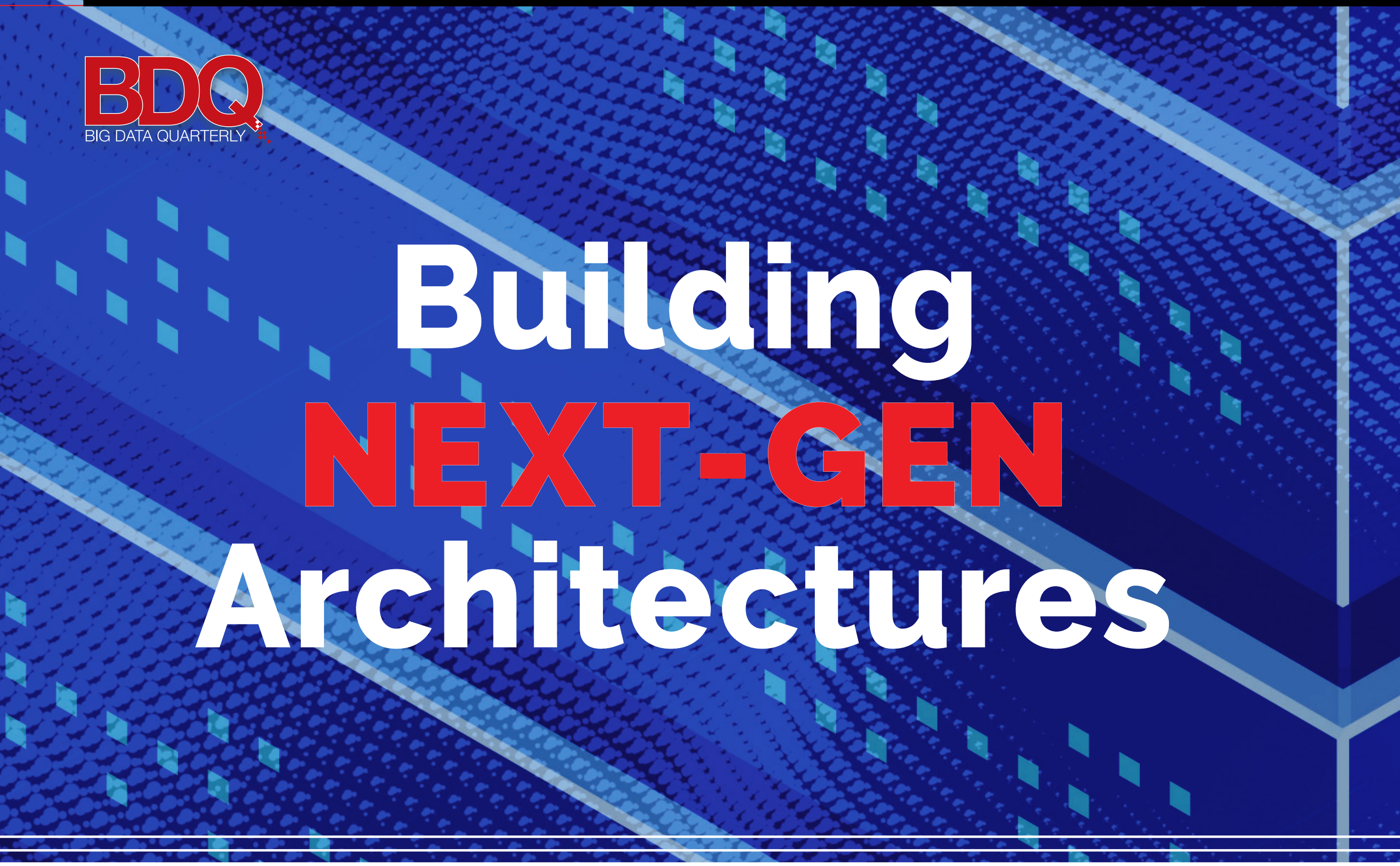
IRI

WHAT DATA LAKE SUCCESS REQUIRES

BDQ
BIG DATA QUARTERLY

Data Warehouses, Data Lakes, AND Next-Gen Architecture

Best Practices Series



Building NEXT-GEN Architectures

Data architecture has come a long way over the past few years. Enterprises are focusing on data analytics to power just about every essential function, from delivering customer services to managing production. As organizations seek to design, build, and implement data and analytics capabilities, they are pressed to reinvent and reorient their data architectures—as well as justify these activities with return on investment measures. The path to next-generation data architectures means adopting approaches such as cloud, data virtualization, and open repositories that enable users and applications to view needed information from anywhere, across or outside the enterprise.

