MODERN DATA PLATFORM PLAYBOOK SERIES

FOR **KOGNITIO ANALYTICAL PLATFORM**

Playbook

Radiant Advisors

This edition published August 2014.

Table of Contents

Executive Summary

CIOs and chief enterprise and information architects are fast realizing that reference architectures and best practices of the past have served well, but are challenged to meet the business demands of today's data intensive and analytical environments. Companies require economical scalability for big data, but want high-performance; they require information discovery and flexibility, but want to govern semantics and enterprise consistency; they want to benefit from advanced analytics and unstructured data, but also want broad accessibility with SQL. Despite the mega-hype of big data, business analytics, and the data scientist, there is real business value and competitive advantage to be gained in these data technologies and skills.

Based on the same fundamental set of data management principles that created past reference architectures, Radiant Advisors' Modern Data Platform (MDP) is a framework for a new reference architecture that meets today's challenges. The MDP strategy incorporates accepted and emerging technologies that allow existing data warehousing (DW) and business intelligence (BI) environments to transform in an agile, iterative process of adopting, integrating, and growing a powerful data platform.

By aligning the strengths and unique differentiators of the Kognitio Analytical Platform with the MDP framework and principles, enterprise and information architects can cultivate a strategy that enables big data and advanced analytics capabilities for the business in ways that are clear and planned within their roadmap. This *Kognitio Playbook for Modern Data Platforms* focuses on understanding the MDP, the Kognitio technology, and its role within a big data strategy to enable today's companies to transform into tomorrow's competitive data-driven organizations.

Modern Data Platform Core Principles

Understanding and accepting MDP first begins with accepting the principles it is based on. For data management principles, this goes beyond data quality, consistency, and security. Rather than absolute "thou shalt" black and white, principles operate in a spectrum of gray and often come with trade-offs. MDP strives to avoid "this or that" approaches and recognizes "this and that" instead. In this section, we explore three interrelated principles that are fundamental for MDP: persistence, semantics, and access.

Taking a position on these three fundamental principles yields the Modern Data Platform framework and strategy.

1. Expanding the variety of different data store technologies is balanced with consolidating instances, clusters, and data stores.
2. Governing and centralizing semantic context away from local physical data stores improves consistency, agility, and navigation.
3. Enabling self-service data access yields the highest value to the business, and transitioning from SQL-based to service-oriented will further increase agility and consistency.

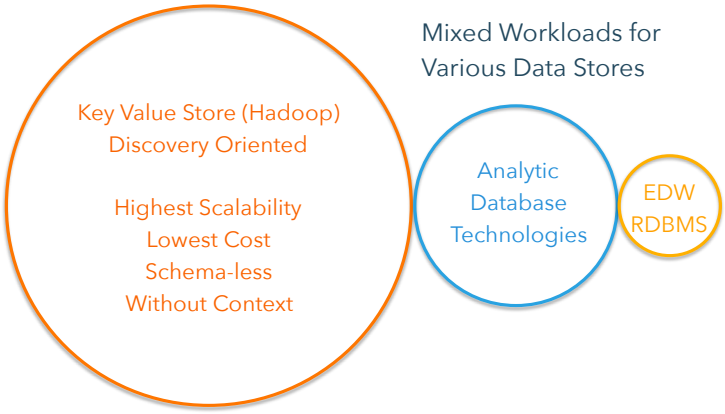
Evolving to the Modern Data Platform takes time, depending on your current environment, number of databases, data growth rates, user counts, and analytic maturity. There is a MDP Transformation Strategy that guides architects through a set of agile principles and techniques for adopting and enabling technologies while minimizing risk and disruption. A company with no service-oriented architecture (SOA) strategy or data virtualization tool will need to decide the path best for them, and centralizing first and relocating second could be a good strategy.

Polyglot Persistence for Analytics

Persistence is a fundamental principle for adopting and managing heterogeneous data technologies to meet the analytic capabilities required of data-centric businesses with information management as a core competitive asset.

Modern Data Platform Core Principles

Recently, the term "polyglot persistence" has emerged to recognize that different data stores (i.e. engines or technologies) have different strengths, and choosing where to optimize -- or persist -- data should be decided based upon the data's primary use. This principle -- sometimes referred to as a "best of breed" approach -- is a major principle within MDP as it embraces trade-offs for environment manageability, data duplication, or navigation. Proponents of the "all-in-one" approach, instead, argue the benefits of manageability, compatibility, stability, consistency, and navigation; however, MDP does not share the perspective that a single data environment is optimized to meet all the use cases for reporting and analytics by an enterprise. Still, while some IT organizations' firm requirements for IT vendor governance for standards, partnerships, or volume discounts may trump a best-of-breed approach, it does not invalidate it.



