Modernizing the ENTERPRISE DATA WAREHOUSE with the CLOUD

Authored by Radiant Advisors

Sponsored by CLOUDERA

UNISPHERE RESEARCH

database TRENDS AND APPLICATIONS

Modernizing the ENTERPRISE DATA WAREHOUSE with the CLOUD

TABLE OF CONTENTS

Modernizing the Enterprise Data Warehouse with the Cloud

Authored by Radiant Advisors
Sponsored by Cloudera

The data warehouse has been serving companies in reporting and analysis for decades with data that has been extracted from an operational system, then cleansed and transformed in business-oriented data models for business intelligence tools. There have been debates over which best practices to use in building them, and there have been many technology optimizations related data integrations, analytic databases, and BI tools that use the DW environments. Expanded business analytics needs and mounting challenges for enterprise data warehouses also arise from increased data volumes, velocity, and the varied data types that may not come as stable or structured data sources.

These challenges create the need for a robust enterprise data analytics platform (DAP) – an environment that can sustain the ever-growing speed, amount, and variety of data types and where the enterprise data warehouse remains a critical component. Whereas Hadoop platforms pioneered concepts of scalability and flexibility a decade ago, the advent and acceptance of cloud computing platforms now make modern enterprise data warehousing a powerful reality. Cloud and the data warehouse both have evolved and converged into a single enterprise DAP that can take advantage of cloud managed services and cloud-native architecture while being deployed as hybrid and multi-cloud architectures. Three foundational concepts are relevant for you as you look at the future of your data warehouse: what it means to be "enterprise," what it means to be "modern," and what the future state of hybrid and multi-cloud mean for your environment.

Enterprise: The Role of the Data Warehouse in an Enterprise Data Analytics Platform

Enterprise data analytics platforms recognize and enable complementary analytics modes, including the data warehouse for developed and governed data models with business subject areas, metrics, and dimensionality. DAPs also facilitate self-service freedom in data discovery and integration, prediction with probabilistic analytics and data science in machine learning, and deep learning analytic models.

The term "enterprise" is essential to describe how the data warehouse, its data processing, and analytics delivery are intended to serve the analytic needs across the entire company and ensure that all data assets are managed with consistency, accuracy, and governance.

Within the DAP framework, the data warehouse is the component for well-understood and properly represented data that serves business users' specific reporting, analysis, and dashboard needs. As a supplemental component of the DAP, the data lake serves as an enterprise repository of all data assets and the foundation for data engineering, which can utilize streaming data hubs and pipelines to deliver data wherever it's needed in near real time. The data lake is the data warehouse's source of operational systems data from which to develop

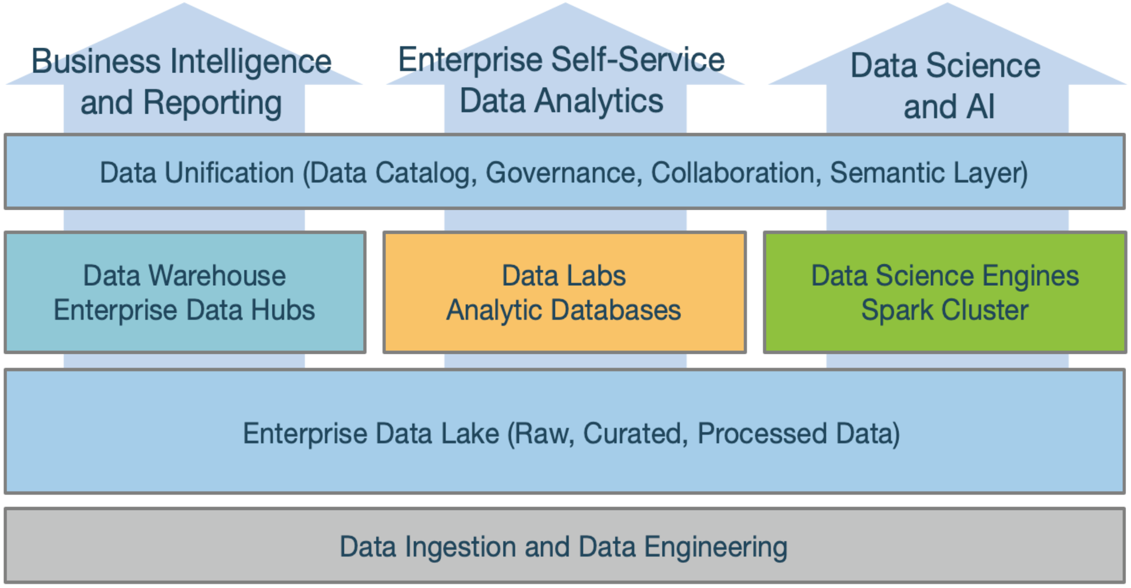

Business Intelligence and Reporting
Enterprise Self-Service Data Analytics
Data Science and AI
Data Unification (Data Catalog, Governance, Collaboration, Semantic Layer)
Data Warehouse Enterprise Data Hubs
Data Labs Analytic Databases
Data Science Engines Spark Cluster
Enterprise Data Lake (Raw, Curated, Processed Data)
Data Ingestion and Data Engineering

