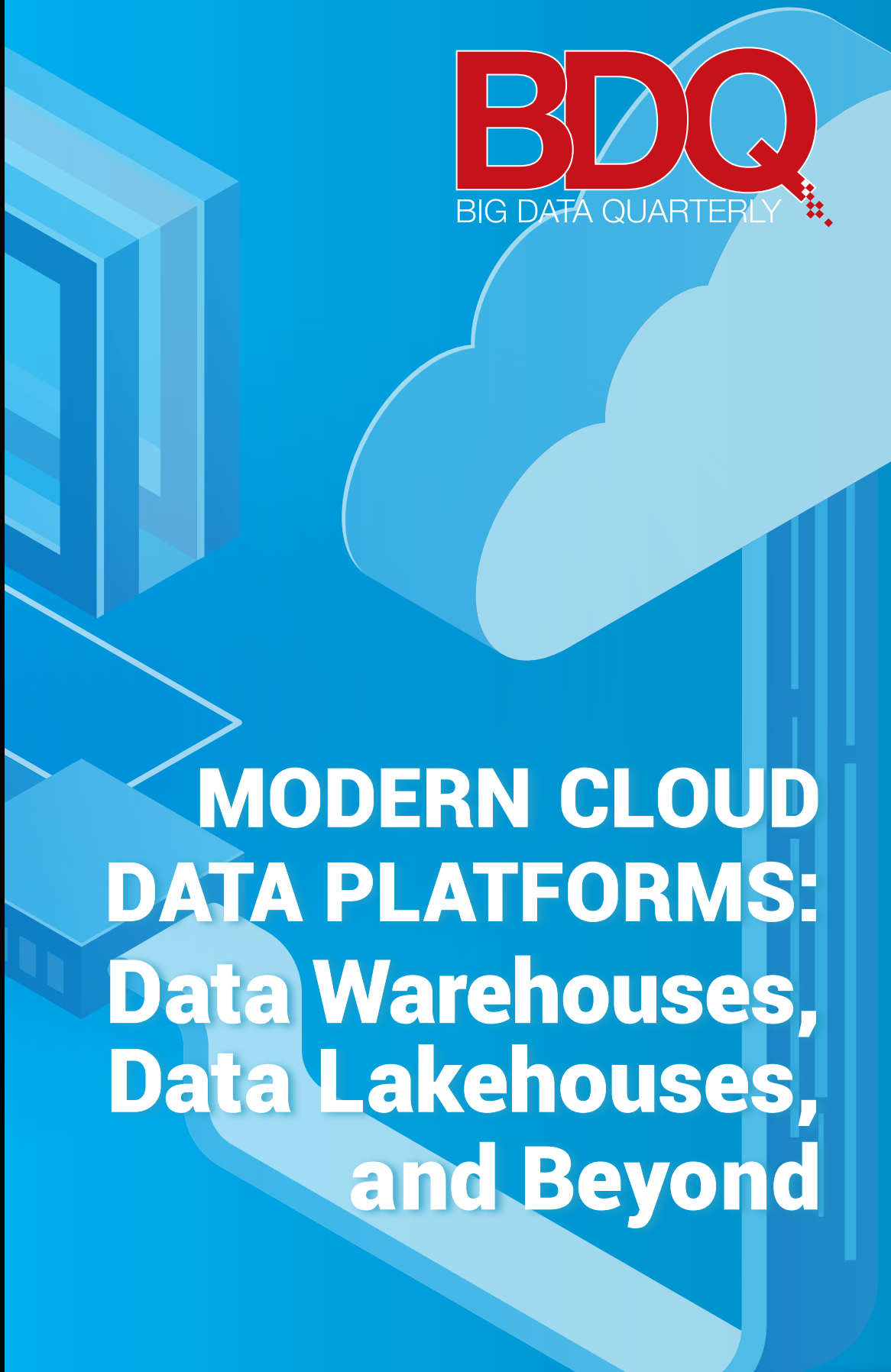
IRI

CLOUD DATA MIGRATION, GOVERNANCE AND ANALYTICS

BDQ

BIG DATA QUARTERLY



MODERN CLOUD DATA PLATFORMS: Data Warehouses, Data Lakehouses, and Beyond

Best Practices Series

BDQ
BIG DATA QUARTERLY

CONTEMPORARY Data Architectures

Without cloud, there would be very few AI and high-end analytics activities underway, with the exception of large organizations that have gigantic information technology and data management budgets. Now, thanks to the availability of fairly cheap and massive processing and storage power, relatively sophisticated data practices are possible for organizations of all shapes and sizes.

However, positioning a data architecture in the cloud is not an overnight process. Rather, it requires carefully considered choices and investments to determine where and how data assets should be managed and maintained. Chief among these considerations is determining how cloud-based data architectures will support use cases and data formats and be supported by infrastructure and available skills.