Accessibility:	Programming	SQL, MDX, UDF	SQL Access
Workload:	Flexible, Scalable	Analytic Optimized	Reference Data Mgmt
Maturity:	Emerging	Accepted	Mature
Structured:	Less	More	Most

Figure 1 - Polyglot Persistence in BI and Analytics

The polyglot persistence principle strikes a careful balance, focusing on the strength of a given data store for its assigned data set. Within MDP's three distinct data classes, each class can have multiple heterogeneous vendors of the same data technology (e.g. there are more than one RDBMS vendors widely in use, but they are all the same relational database technology). Consolidation of physical data stores within a class can still make sense for minimizing complexity, but should share common purpose and service-level requirements. MDP embraces

the strengths of different database types while striving to minimize the number of instances and data movement.

Semantic Context for Information Governance

The unification of databases has traditionally been referred to as "federation of data." The current trend with big data environments (like Hadoop) is to enable masses of SQL-based users and BI tools by making data more SQL-accessible. However, SQL access requires a data schema to be defined. In the simple key-value structure of Hadoop, schema is defined in the abstraction layer above the raw data for the user either in Hive or HCatalog. Programmers can access data by writing MapReduce programs that have data context hard coded into them.

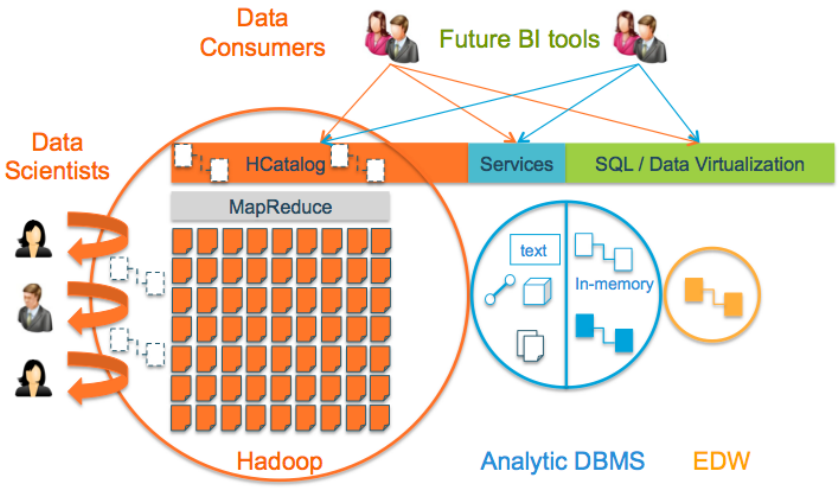


Figure 2 - Semantic Integration at User

Semantic context is a second fundamental principle within the MDP framework to be centralized into an abstraction or virtualization layer or repository (the separation of logical context from physical data). This fulfills several purposes. First, this distinguishes a singular "data platform" to work with from its singular access point. Second, abstracted semantic context has the benefit of metadata definitions, and, therefore, is easy to create and adapt where data itself is unknown or volatile. Finally, centralized semantic context allows for easier management, consistency, and user navigation at the enterprise level.

Modern Data Platform Core Principles

The separation of logical context from physical data happens in several patterns: schemas are defined as tables and columns, and data is organized and loaded into them; views are created in databases mapped to existing physical tables; or virtual tables are created outside the physical database. For Hadoop environments, this is the default as key-value stores are, by nature, schema-lite.

Ultimately, to unify all of these variations, a simplified, consistent, and unifying approach needs to be adopted.

Accessibility and Self-Service for the Enterprise

Today's dominant data access tools are SQL-based. Therefore, seeking to maximize value from the data platform typically equates to how quickly SQL access is provided. Next-generation Hadoop distributions are making data stored in Hadoop more easily SQL-accessible and familiar to users, but this still won't cover all NoSQL data stores within MDP, such as native graph databases or document data stores.

Abstraction is not limited to SQL-based data virtualization technologies: the addition of powerful NoSQL data stores and workloads requires a programming interface, or API layer. MDP embraces a data services layer as the most ubiquitous and complete way to centralize access management, as well as encapsulate semantic context and embed SQL access, if available. This is in line with an SOA strategy that seeks to decompose applications into service-oriented messages for orchestration and reusability, specifically data and analytic services.