Figure 1: Data Analytics Platform Framework

production data transformations and business data models. A helpful way to conceptualize this is to think of the data warehouse as an island in the data lake. This is quite accurate; the enterprise data warehouse contains well-understood and trustworthy data for business management and decision-making based on operational data that can be found in the enterprise data lake.

Enterprise Data Analytics Platforms bring together and manage all of the data assets in the enterprise in order to efficiently enable the spectrum of analytics capabilities needed in each business area. Whereas the EDW development process requires clear and concise business requirements to design data models that produce certified reference data and metrics, the complete DAP establishes flexibility and agility for ad hoc and self-service analytics needs. When it's unclear how or which data can solve business questions, analytics capabilities for data discovery, exploration, and prepping capabilities empower business users to realize how data available from the EDW and data lake can support business decisions. These findings provide quick agile business value and can be analyzed further for governed long-term incorporation into the EDW if applicable. These business analysts can also act as citizen data scientists in preparing data sets that support data scientists with developing advanced analytics models faster. An enterprise data lake with operational data, behavioral and IoT data, external data, and user data supports robust data science analytic models that supply analytics and predictions for desired outcomes in complex business situations.

Modern: The Modern State of Analytics and the Data Warehouse

Enterprise data and analytics platforms need to continually evolve do support the agile, intuitive analytics needs that meets the needs of both developers and businesspeople. The adoption of cloud-native architecture and managed services supports modern data analytics for businesspeople to intuitively work with data analytics software with the data they need.

In the past, technology optimizations such as massively parallel processing and columnar databases for analytic optimizations of the relational SQL database allowed for incremental improvements in data warehouses. However, these were still constrained in cost-effectiveness and elasticity for scaling up and down with workloads due to their deployments on data center servers that operated with physically constrained integration of CPUs, memory, networking, and storage. Data warehouse planning required the purchase of servers and storage based on projected growth plans for several years with the hope to "grow into it" and not underutilize it. Either case was never optimally balanced or maintained over time.

By contrast, cloud-native architecture decouples storage from compute and memory resources in an elastic on-demand service. This allows for data stored in data lakes in open standard file formats such as Parquet or Avro to be affordably scalable. This also allows any data processing engine (such as RDBMS) to run on appropriately sized compute services whenever and for as long as it needs on the open format data.

A modern data warehouse leverages a cloud-native architecture and/or "data warehouse as a service" for data processing and providing access to the same trusted, governed data for business reporting and analysis. The managed services option for cloud-based data warehouse databases converts databases to on-demand, dynamically scalable, elastic versions without needing to maintain its virtual machine. Cloud-native data warehouses and databases persist data separately from groups of computing resources that access it and are dynamically scalable and on-demand. The cloud-native data warehouse can also leverage separate computing groups to isolate and protect workloads or user group service level agreements.

Modern data warehouses overcome the challenges that traditional data warehouses face due to the shared, fixed set of computing resources that are used for overlapping workloads, such as data loading, data transformations, and user queries. Further, some user queries consume a large percentage of resources, thus impacting and delaying other users' queries. With a cloud-native data warehouse, computing groups can operate independently for each workload to work on the same data without effecting the workload performance. This is ideal for data transformations that can be implemented with SQL statements to be executed in the database's compute resources without needing a separate data integration server and compute resources to read and transfer data over the network, process it in the data integration server, then send and write back to the database.

The Hybrid-Cloud Modern Data Warehouse

Hybrid-cloud architectures are, by default, the reality from day one in migration scenarios as companies move their existing enterprise data warehouses to a cloud platform. Unless the data warehouse is new and born in the cloud, the journey to a becoming a modern data warehouse begins with a hybrid architecture, and will mostly remain hybrid over time because a balance will take place as certain data and analytics operate best on-premises near particular data sources.

A variety of challenges exist in migration efforts, including appropriately prioritizing DAP architecture components to align with analytics projects delivery, where to adopt cloud managed services and cloud-native architecture, and assessing the shift of data persistence and workloads between on-premises data sources and cloud SaaS data sources.

