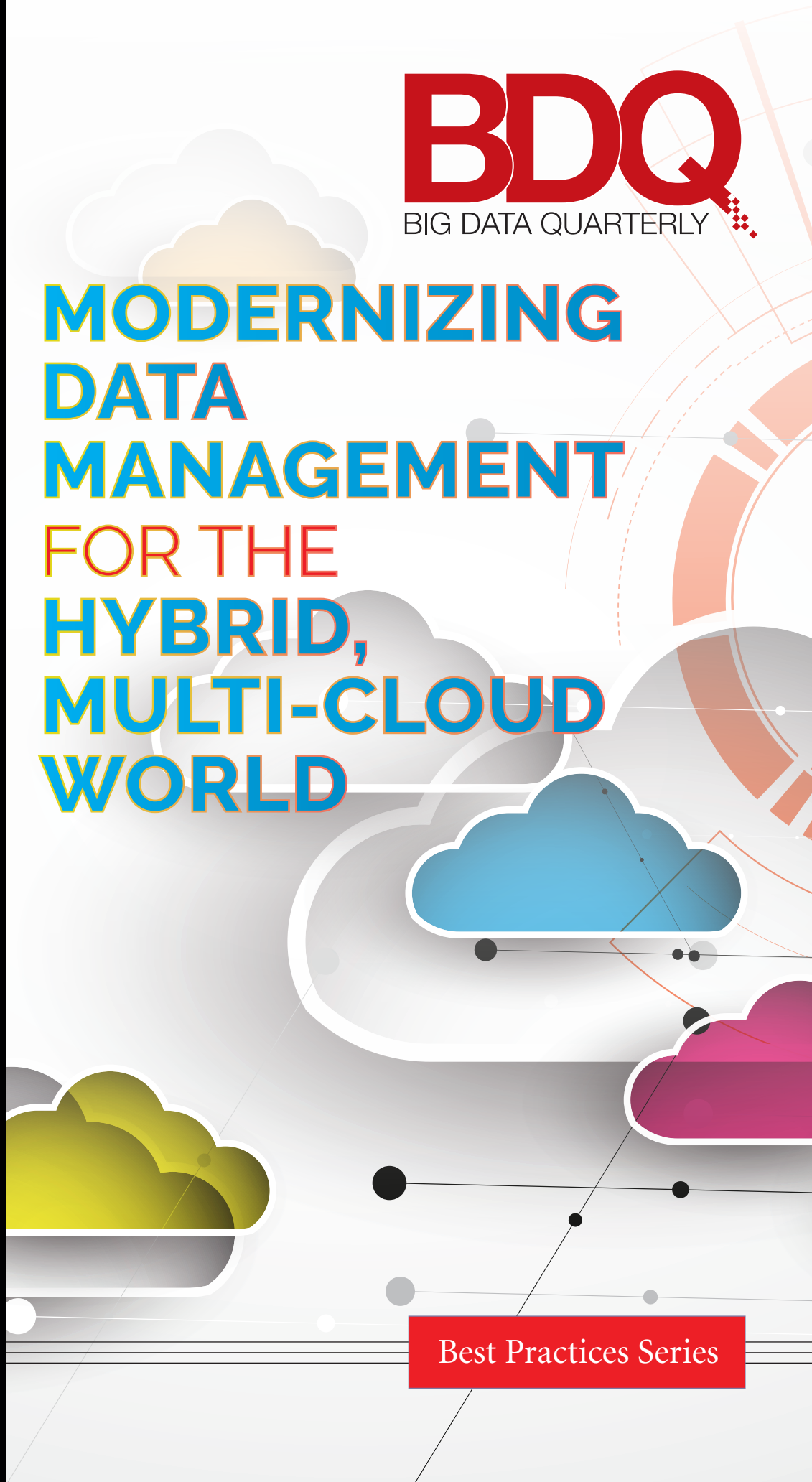
ChaosSearch

DATA LAKE VS. DATA WAREHOUSE: WHAT'S THE FUTURE-PROOF SOLUTION FOR YOUR BUSINESS?

BDQ
BIG DATA QUARTERLY

MODERNIZING DATA MANAGEMENT FOR THE HYBRID, MULTI-CLOUD WORLD



Best Practices Series

BDQ
BIG DATA QUARTERLY

ENHANCING DATA MANAGEMENT FOR NEW HYBRID, MULTI-CLOUD REALITIES

For advanced analytical capabilities, just about everyone is turning to the cloud in some form or another, as well as increasingly relying on cloud services for basic database management. The cloud is the foundation of the emerging class of data-driven enterprises that is actively and aggressively competing on data. Success with cloud, in all its many forms, however, requires greater vigilance and proactive initiatives to ensure data quality, governance, and effective integration.