"Modernizing to a cloud data warehouse" is the next move for upgrading data architectures cited by a majority of 217 data executives (53%) responding to a recent survey by DBTA and Radiant Advisors. In addition, 51% are focusing on real-time data management capabilities. Another 43% are looking at implementing or upgrading a data lakehouse. At least 39% of respondents have approved budgets, and 24% have submitted budgets for initiatives related to modern data architectures in this study (Market Study: 2023 Modern Data Architecture Trends, Unisphere Research, 2023).

AI and machine learning analytics dominate business use cases, cited by 49% of respondents to the Unisphere Research survey.

The question is how to get from here (on-premises data architectures) to there (cloud or hybrid data architectures). Progress has been slow. "It is unsurprising that we find 30% of respondents planning to remain hybrid with on-premises and cloud operations. The cloud modernization initiative has been around for years, with adoption shifting from early adopters to mainstream companies," the Unisphere report states.

That's because there are caveats and precautions to making the choices between these architectural approaches. To date, progress in moving data warehouses and related areas to the cloud has been a slow journey, writes Barry Devlin, founder and principal of 9sight Consulting and author of the book *Cloud Data Warehousing, Volume I: Architecting Data Warehouse, Lakehouse, Mesh, and Fabric*. The proliferation of vendors offering cloud-based data platforms has resulted in a bewildering array of architectural patterns. The vendors tend to only work with their own definitions, Devlin says.

Still, data managers need to carefully investigate cloud-based data warehouses and lakes. Data lakehouses, now primarily cloud-based by necessity due to their scale, should focus on "data warehouse-like functionality in a cloud environment," Devlin cautions. The differences may be slight, he points out.

For example, in his book, Devlin notes that while data lakehouses may be the next shiny new objects in the data manager's world, they also engage the same cloud solutions, with a "data storage approach based on object stores with open-source table management tools on top." Some functions are built on Apache Spark; others are built on relational databases.

While common definitions are difficult to pin down, a recent McKinsey analysis offers suggestions that outline the distinctions

Best Practices Series

within the three key infrastructure approaches and the skills needed to work within these environments:

- **Cloud-native data warehouses:** Cloud-native data warehouses "are a high-performance, reliable, SQL-driven platform for business intelligence and reporting," according to McKinsey. "This archetype accommodates only structured data and offers little possibility to innovate beyond business intelligence and reporting. However, it does allow end users with minimal skills to customize output based on business needs."
- **Data lakes:** Data lakes "offer central, cost-effective, scalable storage for large volumes of structured and unstructured data," says McKinsey. "As the platform evolves, organizations have the flexibility to add analysis capabilities (such as streaming and SQL analytics). Data lakes require users to have a high level of skill and experience to analyze unfamiliar and unprocessed data."
- **Data lakehouses:** Data lakehouses "combine the advantages of a data lake's cost-effective, scalable storage with a data warehouse's reliable and performant reporting offering," the report acknowledges. "They provide central storage for business intelligence and SQL analytics as well as for data applications requiring unstructured or near- or real-time data."

Embracing such architectures within the cloud opens a path to "built-in, seamlessly integrated platform services, such as centralized monitoring and logging functionalities, scheduling, and orchestration," said McKinsey. "Furthermore, compute and storage capacity can be easily tailored to specific industries and enterprises, thanks to a variety of offerings, from optimized compute power to storage nodes."

The leading obstacles to cloud data migration include concerns about cloud billing (49%), data governance (48%), and security (48%), the Unisphere survey shows. In addition, "In the past year, other leading trends are responding to this with movements in data observability and FinOps as companies seek more transparency and control over cloud billing through cloud-controlled optimizations," the report states.

In the survey, data managers also express concern about migrating away from a legacy data architecture (42%) and its corresponding adoption of new technologies, platforms, and services (41%). "As cloud data migration has matured over the years, lessons learned have taught companies about the challenges with this migration and not to underestimate it."

Once underway, pursuing more robust data architectures in the cloud helps organizations move further along with initiatives such as analytics, data engineering, reporting, master data management, data catalogs, and governance, another industry report states. Organizations that are "cloud-powered" are further ahead with these data capabilities, according to PwC. About 10% of organizations surveyed by the consultancy "have reinvented their businesses through cloud, they report fewer barriers to realizing value and they're doing so at a rate twice that of other companies."

All of these cloud-powered organizations (100%) have more highly developed data, analytics, and AI strategies, the PwC report observes. "Executives at cloud-powered companies understand that data is at the heart of transformation efforts," the report's authors state. "And they need to address the all-too common issue of siloed—and mushrooming—data that is not only untapped but also ungoverned. Cloud-powered companies are much more likely to have an enterprise-wide data strategy than other companies (88% versus 59%). This means they develop a streamlined architecture to modernize their data into an integrated view, create governance structures, and concentrate on building the skills and operational changes needed to become a data-driven organization."