Companies typically take one of two approaches when incorporating a cloud computing platform for their enterprise data warehouse, and each has its pros and cons. With either approach, a fully optimized modern data warehouse can take months or years and will continuously evolve for an optimal balance.

The migration strategy commonly referred to as "lift-and-shift" involves the least amount of changes to the existing on-premises data warehouse architecture and tools to be in the cloud. This approach is typically chosen when facing a pressing migration deadline. Here the priority is to get into the cloud quickly, then work in the cloud environment to distill the data warehouse environment into cloud services. This is a typically a three-step cloud migration strategy, beginning with the shift first, then optimizing with cloud services, then building out the broader enterprise DAP architecture. While leveraging the cloud as an Infrastructure as a Service model does not take advantage of the full potential and value of the cloud platforms, it does allow time to develop skills and competency for working in the cloud for DW/BI and IT operations teams.

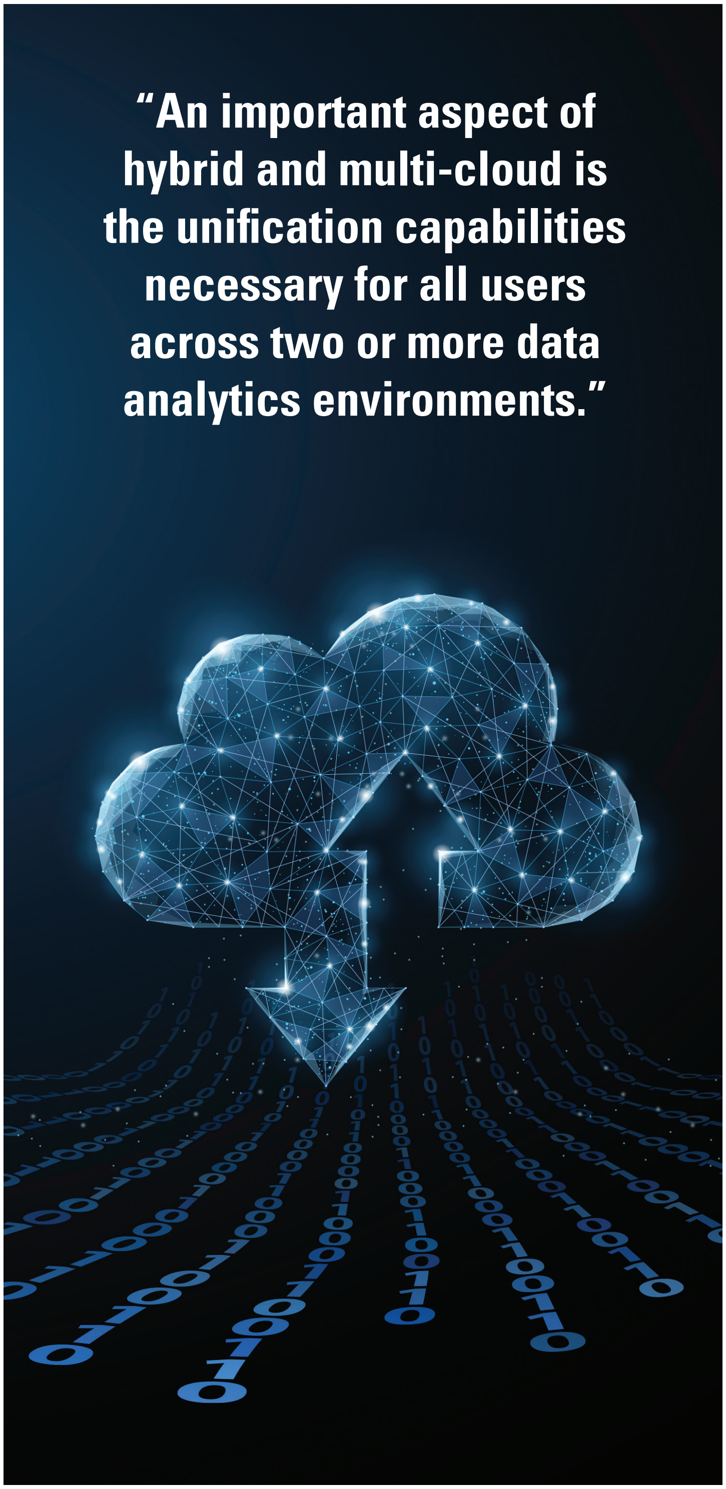
A second migration approach maintains focus on business analytics project delivery and leverages that to prioritize which components from the enterprise DAP to implement (such as data lake and ingestion) and which cloud services to migrate to for the long-term modern data warehouse based on cloud architecture patterns and standards. The business analytics projects can vary for end users to have a new dashboard, become empowered with self-service data analytics, or tackle a business challenge through data science. Established on-premises BI and data visualization tools remain on-premises for early projects until the data warehouse has a significant portion of its data in the cloud. During this time, data unification technologies such as data catalogs, data virtualization, and semantic layers will span both environments for users and play an important role in allowing them to access data regardless of where it is.