Not too long ago, cloud mainly meant cost savings resulting from access to compute resources on a subscription basis. Now, it is seen as a gateway to support a wide range of sophisticated functions, from analytics to AI. In addition, cloud platforms offer greater flexibility since applications and data can be moved between platforms as necessary. A recent study by Unisphere Research, a division of Information Today, Inc., found that 66% of respondents now use or plan to use cloud as a strategy to reduce the time and money spent on ongoing database management activities, and 30% consider a cloud or cloud-like experience to be an important consideration in selecting their database infrastructure.

In a hybrid, multi-cloud world, data management must evolve from traditional, singular approaches to more adaptable approaches. This applies to the tools and platforms that are being employed to support data management initiatives—particularly the infrastructure-as-a-service, platform-as-a-service, and SaaS offerings that now dominate the data management landscape.

The bottom line is that data-driven applications such as analytics and AI require well-managed data to deliver value. Building these applications on hybrid, multi-cloud platforms is advantageous but also introduces new complexities. The problem that data managers and their organizations face is dealing with large volumes of a variety of data from sources, which are changing from week to week, if not day to day. For larger organizations, this can mean managing information streaming from hundreds, or even thousands, of data sources—especially with the increasing flow of data from devices and systems across the Internet of Things.

Data managers still need to grapple with many of the same challenges and tasks they faced before transitioning to cloud-based platforms. If anything, data volumes, varieties, and application workloads may increase significantly. As a result, data may still be dispersed widely across an assortment of cloud sites, on-premise systems, and devices.

The factors data managers must address as they move into cloud realms include the following:

Skills: Data managers and their organizations need to prepare for highly diverse scenarios. One of the most critical issues is having the skills to build and deploy these environments. Data scientists and AI developers are necessary to create models and algorithms that leverage or test datasets, but organizations also require data engineers and administrators to ensure the availability, viability, and quality of the data being fed into these sophisticated

Best Practices Series

systems. At the same time, organizations may need to hang on to the skills associated with the on-premise data environments that will be part of the picture for some time to come. Larger companies may have access to cadres of these various specialties, but many organizations do not have these skills in-house.

Data governance: Governance also needs to be reconsidered as enterprises move data and applications into hybrid and multi-cloud settings. With global regulations affecting how and where data is stored, data managers have to understand the physical architecture of their cloud providers. Other issues may be tied to internal corporate policies. This includes who has access to data and how data is to be used for specific customer groups. Some departments may even be using their own cloud services without the guidance of IT departments. Enterprises are basing more and more key decisions on data, and that data needs to be trustworthy, timely, and secure. This can become difficult as data keeps flowing in.

Data integration: Until recently, data was often managed manually, bound by written scripts and patchworks of interfaces. Along with automated capabilities, many cloud platforms and services automatically provide for connectivity and integration, reducing the need for such manual work. At the same time, the focus of data managers' work may move to keeping cloud services aligned rather than concentrating exclusively on internal applications. Data managers and their business

Data managers and their organizations need to prepare for highly diverse scenarios. One of the most critical issues is having the skills to build and deploy these environments.

counterparts need to sit down and design systems and infrastructures that allow for the rapid and free movement across platforms, whether they are on-premise, in the cloud, or a combination of both.

Data security: This is another area of concern as data moves into hybrid and multi-cloud environments. Moving to the cloud while still maintaining on-premise applications increases the attack surface for hackers and malicious viruses. This requires additional attention to backup and recovery processes to maintain data availability, as well as to enable data encryption.

Data availability: Data availability also becomes an issue as the move to cloud intensifies. The variety of places where data can land and be stored—on-premise, in the cloud, or on the edge—results in the need to manage multiple venues. From the perspective of the business, it's important that data be rapidly available from all platforms, and data architects and systems planners must design for instantaneous failover and recovery in such a way that any mishaps are invisible to end users.

Performance: Performance is challenging in hybrid and multi-cloud environments—both in terms of visibility and observability. Performance needs to be monitored across the various platforms employed, especially in terms of configurations, usage, and costs.

Tools: Part of the difficulty with managing on-premise environments along with multi-cloud offerings comes from the need to use multiple tools and platforms to manage end-to-end processes. Cloud services provide robust tools associated with their platforms, but there remains a need for data managers to align these environments to provide end-to-end flows of data and associated applications—so there are still manual tasks often required at this level.