Centralizing access can be chosen as different degrees of separation from the data itself. First, would be views within the same database schema accessing remote native SQL databases or SQL interfaces to Hadoop. Second, would be to centralize access outside of the database in data virtualization technologies accessing heterogeneous SQL databases and data services. Third, is to have a data services layer, and fourth to centralize access in the application layer -- such as BI tool catalogs and meta layers. Finally, the fifth is at the user level, where they access any database, define their own context, and share or collaborate result sets and discoveries.

Analysis of Kognitio Strategy

While some of today's newer data technology companies benefit from fresh new mindsets, others benefit from years of deep experience in a specific area, nurtured by embracing and evolving alongside industry trends and technologies.

At a time when RAM prices were exorbitant, Kognitio engineered an in-memory database that successfully solved performance challenges in databases, believing that price would eventually become more economical. Kognitio then spent years solving complex issues of in-memory scale-out and complex query concurrency that still challenge some vendors today. Kognitio expanded to the US market beginning in 2005, when businesses were readily embracing massively parallel processing (MPP) and scale-out database architectures as part of the data warehouse appliance trend. Kognitio's engineering head start allowed it to easily offer the Data Warehouse-as-a-Service (DaaS®) hosted model that later evolved into Kognitio Cloud. A decade later, Modern Data Platforms require the combination of these data technologies and service-oriented delivery to enable big data and business analytics for critical competitive advantages in a data-driven world -- whether large enterprise or small-medium enterprise (SME) scale business.

Terminology

Kognitio encapsulates its engineering and service abilities as the Kognitio Analytical Platform. Because the MDP embraces polyglot persistence, the adoption of best of breed data technologies designates the Kognitio in-memory MPP Analytical Platform as capable of delivering high-performance data analytics.

Being an analytics and capability driven architecture is a core principle within MDP. Class 1 emphasizes those data technologies whose strengths are affordable scalability with the flexibility to handle volatile and unstructured data; Class 3 emphasizes EDW and MDM that leverage data technologies from the entity-relationship and normalization paradigm for reference data management from relational database management systems (RDBMS). Some EDW/MDM solutions operate on RDBMS technology that is SMP from the transactional world, and others have embraced RDBMS technology that is from the MPP analytics world. However, because of primary purpose and strength, the Modern Data Platform separates the EDW/MDM implementations from more analytic intensive data technologies.

Analysis of Kognitio Strategy

Class 2 is reserved for analytic-optimized best of breed data technologies, like Kognitio, that efficiently deliver the time-sensitive analytic capabilities required by the business for higher levels of optimization, insights, innovation, and risk avoidance that play a part in today's data science and discovery environment. This class is subdivided into SQL-based (optimized for analytics) and NoSQL (specialized for analytics) categories. Kognitio fits into the structured schema-based (and widely accepted) SQL category and paradigm.

Caveats in the analytic database world run deep, with nearly every optimization trading off with something else. Benefits gained from columnar-oriented data play against the challenges of other workloads, such as data loading and more complicated SQL statements. Key-value stores that are basically schema-less benefit from flexibility and scalability in storing data, however this plays against the challenges of performance in data access -- and in the HDFS case also with the ubiquitous nature of SQL access. OLAP cubes are consistently fast performing, yet struggle with requiring hours to load and pre-calculate data -- and with scalability.

Kognitio's engineering heritage has refined in-memory database with MPP scale-out to provide both performance and scalability, while still allowing the falling economics of DRAM to leverage the price performance gains cited in Moore's Law for large-scale implementations. Kognitio, too, has focused on a row-based data architecture, citing that data load latency and throughput is vital for an agile data analytics platform supporting both SQL and not-only-SQL processing.

By retaining the ability to execute complex SQL for advanced analytics and models with row-based orientation, Kognitio provides a powerful combination of the performance, scalability, and ease of use required for advanced analytic capabilities. The MPP architecture itself is also leveraged for analytics written in a variety of modern analytics programming languages with full parallel execution from simple SQL data access statements.

Analysis of Kognitio Strategy

Advanced Analytic Capabilities

Data Science and advanced analytics have received much hype and attention over the past few years, yet remain ill-defined in both process and role for driving business value. However, business analytics driven and data science processes leverage data as a way to solve real business problems and create new opportunities through a highly-iterative process that requires high-performance in combining big data sets from Hadoop with strong reference data sets from an EDW or MDM.