Within next-generation enterprise data architectures are a range of solutions, including cloud data warehouses, data lakehouses, data fabric, and data mesh. The objective of these approaches is to enable highly agile and responsive data environments that evolve with the fast-changing needs of today's organizations. Ultimately, new architectures may finally break down the silos that have defined enterprise frameworks for years, if not decades—but were originally created for a previous purpose. An effective next-generation architecture should support seamless data intake, storage, real-time processing, and accessibility across enterprises—as well as addressing compliance.

Many companies are now cautiously eyeing these technologies, a survey of more than 200 enterprises by Unisphere Research, a division of Information Today, Inc., shows. Adoption of next-generation data solutions—data lakehouses, data fabric, and data mesh—will better position organizations to embrace AI, machine learning, and Internet of Things networks. At the same time, it means greater complexity and skills issues ("The Move to Modern Data Architecture: 2022 Data Delivery and Consumption Patterns Survey," May 2022).

There are many definitions of next-generation data architecture, along with many variations, but the common denominator is a commitment to employing the latest methodologies and platforms to manage and analyze data in a more cohesive way. Moving to a next-generation data architecture is a journey, not an overnight sprint. Here are some considerations for succeeding along the way:

Recognize that legacy data architecture is going to be prominent for some time to come. While the promises and capabilities of next-generation enterprise data architectures can be extremely compelling, a majority of enterprises in the Unisphere Research survey, 82%, are still employing more traditional data warehouse solutions to support business-critical applications—and 36% said *they rely exclusively on data warehouses* and had yet to implement other kinds of data environments. Emerging components of next-generation architecture—data lakehouses, data fabric, and data mesh—are still relatively new on the scene. Only 43% plan to increase investments in data lakehouses, 47% for data mesh and fabric. Still, among respondents already using these environments, 94% will ramp up investments in data lakehouses and 92% will boost their data fabric and data mesh spending.

Best Practices Series

There are many definitions of next-generation data architecture, along with many variations, but the common denominator is a commitment to employing the latest methodologies and platforms to manage and analyze data in a more cohesive way.

Reorient to the cloud. While many enterprises still support on-premises data warehouses, the shift to the cloud is happening at a rapid rate, and this is transforming data architectural approaches as well. While 73% of enterprises in the Unisphere survey indicated they were running data warehouses on-premises, only 34% indicated they would remain entirely on-prem within the next 3 years.

Rethink your data culture and processes. The ultimate business goal of a next-generation data architecture is to facilitate the transformation from a static database environment to a data-driven enterprise. The technology to make this transition is available and ready, but a data architecture is built on culture as much as it is software solutions. This is a process that begins with making decisions that have implications for decentralization, scalability, and building a data-driven culture. This requires new ways of thinking about data—at all levels. For example, who is responsible for the datasets supporting the enterprise—and how will the data be used? Deciding the appropriate tools will be next; for many organizations, data management and architectural planning is still confined to data management or IT departments, with sporadic input from business users.

Understand the different aspects of next-generation components. For example, while data fabric is a centralized approach that supports existing data environments, data mesh is a highly decentralized self-service structure. A data lakehouse supports the schema and structure enabled with data warehouses, while also supporting both structured and unstructured data, as does a data lake. Data fabric is intended to centralize data resources from across various sources, while mesh is decentralized. Though cloud applications have been on the scene for well over a decade, data lakehouses, data fabric, and data mesh are only at the beginning of their adoption stages.

Recognize there is no single, one-size-fits-all approach for the journey to next-generation enterprise data architecture. Moving to next-generation data architecture will be an uneven process. Every business has its own unique needs and ways of doing things and likely has investments in its own arrays of technologies. Larger organizations will have more resources and bigger technology budgets for such transformations than their smaller counterparts. In addition, some organizations are already more advanced with data analytics solutions. Many organizations, as shown in the Unisphere survey, have existing data warehouses that are moving to the cloud and may build out from there. In other cases, organizations that have not invested in data warehouses may adopt data lakes and data lakehouses that provide repositories for both structured and unstructured data. Some organizations are built around centralized departments where skills—data engineers, data scientists, and analysts—are concentrated accordingly. Moving to a decentralized environment may require new skill sets.

