Informatica

Turning a Data Lake into a Data Marketplace

Content

Tip: click to jump straight to any section.

Introduction

Introduction

There's no doubt that data lakes represent a huge opportunity for you to deliver new insights from vast amounts of data delivered from old and new sources.

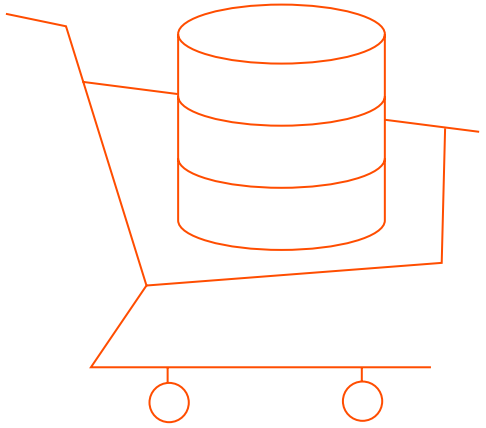
However, organizations struggle with the construction, maintenance, and effective use of data lake environments. As a result, they are failing to capitalize on data-driven insights, potentially missing out on capitalizing on new opportunities. Meanwhile, data lakes risk becoming passive spaces for storing data rather than active spaces for retailing data to engaged data consumers.

A new set of technological capabilities and organizational practices are emerging to form the basis for turning data lakes into data marketplaces. Central to this are the principles of designing for agility, building a data supply chain machine, and organizing for faster and more collaborative success.

What's a data marketplace?

A data marketplace is a new type of information management architecture that extends traditional notions of data lakes to combine a standardized and industrialized process for curating raw data assets into trusted information, with a collaborative and self-service mode of engagement with end users so that data consumers can quickly and easily shop for the data they need.

The aim of this workbook is to share the advice and best practices needed to maximize the value of your organization's data lake environments and capitalize on the potential for data-driven intelligent disruption through a data marketplace.



Part One

Design for Agility

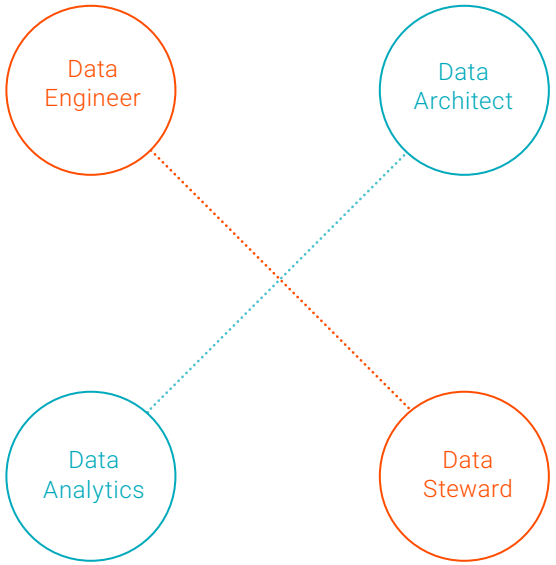
Obstacles to effective data marketplace implementation

Data marketplace environments promise to facilitate rapid discovery of new insights. However, organizations that embark on these initiatives often find themselves unable to extract maximum value from them, for a number of reasons:

- **Antiquated data management processes often inhibit speed, flexibility, and collaboration.** Complex requirement gathering processes and long development cycles cause delays for lines of business to get the insights that are needed to prove value and build the necessary momentum.
- **Excessive IT controls**–too many can slow down projects since IT is often unnecessarily involved in operations.
- **A lack of effective tools for collaboration.** Without these, teams are unable to reap the benefits from the work that other teams have already created.

Once these obstacles are cleared, the importance of cross-functional teams made up of data engineers and data architects from IT, as well as line of business stakeholders from analytics and stewardship teams working toward a common business program is paramount. Team members are empowered to represent the needs of their function while the group collectively executes the scope of a project from start to finish.

The main benefit to building cross-functional teams 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, as well as ensuring that everyone is aligned to a common understanding of the data available.



Questions to ask yourself:

Are your data management processes as efficient as they should be? If not, how can they be optimized to ensure they don't inhibit time to value?

How accessible and timely is data? Are the controls you are using too stringent?

Have you set up a cross-functional team with relevant stakeholders from across your business, who can provide multiple perspectives and ensure the right outcomes are delivered from your data marketplace project?

Have you defined roles and responsibilities, and provided supporting tools to ensure effective collaboration?