Data discovery and visual discovery are new analytics capabilities being added to the BI spectrum -- alongside reporting, monitoring, analysis, and decision support. In discovery there are two unknowns to be discovered (or insights to be gained): semantic context and an analytic model. The challenge has been to first define and model semantic context for data then to extract, transform, and load into it. This process has been used for the development of EDWs -- whether in traditional BI project management or agile BI development. Discovery is based on "discovering the context," and therefore working with data requires a defined abstraction layer and approach. Both of the above are time-sensitive, as modern business is dynamic and changing, identifying context and defining models needs to be as quick as possible to extract the highest information value from the data as information has an ever shorter shelf life. Users need a freeform approach with on-the-fly access to raw data and a suitable analytic platform to rapidly create in-memory data models and experiment with variants and iterations of analytic models to tease out high value answers.

Advanced SQL capabilities are critical for analytic databases

"Kognitio fully supports the core features on ANSI SQL:2008 enhanced with optional features present in the most recent SQL:2011 standard ...along with MDX and excellent compatibility with Oracle SQL." -- Kognitio Technotes

Architecture Roles for Kognitio

Advanced analytics and data discovery are key analytic functions that benefit from the combination of a data technology that incorporates in-memory, MPP, and fast to load, row-based simplicity offered by technology such as Kognitio. As a Class 2 analytic-optimized technology, several potential architectural roles exist for Kognitio within the Modern Data Platform framework.

Kognitio as the Big Data Accelerator

Very similar to how generalized RDBMS-based EDWs leveraged OLAP MDDB cubes in order to meet performance requirements in analytics, the high scalability and flexibility of Hadoop data stores will be complimented by feature rich, easy to use, scalable in-memory databases that compensate for Hadoop's inherent weak response-time performance and coding complexity. So, the first role we define is a two-classed architecture with a scalable, robust storage class orders of magnitude larger than its high-performance delivery class.

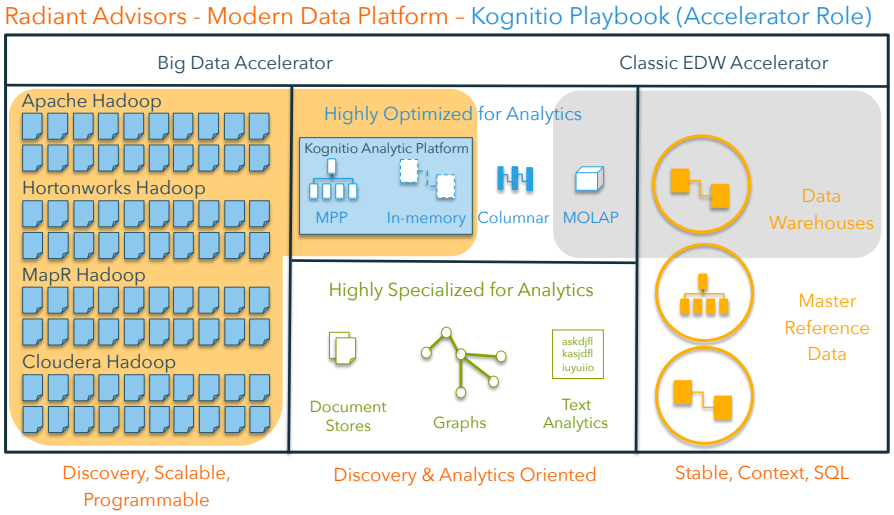


Figure 3 - Kognitio Big Data Accelerator in MDP

Projects that require increasing performance in Hadoop implementation can quickly embrace an MPP in-memory technology, tightly coupled with Hadoop MapReduce and HDFS, for a seamless user environment. The challenge, however, will be dealing with the capacity difference of the in-memory database compared to the Hadoop cluster user data volume. As for data discovery and visualization

Architecture Roles for Kognitio

tasks, sets of data need to be rapidly brought into RAM for assessment to guide development of a refined data model. Quality Hadoop integration pushes that filter access down into Hadoop and turns a single user command into a complex MapReduce job underneath the covers ensuring the analyst does not have to be a MapReduce programmer as well. Ultimately, how large a chunk of data is assessed is determined by need and budget sizing the in-memory component.

