Informatica

White Paper

The CDO's Guide to Intelligent Data Lake Management

Nine principles for delivering accurate and consistent insights

ABOUT INFORMATICA

Digital transformation changes our expectations: better service, faster delivery, greater convenience, with less cost. Businesses must transform to stay relevant. The good news? Data holds the answers.

As the world's leader in enterprise cloud data management, we're prepared to help you intelligently lead—in any sector, category or niche. To provide you with the foresight to become more agile, realize new growth opportunities or even invent new things. With 100% focus on everything data, we offer the versatility you need to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption. Not just once, but again and again.

Table of Contents

Executive Summary 4

The Promise of Intelligent Data Lake Management 5

Why Data Management Matters for Your Data Lake 5

How Data Lakes Add Business Value in Practice 6

Top Nine Design Principles for Data Lake Management 8

Conclusion 12

Executive Summary

Transformative businesses and progressive data leaders use data lakes to drive disruptive business insights and deliver data-driven digital transformation outcomes. Thanks to the efficiency and scalability of data lake environments, your business can quickly discover new insights and make possible what never existed before.

A new set of technological capabilities and organizational practices, also known as intelligent data lake management, are forming the basis for maximizing the value of data lakes. By approaching the data management of big data in a systematic way, you will manage a data lake project that is fully automated with significantly fewer manual resources. A systematic approach to data management not only provides you with the confidence of consistently accurate insights, but it also paves the way for the data organization to win the political capital necessary to secure budgets and expand the scope of your efforts.

It's time to challenge yourself to think differently about big data. This white paper gives Chief Data Officers (CDO) and other data management leaders, like you, the guidance you need to get the most out of data lakes. We will show you the key design principles to manage data lakes that deliver results–not just once, but again and again.

What is a Data Lake?

A data lake enables you to store and process all of your data (including big data) from multiple data sources (cloud, on-premises, and hybrid) without having to pre-structure it. You can fill your data lake with all data types—structured, unstructured, or multistructured—which means your business leaders and axnalysts can drive more innovative analytics from more data.

The Promise of Intelligent Data Lake Management

Data management principles are the foundation for delivering trusted data to the right people at the right time. They ensure all critical data management processes, from collecting data to preparing and governing it, are fully automated for the data lake. Moreover, an intelligent machine-learning based approach to data management enables businesses to address a greater volume and variety of data with radically less manpower than traditional approaches. Your organization can now turn more raw data sources into trusted and timely sources of high-value and fully trusted information for more relevant analytics.

Why Data Management Matters for Your Data Lake

Data management is an even greater concern in the data lake environment than it is in more traditional data environments. There are a couple of reasons for this.

The need for speed

- As competitive pressures and the pace of business increases, there is a need to get access to data faster than before.
- New types of data are inherently time bound, such as data about customer service deterioration, and have very short time windows for effective remediation.
- Antiquated waterfall processes of the past are too slow for the modern needs of business.
- Complex requirements gathering processes and long development cycles only cause delays for lines of business to get the insights they need.
- Antiquated hand-coding methods that require the continuous hiring, staffing, and retraining of large and growing pools of headcount inhibit long-term maintainability and burden teams with rework and churn.

The need for self-service

- The business context around datasets is inherently more distributed than before.
- Data consumers within the lines of business are also now data producers providing their own datasets for enterprise projects.
- Unnecessary control by IT causes delays for the line of business to get the data they need.
- With business conditions changing quickly, business context is increasingly necessary for data projects.
- The teams that manage and consume data are often distributed.

The legacy of traditional data management paradigms, combined with the extreme efficiency of data lakes, can leave organizations with a proliferation of so-called “data swamps” that do not lead to sustainable value for the organization at large. CDOs must adopt a systematic and repeatable platform and set of processes to facilitate both agility and collaboration. A systematic approach is the only way that big data sources can be consistently turned into useful insights.