The Future: The Evolution to a Multi-Cloud Modern Data Warehouse without Limitations

A multi-cloud architecture can be a factor early in migration strategies, as different parts of the company may have already adopted various cloud platforms for their business systems or data analytics. However, for companies where this is their first evaluation and decision for a cloud provider, a multi-cloud mindset will provide perspective regarding the architecture and migration because it's likely that other cloud platforms will be a factor in the future.

Multi-cloud architecture considerations include a mindset for a globally oriented, distributed data architecture and naming standards. An important aspect of hybrid and multi-cloud is the unification capabilities necessary for all users across two or more data analytics environments. The top priority is to ensure that governance and security are centralized and then facilitated on-premises and in the cloud before being extended to include other cloud platforms. Secondly, as an enterprise platform, end users and developers will require unification capabilities that enable them to be efficient in working across the many data sets, data processing routines, and analytics that span multiple platforms. A data catalog can help people find data, collaborate to share information about the data, and establish galleries to leverage reuse and consistency of data analytics.

For companies that are concerned about "vendor lock-in" with any particular cloud provider, the multi-cloud architecture can be viewed as an



"An important aspect of hybrid and multi-cloud is the unification capabilities necessary for all users across two or more data analytics environments."

open architecture mindset. The challenge is that valuable cloud managed services may be specific to that cloud provider and not easily portable to other cloud platforms, which could potentially limit the optimization of the enterprise DAP. Some providers are beginning to introduce cloud services (such as Google Anthos, for example) that are made for execution on other clouds as well as their own, and a category called cloud management software focuses on the ability to deploy and operate software anywhere. There's also increased adoption of Kubernetes and Dockers frameworks to deploy containerization of software packages that can operate on any infrastructure. These self-contained software deployments typically require expanded skillsets for "full-stack developers," and it remains to be seen if MDW and DAP data engineers will adopt this approach into the mainstream.

Data ingestion, distribution, and persistence will require data architecture adjustments for a multi-cloud architecture that starts in a hybrid-cloud mode. Consider that in a hybrid-cloud architecture, data ingestion can be from on-premises data sources and SaaS application APIs into the cloud platform. While data may be ingested and flow into the cloud data lake initially, sending the same raw data or its curated output to other cloud platforms for end users to perform analytics can become costly due to cloud egress charges (a fee charged when companies send data out of the cloud platform). Careful consideration is required for modern enterprise data architectures regarding multi-cloud deployment, since restructuring large data volumes later could become costly and time-consuming.

Accordingly, architects should think about a distributed logical data architecture that allows for data lakes and modern data warehouses to reside in different cloud platforms depending on their data gravity with data sources, cloud services, and end users. Also consider data broker hubs that can be outside of the cloud platforms, on-premises, or in private data centers that can publish data to cloud platform subscribers without pushing extra data from one cloud to another. Apache Kafka is a good example of a data ingestion and publishing hub that can operate on-premises and in every cloud with the same configuration and portability (versus proprietary native cloud data streaming services and hubs).

Conclusion

As enterprise data warehouses evolve to become modern data warehouses in the cloud, they still hold a significant role for enterprise analytics as a vital component of an enterprise data analytics platform. The reality is that this evolution will be a hybrid-cloud architecture that requires shared and unified capabilities to represent both cloud and on-premises environments as a single data analytics platform for the business. A multi-cloud architecture will be likely for many companies as data gravity from more data sources, users, and applications shifts data processing among clouds, requiring open data architecture principles and furthering the need for enterprise data unification and governance.

About Cloudera:

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open-source community, Cloudera advances digital transformation for the world's largest enterprises. Store, analyze, and manage all your data in all its forms in a modern data warehouse wherever it works best for you. With Cloudera Data Warehouse you're in control. Run on-premises, in the public cloud, or any combination you'd like. With Cloudera the choice is yours. Visit the Cloudera Modern Data Warehouse Kit Hub to learn more.
Cloudera Modern Data Warehouse Kit Hub

About Radiant Advisors:

Radiant Advisors is an independent research and advisory services firm that delivers innovative, cutting-edge research and thought leadership to transform today's organizations into tomorrow's data-centric industry leaders.