Kognitio as the Analytic Sandbox

A new and important component of the Modern Data Platform performs the role of an analytic sandbox, or discovery environment. The discovery process itself requires a database that is extremely high-performance and advanced analytics capable in order to enable data scientists, business users, and data analysts to work quickly and intuitively in a highly iterative and SQL-driven fashion to gain insights and discover data relationships and possible new business value. Data warehouse appliances began popularizing this role via marts nearly a decade ago as an independent environment wherein business units could work with more data more self-sufficiently.

With its in-memory database performance capabilities and feature rich analytic SQL and parallel R capabilities, Kognitio is well suited as an analytic sandbox. Kognitio offers two forms of tightly coupled Hadoop data integration with parallel data loading into memory – data is pulled into RAM via Kognitio SQL.

If the files are well organized within the HDFS via directory structures, Kognitio can parallel load files straight into memory directly from the HDFS, providing low latency access. With larger data sets in Hadoop, Kognitio integrates via a MapReduce connector to specify the projection or selection of data elements via SQL filters that get pushed down into the Hadoop store. This is an active integration, but since the underlying MapReduce is batch-oriented, these tasks take longer to execute and are best suited to large filtered fetches.

On an equivalent basis Kognitio offers ODBC integration for existing RDBMS and data warehouse appliances, allowing data to be pulled into Kognitio memory via

Architecture Roles for Kognitio

simple SQL commands. These external tables are pinned in-memory as views and are treated as read-only snapshots that require periodic refresh process. For data stored in the local Kognitio storage system, tables and associated memory images are fully updatable and are synchronized between DRAM and local persisted disk for resilience. View images are similar to standard database views, but with the result of the view object pinned into memory. This use of views aligns neatly to the accessibility and self-service capability defined earlier in this playbook.

Typically there is more than one analytic sandbox within a company. On average, five to six analytic sandboxes for various business units or departments may exist, each with their own Hadoop clusters -- and sometimes with different distributions from different vendors. Here Kognitio's heritage as a DaaS and MPP product is a valuable asset: with MPP database architectures, the more nodes that hold data in a MPP array, the better the performance is -- and it is a linear performance gain by nature. However, keep in mind that companies may also have a separate production analytics environment for operational analytic and data services that require higher levels of availability and fault tolerance.

Kognitio as Production Analytics and Data Services

Once a new advanced analytics model has been developed it must be promptly operationalized to return value to the business. By promoting it to the production analytics environment, a suitably scaled platform can be brought to bear, including on-demand cloud deployments for agile on-demand utilization. This is separate from the discovery and data science processes that require high-performance and scalability for iterating through rapidly changing sets of data quickly to test hypotheses or insights.

MPP database and in-memory architectures also benefit from the fact that they can scale very easily and quickly by simply adding as many additional servers nodes as necessary to maintain the operational capacity of ever-increasing analytic models in production. In the past we have seen companies generate a few to a dozen analytic models per year. Today, hundreds of analytic models are being developed and operationalized per year. Commensurately, businesses tend to start with a few

nodes of analytical platform and rapidly scale to tens or more of nodes as insight flows and time-to-insight drops.

Kognitio as the OLAP Virtual Cube

Though OLAP cubes and MDDB technologies have historically served as the performance accelerators of EDWs, in-memory databases -- and especially those that have MPP scalability -- are quickly supplanting them in the Modern Data Platform. Atomic level business performance metrics and multi-dimensionality can efficiently be loaded in-memory for high-performance analytics without the penalty of pre-calculating aggregates at load time. Over time, we predict that companies will increasingly adopt in-memory databases while decommissioning existing OLAP cubes.

Kognitio has the unique capability to assist with the over-time migration of existing legacy OLAP cube infrastructure to newer in-memory database technology. The Kognitio Cube Builder Tool allows for designing logical OLAP cubes that can be made available for MDX language via ODBO or XLMA without disruption to existing applications. Some EDW architectures that support Relational OLAP will be able to load table images or fragment table images into memory for cube-like performance.

Architecture Pattern for Kognitio

In the Modern Data Platform, data integration and data flow patterns are just as important as the framework of analytic databases. While many data flow patterns can be evaluated and adopted over time, data virtualization, in-database federation, or MapReduce programs can replace legacy ETL programs.