A HOLISTIC APPROACH

Developing a holistic approach to data management is critical. Data applications are no longer confined to back-end data warehouses and data lakes, accessible only to in-house analysts and executives. Data is required to improve customer experience and perform predictive analysis to drive automated systems. Data governance and proactive management need to be built around a forward-looking strategy that maintains consistency across the enterprises. New players who can serve as business-focused data stewards must be brought into the equation to help organizations realize the value of their data assets.

—Joe McKendrick

Data Lake vs. Data Warehouse: What's the Future-proof Solution for Your Business?

DATA WAREHOUSES and data lakes represent two leading solutions for enterprise data management in 2021. Data warehouses were born in the 1990s as on-premise solutions, and in recent years have re-emerged in the cloud to support digital transformation. Data lakes were also initially built to run on premise, on Apache Hadoop, in the early 2000s. But with the rise of secure, resilient public cloud storage offerings from the likes of Amazon, Microsoft, and Google, they too have found new footing in the cloud.

Data warehouses and data lakes share some overlapping features and use cases, and both have embraced modern approaches to data management by operating in the cloud. However, there are fundamental differences in their data management philosophies, design characteristics, and ideal use conditions that should be considered as you develop your data management strategy.

WHAT'S THE DIFFERENCE?

A **data warehouse** is a data management system that provides business intelligence for structured operational data with clear and defined use cases, usually from a relational database management system (RDBMS).

Data warehouses follow a schema-on-write data model; source data must fit into a predefined structure (schema) before it can enter the warehouse, where it is then connected to downstream analytical tools that support BI initiatives. This is usually accomplished through an ETL (extract-transform-load) process. This connection between data ingress and the ETL process means that storage and compute resources are tightly coupled. If you want to ingest more data into the warehouse, you need to do more ETL, which requires more computation.

The data warehouse is all about functionality and performance. These functions are all essential, but the data warehouse paradigm of schema-on-write, tightly coupled storage/compute, and reliance on predefined use cases makes data warehouses a sub-optimal choice for big, multi-structured data or multi-model capabilities.

A **data lake**, on the other hand, is a centralized repository where multi-structured data from a variety of sources can be stored in their raw format. This encourages a schema-on-read model where data is aggregated or transformed at query-time. Bypassing the ETL process means you can ingest large volumes of data into your data lake with less time, cost, and complexity.

Data lakes provide a less restrictive philosophy that's more suited to the demands of a big data world: schema-on-read, loosely coupled storage/compute, and flexible use cases that combine to drive innovation by reducing the time, cost, and complexity of data management.

But data lake solutions don't inherently include analytic features. They're often combined with other cloud-based services and downstream software tools to deliver data indexing, transformation, querying, and analytics functionality. Without warehouse functionality, governance, or integration with known ETL or analytics tools, the data lake can become a "data swamp"—a murky mire of data that's impossible to sift through. It accumulates and sits stagnant because users don't know how to effectively access or glean insights from the data. Smaller datasets are duplicated and pushed to end user tools for analytics, creating silos.

Progress is being made, though. Today's data lakes are built on cloud object storage and can be activated directly to support multi-dimensional use cases including full text search, relational queries, and machine learning.

In fact, Gartner's Hype Cycle for Data Management, 2021 reveals that data lake technologies are poised to exit the Trough of Disillusionment and enter the Slope of Enlightenment.

According to Gartner, "A data lake, when designed properly, can provision data for the diverse exploration requirements of multiple user types and use cases… Today's data lake is on cloud, and it supports multiple analytics techniques (not just data science)."

CRAFTING A MODERN DATA MANAGEMENT STRATEGY

Neither a data lake, nor a data warehouse on its own, comprises a Data & Analytics Strategy—but both solutions can be a part of one. Enterprises continue to rely on a variety of solutions to meet their needs, including RDBMS, operational data stores, data warehouses and marts, Hadoop clusters, and data lakes.

While most of these solutions have been around long enough that their shortcomings are well-known, newer alternatives like data lakes are still reaching maturity and showing their potential for the future of scalable, flexible, and resilient data management in the cloud.

Across the board, a modern data management solution must be cloud-native, simple to manage, and interconnected with known analytics tools to deliver value.

KNOW BETTER™ WITH CHAOSSEARCH

At ChaosSearch, our goal is to help customers prepare for the future state of enterprise data management by bridging the gap between data lakes and data warehouses. ChaosSearch activates the data lake for analytics; We publish analytic APIs that a data warehouse would also provide, indexing data within your cloud storage environment, rendering it fully searchable, and enabling analytics at scale. With its revolutionary approach delivered in a fully managed service, ChaosSearch overcomes the cost and complexity of competitive solutions, delivering unlimited scalability, industry-leading resiliency, and massive time and cost savings.

ChaosSearch
www.chaossearch.io