data in many ways, including non-fixed variable schema, variable data types, and the ability to have "nested" data structures. This presents a major hurdle when companies need to access this data for analytics purposes.

Analytics tools expect the data to be in a fixed tabular form (think spreadsheet) of rows and columns. In order to do this, the data needs to be transformed from the JSON model to the tabular model. Traditional ETL/ELT solutions cannot handle complex JSON well, if at all. To be clear, they all claim that they handle JSON, but what they really mean is that they can do some very simple things, and then require engineers to write complicated code to solve the rest. Most companies don't even bother trying to use commercial ETL/ELT software and simply have highly paid data integration engineers write custom code to transform their JSON data. This approach is slow, complicated, expensive, and not self-service in any way.

## FINALLY, A SOLUTION FOR TRANSFORMING JSON

SlamData REFORM is a revolutionary solution lets ANY user visually prepare analytics-ready tables directly on the JSON data, regardless of complexity. This means ZERO CODING and no waiting on Data Integration Engineers. Users can curate out custom data sets in minutes, and then iterate over them at any time as their data needs change. These tables can be streamed into all popular data warehouses, including Redshift, Snowflake, and Teradata, or pushed to any other destination you choose.

- Zero coding solution
- Any users, not just engineers can make complex JSON data analytics-ready
- Any data (JSON,CSV,XML), regardless of complexity
- More agile and accurate than custom coding
- Fast high-performance streaming engine for large amounts of data
- Adjusts automatically to changes in data

## WHO CAN USE SLAMDATA REFORM?

**Data Integration Engineers**—Makes their job easier, less coding, ability to respond to users' needs faster

**Data Architects**—Makes their job easier, less coding, ability to respond to users' needs faster

**Business Analysts**—Lets them have REAL self-service against complex JSON data (nobody else can really say this)

**Data Scientists**—Lets them have REAL self-service against complex JSON data (nobody else can really say this)

## FAST AND EASY TO INSTALL

SlamData REFORM is a software tool (so no added SaaS compliance or security issues) and is also available in the AWS Marketplace. Users can install the solution within their existing infrastructure and use it as they need, securely.

REFORM supports JSON data stored in AWS S3, Azure Blob Storage, Wasabi, and MongoDB. We can add a new connector to any JSON data source quickly with our advanced Lightweight Connector Technology (LWC).

Learn more about SlamData REFORM at http://slamdata.com or see it in action in this informative video. ■

SLAMDATA
https://slamdata.com/

# Next-Gen Data Warehouses to Power Intelligent Enterprises

**In today's digital** era, consumers around the world are driving organizations to transform themselves into intelligent enterprises by embracing technological innovation in artificial intelligence (AI), cloud, and Internet of Things (IoT). These innovations can radically impact businesses with adoption of right strategy to harness the power of data and analytics to aid digital transformation. The need of the hour is to move from a "system of records" to "actionable insights" through successful delivery of intelligent data platforms that can aid real-time analytics, providing the right data, on demand. The foundation of a successful, intelligent enterprise will be next-generation data warehouse platforms, which can enable any kind of data provisioning in a digitally disrupted world.

Traditional data warehouses served the need of descriptive analytics on core transactional systems capturing only 20-25% of all enterprise data. These warehouses cannot keep pace with business disruption and are a big impediment to agile business analytics and digital computing.

Some fundamental limitations to the traditional data warehouses include:

1. Increased operational risk and threat of data breach
2. Lack of scalability, affecting business agility and time-to-market
3. Increased latency issues as data volumes grow with complexity
4. Lack of accuracy in ROI quantification
5. Tightly coupled platform and integration affecting agility
6. Provisioning for structured data only

Today, data processing has become more evolved and complex with mobile, social media, cloud, machine, and sensor data integration. These new data sources have tremendous business value to be unearthed and monetized. Business need has evolved from descriptive/diagnostic to predictive/prescriptive analysis. This change in analysis is possible only when data is captured in its most native form through streaming, in near real-time, and merged with historical data amounting to massive volumes of data in terabyte/petabytes. Such volumes facilitate in-depth analysis and computing on a large scale to build various forecasting models, empowering businesses with actionable insight. Harvard Business Review Analytic Services recently published a report on the advantages real-time data and analytics can bring to an enterprise, helping to build a truly data-driven intelligent enterprise.

Next-generation data warehouses are on-demand, secure, and scalable self-service data centers that fully automate the provisioning, administration, tuning, backup, and recovery of data. This accelerates analytics and actionable insights while minimizing administration requirements. Next-generation data warehouses also provide real-time, complete access from surface-level analytics components to the core in-memory platform. This allows businesses to ingest and store structured and unstructured data, and also transform raw data assets. A complete portfolio of data exploration, reporting, analytics, machine learning, and visualization tools can be enabled on the data for accelerated analytics without replicating data. With next-generation data warehouses, organizations do not need an innovation-limiting, pre-defined schema that limits their ability to harness insights from available information.

## THE ADVANTAGES OF NEXT-GENERATION DATA WAREHOUSES

Cloud is the cornerstone for next-generation data warehouses, given the advantages in cost, scalability, performance, anytime/anywhere access, security, and ease of administration. Many enterprises have started their data-to-decision transformational journey enabled by hybrid, public, and private clouds. With the advantage of hybrid and cloud-native platforms, next-generation data warehouses are becoming smarter in all three dimensions—storage, computing infrastructure, and services. Additionally, built-in resiliency, enterprise-grade security, and protected data-sharing capabilities are making them intelligent enough to empower users for generating insights in a self-service consumption model. With the advent of AWS, MS Azure, and Google Cloud, immense business benefits can be realized that include:

- Creation of a data-driven customer journey, resulting in increased customer satisfaction
- Enhanced business agility and faster time-to-market, enabling improved and faster decision making
- Reduced infrastructure, maintenance, and admin overhead costs, resulting in improved ROI
- Anytime/anywhere access, enabling self-service BI capabilities
- Automation based on AI/ML

With the tremendous growth that analysts are predicting in analytical database management over the next three years, the next-generation data warehouse market will be shaped by the following forces:

- The emergence of data warehouses in the cloud or data warehousing-as-a-service (DWaaS)
- The need for data warehouse infrastructure to support big data
- Increasing demands for low latency and high-speed analytics
- The increased role of business intelligence in enterprise management
- The commoditization of data warehouse software and hardware

With the evolution of data warehouses in the cloud, it is time to take away the complexity traditionally associated with business intelligence infrastructure and democratize data. Next-generation data warehouses have the ability to truly enable a big leap forward in enterprises, allowing on-demand access to make informed business decisions. ■

WIPRO LTD.
Visit: https://www.wipro.com/analytics/
Email: ask.analytics@wipro.com