This gives them an edge with emerging initiatives, including AI and machine learning. "With their data in order, cloud-powered companies can turn their attention to leveraging machine learning and AI to reduce costs, add intelligence and get things done more quickly," according to PwC. "Cloud-based AI services can be used to wrangle data, reengineer processes, make automated decisions in real time and drive simulations to model different strategies and outcomes. To do this, companies will need the business and data science talent who can make it all happen and make sure it's done responsibly to reduce bias."

Once these pieces are put into place, it's conceivable that within a few years, data management and delivery will be an entirely cloud-based service. But this is very much an industry in transition, and new capabilities and approaches—made possible through the cloud—will continue to boost the capabilities and attractiveness of data warehouses, data lakes, and data lakehouses.

—Joe McKendrick

Cloud Data Migration, Governance and Analytics

IRI Voracity
An Insatiable Appetite for Data

THE TRANSITION TO CLOUD COMPUTING HOLDS THE promise of big data analytics, enhanced operational efficiency, and digital transformation. But it also presents a range of challenges, like locating and filtering data from multiple source systems and silos and understanding the conversion requirements for the cloud destinations.

Other non-trivial challenges include:

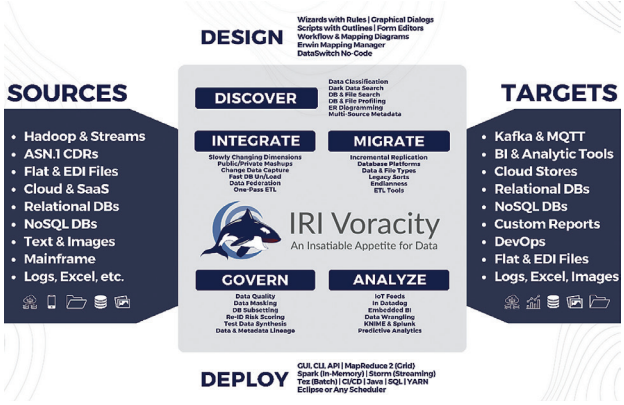
1. Data Transformation and Quality: with data in different formats and with redundancies, gaps, or inconsistencies. This diversity and disparity in data can render it unusable for analytics without substantial transformation and cleansing.
2. Data Migration: Migrating data to the cloud can be a complex, error-prone task, especially when dealing with large volumes of data.
3. Standardization Issues: Inconsistencies in data formats and schemas among different cloud services necessitate frequent modifications to data connectors or adaptors.
4. Breach Risk and Test Data: Ensuring data security and compliance with regulations like GDPR and HIPAA is a major concern in internet-based cloud services vulnerable to data leaks, hackers, worms, etc.
5. Workflow Speed and Complexity: New ETL implementations involving data cleansing and transformation are inconsistent across cloud providers and rely on slow (e.g., SQL) or complicated (e.g., Spark) engines.
6. Learning Curves: New databases, applications, and infrastructure require trained personnel to mitigate migration failures, compute costs, and security gaps.

Navigating these challenges in cloud migration and operations suggests a strategic approach based on incremental adoption, expert training, and simplifying technology. A single, ergonomic, and affordable data management platform that can support migrations to, and mission-critical data operations in, the cloud solves many problems.

The Voracity data management platform uses a familiar job design IDE and powerful data mapping engine to optimize and combine common tasks involved in the discovery, integration, migration, governance, and analytics of large and/or sensitive data in both on-premise and cloud silos.

The platform supports structured, semi-structured, and unstructured sources. Its functional versatility ensures that you can manage the data lifecycle more fully using software that's compatible with your infrastructure but customizable for unique needs:

Voracity's ability to handle complex data transformations and migrations makes it an ideal choice for transitioning to cloud data warehouses and lakehouses. At its core is the



IRI CoSort engine, which has made its bones for decades modernizing mainframe sorts, legacy formats, and ETL jobs while wrangling data for internal or external BI.

To address the 6 challenges above, Voracity delivers:

1. Automated, ongoing data cleansing and transformation for data accuracy
2. Speed and consistency in high data manipulation and movement
3. Standardized driver connections and metadata definitions for structured data in both on-premise and cloud systems
4. Built-in PII classification, discovery, and masking functionality for production data moving to the cloud, and multiple test data management features for prototyping
5. The decades-proven CoSort engine for manipulating and mapping large datasets
6. An ergonomic Eclipse UI with graphical wizards and workflow diagrams front-ending self-documenting metadata common to every structured data job

Voracity also aligns with modern data architecture objectives, including the production analytic platform, data mesh, and data fabric as it facilitates domain-driven design, data as a product, self-service data infrastructure, and federated computational governance.

It also scales large jobs vertically or horizontally to improve production data accessibility and test data/development agility. Robust data discovery, quality, masking, and subsetting features support data governance, privacy law compliance, and test data management initiatives.

Bottom line: As the digital landscape evolves, so does the complexity of data environments. The IRI Voracity data management platform shrinks the problems of multiple tools and big data, speeding your way into the cloud and the results you need from it.

IRI iri.com