How Data Lakes Add Business Value in Practice

Data lakes are being used across industries and business functions to turn raw data into new sources of profitability. Knowing what's possible with data lakes in real-world scenarios will help your teams better understand the business context and build a data lake that delivers real value to your business decision makers.

Fraud data lake

Data-driven banks use data lakes to identify fraud or risks of anti-money laundering across a variety of threat vectors. Banks collect general ledger and other financial data from mainframe systems for easier access. They then combine it with external datasets, like watch lists. Financial organizations will identify anomalies and trends using advanced analytics approaches to fraud detection. When security and governance are built into the fraud data lake architecture, data analysts get trusted insights from more data without incurring additional risks–thanks to security and governance capabilities built into the architecture. This affords data analysts a trusted 360-degree view of risk across the organization, allowing them to deliver safer and more profitable services.

Marketing data lake

For years, marketing data has been spread across organizations making it impossible for marketers to get a full understanding of their prospects and customers. With a marketing data lake, marketers finally connect the dots through self-service access to an integrated, end-to-end view of their marketing data across all platforms, silos, and channels. Everything marketers want to know about their customers' and prospects' relationships with their companies is analyzed for trends and patterns—to better cross-sell and up-sell new products to customers. Increasingly these data lakes are being used to predict the "next best step" to lead customers though the buyer's journey.

Healthcare data lake

Healthcare organizations have more data and types of data than most other industries and must conform with tough industry and data privacy regulations (i.e. HIPAA). Data governance and security are critical to any healthcare company working with large amounts of sensitive data across departments and with partners. A healthcare data lake managed by intelligent data lake management improves collaboration for managing the myriad of healthcare data used in big data analytics to predict patient outcomes, lower healthcare costs, and optimize staffing and supply chains. By collecting patient data from hospitals and combining it with insurance data, healthcare organizations are building a single view of patients and making better decisions that reduce the cost of care while delivering better patient healthcare outcomes.

Industrial data lake

Manufacturing organizations need to drive profitability and lower risk by running more efficient supply chains. Sensors attached to machines generate data that is used to develop a more unified view of equipment health to predict potential maintainable failures and proactively fix them at much lower costs.

Compliance data lake

Financial services organizations are under pressure to demonstrate compliance of many regulatory rules, such as GDPR. Ingesting all compliance related data into a data lake enable compliance analysts to more holistically ensure and prove compliance against regulatory requirements.

Top Nine Design Principles for Data Lake Management

As a CDO planning a data lake initiative, there are nine design principles to consider in order to maximize the value of your data lake environments. Let's discuss how to do this in detail.

1. **Leverage a small, agile cross-functional BizDevOps core team to execute the project**

   There's a lot of talk about agile development, but the biggest missing link is around the importance of cross-functional teams. There are several benefits to using cross-functional teams with data lake projects. First, is the ability to integrate functional domain knowledge from multiple sources. Data lake projects require implementation knowledge from data engineering, business context from data stewards, as well as analytical expertise from data scientists and analysts. Having multiple perspectives encourages the timely development of accurate and consistent business insights that effectively meet business demands. It also ensures that everyone is aligned to a common understanding of the data available. Data lakes with cross-functional team ownership generically achieve project objectives faster and with fewer data quality issues.

2. **Empower your data scientists to quickly get the data they need to assist in data preparation**

   Self-service visualization tools, like Tableau, Qlik, Kibana, and Zoomdata have become very popular over the years as more business analysts seek to have direct access to data. But, visualization alone leaves business users waiting for the data they need from IT. This is where self-service data preparation comes in. Self-service data preparation provides the ability to allow knowledgeable business analysts to merge, transform, and cleanse relevant data into more trusted and certified forms prior to analysis. Sophisticated tools enable users to publish their prepared datasets back into collaborative workspaces, so that multiple business stakeholders can access and prepare the data together. Furthermore, machine-learning techniques within tools provide a guided experience for business analysts as they explore and discover data in the data lake.