Recognize that transforming to a next-generation data architecture is a journey, not an overnight sprint. Evangelizing and planning with executives and managers across the enterprise are part of an ongoing, deliberate process. These new approaches tend to introduce a great deal of complexity, which may introduce challenges in their design and implementation. Transforming processes and associated technology means working closely with the business to identify its needs, the data needed by the business now and into the future, and getting buy-in from stakeholders and executives across the business. In addition, organizations moving to next-generation data architectures will need to actively recruit and train for essential skills. Needless to say, this is a process that may take months to see through.

Incorporate data security and governance into the design and planning of next-generation enterprise data architectures. Overhauling a data architecture with new configurations and systems means introducing potential exposures and risks to sensitive data. That means rethinking and enhancing security measures, which invite more complexity. In addition, depending on the industry, organizations may be subject to laws and regulations that affect the way data is deployed.

—Joe McKendrick

IRI Voracity
An Insatiable Appetite for Data

What Data Lake Success Requires

THIS ARTICLE EXPLAINS WHAT A DATA LAKE IS, and how you can fish it for value in an architecture optimized for your needs. IRI Voracity users should also be able to assess this approach relative to other paradigms the platform supports, including LDW, EDW, ODS, and self-service BI.

WHAT IS A DATA LAKE?

Data lakes are (local or Hadoop) file system, or cloud, environments where data is gathered and stored for experimentation. They consist of both raw and transformed enterprise (source) data that can be used for reporting or analytics without, per Gartner, "the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse)."

Some want the data lake to replace the traditional data warehouse, while others see it as more of a staging area to feed data into existing data warehouses. Regardless, consolidating siloed sources can increase data sharing and information use, and reduce the costs of hardware and software holding data now.

Companies that build successful data lakes can mature their lake as they figure out which data and metadata are interesting to their organization.

RED FLAGS—GOVERNANCE AND EFFICIENCY

Data lakes also provide the "opportunity" to not have to prepare and protect all the data an organization gathers, and instead, just "save it for later" when new needs or ideas arise. Many organizations also turn their data lakes into data graveyards because nothing is tracked on the way in. Others can't even find the right data in the first place, collaborate on needs and access security, or refine the data through quality or lineage.

Accordingly, governing data in a data lake is important, which means, dealing with veracity, security, and metadata lineage issues, to name a few. The other issue is performance. Most tools and data interfaces cannot ingest, process, or produce information in an unmanaged lake. Thus, the consistent semantics and CoSort engine in Voracity would help.

STOCKING THE LAKE

The Voracity data connection and curation platform can help you populate a data lake by connecting to, profiling, and collecting data from different LAN or cloud sources. During or after movement, you can select, transform, reformat, and report on data from those sources using jobs defined by scripts, wizards, dialogs, or diagrams in Eclipse.

FISHING THE LAKE

You can only consider what experiments to run on lake data after you discover what you have. Voracity discovery tools profile, classify and diagram data, and find data using RegEx patterns, fuzzy/dictionary matches, NER/OCR models, and structure filters.

After identifying worthy data, even data scientists can struggle deriving value from it without semantic consistency or managed metadata. It is much harder to manipulate or analyze data without those.

Voracity wizards auto-create metadata and use it in ETL, federation, masking, reformatting, and/or reporting jobs that filter relevant data from the lake, transform it into useful information, and display it.

DE-MUCKING THE LAKE

Recall that a key problem with data lakes, as with real lakes, is that people don't know what's in them, or how clean they are. And if there's no structure or understanding around the data, it's also hard to envision its potential.

Lake data should thus be organized and cleansed, so it is ready and trustworthy. Voracity can catalog data and enrich it with data quality features like filtering, unification, replacement, validation, regulation, synthesis, and standardization.

Finding, classifying, and de-identifying personally identifiable information (PII) in the data sets is also critical. Voracity provides data security through rule (and role)-based masking functions applied separately or during data transformation, that preserve realism, referential integrity, and reversibility as needed.

MANAGING GROWTH

Administrative management of the data lake through persistent, centralized, shareable, and modifiable metadata is very helpful. And if you can automate processes that prepare and report on data, you can re-run what-if analyses to improve results faster.

Voracity simplifies metadata management through automatic creation, self-documenting syntax, and Git integration for lineage, security, and version control. Built-in task scheduling allows you to sequence—and fine tune the repetition of—integration, cleansing, masking, reporting, and/or other jobs you might want to run on lake data.

IRI, The CoSort Company
www.iri.com