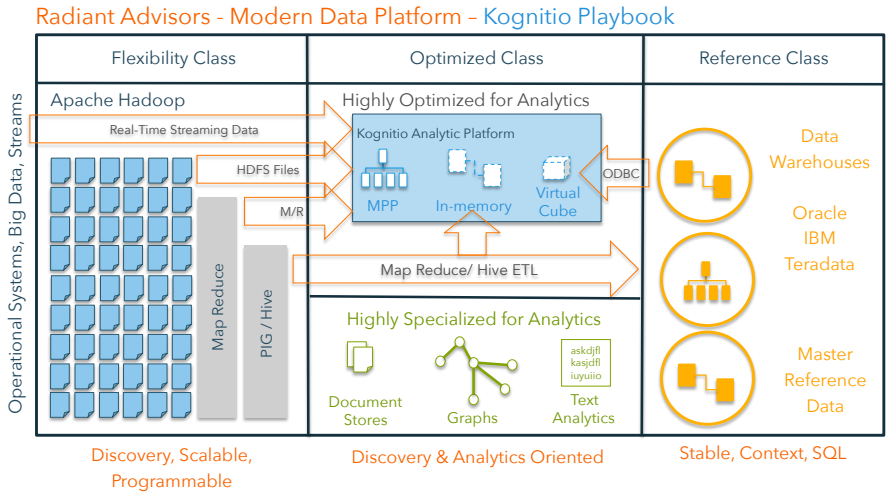


Figure 4 - Modern Data Platform with Kognitio Analytical Platform

With the addition of Hadoop into the Modern Data Platform the typical flow of corporate data is changing as data flows into the data lake. The Kognitio Analytical Platform simply reaches in and draws data from the lake via its HDFS and MapReduce connectors with no recourse to complex ETL steps or Hadoop programming – this smooths and speeds the overall flow of data and thus improves the time to insight.

Additionally, some EDWs have been online archiving structured historical data into the Hadoop cluster. This data remains available to users via data virtualization technologies or semantic layers within BI tools to union both data sets when need. Kognitio can tap into this zone of the data lake as well and pull historical data on-demand into memory alongside recent operational data for comparative analytics or time-based analytics.

Conclusion

The Kognitio Analytical Platform plays an important role in the Modern Data Platform. With the introduction and rapid maturing of Hadoop to ingest and persist vast amounts of data and the incumbent enterprise data warehouse, advanced analytics requires a number of optimizations to the relational database management system, or new specialized data stores like graph databases. Optimized for analytics databases include MPP data warehouse appliances, columnar databases, in-memory database, and even traditional OLAP multi-dimensional databases. The robustness of SQL varies from basic reporting SQL, to multi-dimensional expressions, to the advanced analytics features found in SQL 2011 on to the execution of modern preference analytics languages. Kognitio is a rare find with scalable in-memory technology and wide range of analytics capabilities from SQL through to MPP execution of R, Java, and other languages.

Modern Data Platforms require a balanced approach of diverse data technologies with the ability to manage a complex environment. Kognitio can not only meet the future needs of accelerating the Hadoop workflow, but it can also simplify the environment by eliminating the need for legacy and difficult to maintain MOLAP cubes, too.

There is a major drive towards solving the lack of performance and SQL accessibility of Hadoop environments to unlock the value of big data. Timeliness and low latency now count for a lot, and the ability to iterate and experiment with data is vital. Hadoop needs complementary best of breed of technology as defined by polyglot persistence. Strong SQL access is required to quickly enable many users with minimal re-education, and basic Hadoop Hive is limited, slow performing, and thus frustrating. Enterprises and SME business are also more cost-conscious and look to low-cost analytics solutions, which need efficient MPP acceleration for big data use cases.

Some of the newer distributed in-memory technologies arriving may take years to mature, while others like Kognitio, with its historical investment in engineering, can be implemented today on a production basis.

About the Author

John O'Brien, Principal Advisor and CEO, Radiant Advisors

With over 25 years of experience, John O'Brien is a recognized thought-leader in data architectures. As principal advisor and CEO of Radiant Advisors, he guides Radiant Advisors in providing research, strategic advisory services, and mentoring for companies in meeting the demands of next generation information architectures, and emerging technologies.

About Radiant Advisors

Radiant Advisors is a leading strategic research and advisory firm that delivers innovative, cutting-edge research and thought-leadership to transform today's organizations into tomorrow's data-driven industry leaders.

Boulder, CO USA
Email: info@radiantadvisors.com

To learn more, visit www.radiantadvisors.com