3. **Use the wisdom of crowds with crowdsourcing and tagging to govern data assets**

   With organizations adopting data lakes for processing sensitive data, like patient or consumer data, effective methods of data governance are necessary. With data being increasingly distributed across the organization, self-service methods by data users themselves need to augment traditional IT driven methods for data governance. Crowdsourcing of data governance enables the organization to tap into the wisdom of business users for the knowledge, context, and expertise that collectively enhances the quality of the data. In a self-service environment, every user has the power to apply their subject matter expertise to improve the quality and structure of data. As an example, business analysts should be able to contribute their knowledge, through tags and other classifications, so that key data elements and data assets are continuously increasing in quality. Collaboration then becomes a mechanism for enabling business analysts to help one another towards a common enterprise-wide goal of delivering trustworthy data assets.

Machine learning is also an approach to automate data domain discovery with classification algorithms. Machine learning can also use clustering analysis to proactively discover similarities between data sets. This enables the system to automatically treat new data in the way similar past data was managed, relieving data management professionals from repetitious work. Analyzing the knowledge and behavior of the crowd also significantly increases the effectiveness of data governance.

4. **Automate the collection and transformation of data**

   Manual ingestion and transformation of data is a complex multi-step process that leads to unrepeatable and inconsistent results. There is no business benefit to bottlenecking the collection and transformation with manual or complex processes. Successful organizations take advantage of pre-built connectors and high-speed data ingestion platforms to load and transform datasets into the data lake. This enables data lakes to quickly accommodate new types of data and scale to increasing volumes of incoming data. Automation also accelerates the fast iteration and flexibility required for agility because changes can be made to automated processes very quickly without any risk of bugs.

5. **Leverage rule-based data validation and data scoring to identify data quality issues early**

   As executives know, problems not caught early cause larger downstream issues. Likewise, with data lakes, data quality errors not identified early, dramatically affect business insights due to inaccuracies or inconsistencies between different data assets. Artificial intelligence (AI) applied to metadata and data profiles automates data quality processes and business rules. Data lakes with rule-based data validation, infused with AI, automatically detects and corrects incomplete, inaccurate, or inconsistent data. Detecting and fixing these anomalies earlier on has a dramatic impact on the accuracy and consistency of business insights.

   A system of rules is used to profile and filter data as it is collected and transformed in the data lake. When rules identify data that are outside threshold limits and cannot be automatically fixed, these specific data issues can be triaged and escalated for follow-up by data engineers and data analysts. This type of rule-based data validation and data scoring focuses the limited time team members have by highlighting the priority areas where data may have the greatest business impact. Thus, data quality scorecards and dashboards will help drive visibility and understanding into where the manual effort should be focused.

6. **Let artificial intelligence and machine learning do the work of data discovery, data security, and data stewardship**

   With data volumes exploding, one of the largest challenges for a CDO is simply getting visibility into what data assets are even available. While the act of building a data lake helps centralize key data assets into a singular environment, there is still a question of discovering what assets to collect into the data lake in the first place.

AI can be used to automatically discover the structure in unstructured data. This understanding can then be used to automatically onboard other, similar, unstructured data. The result is a great boost in productivity for a task that is very time-consuming.

Similar to the way web search engines crawl and index the web, automated data scanners are used to proactively search and index new data assets throughout the enterprise. Machine-learning techniques then identify correlations and similarities between different data assets. This helps build a holistic view of data assets for data security and data stewardship. Rather than relying on manual approaches to detecting undesirable proliferation or failure to comply with data regulations, a machine-learning based approach proactively monitors and detects all the data across the enterprise to ensure maximum protection and compliance.

Moreover, a holistic view of data assets is used to form an intelligent catalog of all data assets and infer relationships between them. Data consumers, like business analysts, then use the catalog to identify new assets that may be of interest to them—in fact some catalogs will go as far as to recommend data assets based on machine learning techniques.

7. Design for a single source of truth available to a federated organization

The legacy of departmental data marts still haunts many organizations. This legacy has arguably led to the creation of data swamps where departmental teams build siloed data lakes that are inconsistent and duplicative of other environments in the organization.

The principle of co-location is essential to maximize the benefits of a data lake. You should look to a limited number of large data lake environments that are comprehensively organized around critical business domains. This principle of co-location ensures that data lakes truly reflect single views of truth, increases the power of predictive analytics, and minimizes unnecessary duplication across the organization.

Furthermore, data lake management approaches that exploit data sharing, data tagging, and project workspaces facilitate essential collaboration among data scientists and analysts. Data consumers should view one another as cohorts on analytical journeys where the work by one analyst in the data lake is published and shared with other analysts to reuse.

8. Standardize the process and drive consistency in the architecture

Teams often report the dilemma of re-inventing the same data management problems over and over. The absence of standardization permanently damages data lake efforts as demands continue to increase, because environments are simply not built for scale. Standardization and consistency are critical for long-term scale.

There is also an issue of re-use. You want your data management professionals and business analysts to re-use existing approaches to data management rather than creating new ones. The problem has always been that it is easier to create new “code” than to find existing “code.” With integrated AI, the data integration platform can proactively recommend existing “code” (i.e., rules, logic, policies) to re-use.

"By the time we go live with a new technology, it's out of date! The Informatica platform isolates us from the underlying sea of change that's going on." —Chief Data & Analytics Officer, Ford.

A standardized process and consistent architecture ensures that your data scientists and business analysts focus on innovation and analytics, and not on data management. The more that IT and LOB stakeholders focus on data management, the less they focus on driving the data-driven innovations that are so valuable to your business.

9. Establish policies, taxonomies and classifications, so all teams are aligned

One of the largest bottlenecks to speed, agility, and collaboration is the absence of common policies and a language or glossary of terms to provide business meaning and context. If everyone in the organization does not adhere to standard policies and recognize data assets consistently, siloed misunderstandings of data are created. In return, problems for enterprise-wide use will occur. Moreover, data consumers like data scientists often report spending too much time on cleaning up inconsistencies in data—instead of being focused on the value-added efforts of analysis.

Standardized processes, taxonomies, and glossaries are a way to ensure everyone on the project team is speaking the same language. Creating simple procedures earlier on in the process to establish what the key data assets are—and how they will be managed and referenced—eliminate churn and frustration. Standardized taxonomies and policies radically simplify auditing and lineage tracking for compliance reasons, so that you can always understand the provenance of data and proactively protect sensitive data.

Conclusion

Data lakes offer a unique opportunity to deliver radically new business insights very quickly and efficiently. The best practices outlined in this white paper will help you to avoid many of the common pitfalls that inhibit success and get your data lake environment started the right way.

Informatica Intelligent Data Lake Management is the industry's most complete and integrated end-to-end solution for data-driven digital transformation. Informatica enables companies to unleash the power of big data to become more agile and realize new growth opportunities through innovation that lead to intelligent market disruptions.

Informatica Intelligent Data Lake Management enables you to find, enrich, prepare, catalog, master, govern, and protect the big data you need to deliver accurate and consistent insights--empowering quicker business decisions. Based on Informatica's unique metadata-driven artificial intelligence technology, known as the CLAIRE™ engine, organizations systematically find any data, discover data relationships that matter, quickly prepare and share the right data with the right people at the right time. And, ultimately deliver more innovative, timely, relevant, and personalized data.

The best practices shared in this white paper should help you unleash the value of your next data lake initiative.

Next Steps

Go to www.informatica.com/bigdataready for resources, including eBooks, analyst reports, and webinars that provide best practices for managing your data lake.

Informatica

Worldwide Headquarters 2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871

www.informatica.com linkedin.com/company/informatica twitter.com/Informatica

IN09_0218_3401