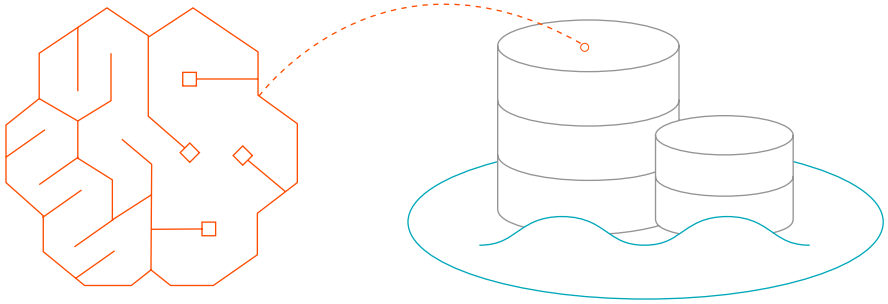
Empower your data scientists to access the data they need to assist in data preparation

Self-service data visualization tools, like Tableau, Qlik, and Zoomdata have become very popular over the years to give business analysts direct access to data. Self-service initiatives are one of the core tenets of building a data marketplace that brings data out of the shadows of a warehouse into the consumer-facing shelves of an organization. Enabling lines of business users to directly "shop" for fit-for-purpose data in the marketplace empowers them to engage in the process of preparing data in trusted assets.

But often the problem with these self-service initiatives is that business users either wait for the data they need from IT or are forced to adopt manual processes for curating and cleaning data to get it into the form they require–often taking place in spreadsheets.

This is where self-service data preparation comes in. It enables knowledgeable business analytical users 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 the data together. Furthermore, artificial intelligence and machine-learning-enabled techniques within tools can provide an automated and guided experience for business analysts as they explore and discover data in the data lake.



Use crowdsourcing and tagging to govern data assets

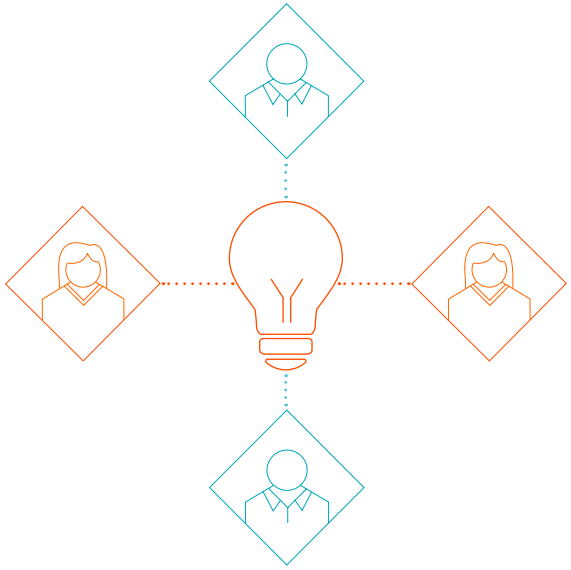
There is often a belief that data lakes can be left ungoverned. This is a dangerous myth: with organizations adopting data lakes for the processing of sensitive data–concerning patients or consumers, for example–effective methods of data governance are paramount. However, slow and centralized forms of governance can negate the agility benefits a data lake promises.

Online retail marketplaces have leveraged the so-called "wisdom of crowds" to enable consumers to share feedback and reviews to empower future consumers to benefit from previous experience. As such, this type of collaborative filtering and crowdsourcing of wisdom is another critical tenet of a data marketplace.

Data governance is intended to be a value-added function that increases the quality of data and ensures compliance with standards and protection of sensitive data. As such, business analysts and other consumers of the data have as much of an interest in the governance of the data as their stewardship peers. This is where the concept of crowdsourcing data governance comes in.

Crowdsourcing is the ability to tap into the business analyst user citizenry for knowledge and expertise that collectively enhances the quality of data. In a self-service environment, every user has the power to apply their subject matter expertise to improve the quality and context of data.

Business analysts should contribute their knowledge, through tags and other classifications, so that data assets are continuously increasing in quality. Collaboration then becomes a mechanism for ensuring self-resiliency, as business analysts help one another improve the quality of data assets.



Questions to ask yourself:

Have you made it as easy as possible for users across the business to prepare data from your data lake without involving IT, or does data preparation still require cumbersome manual curation and cleansing?

Have you implemented sufficient data governance policies to ensure you can keep sensitive data protected, and ensure the quality of data is fit for use in key decisions?

Are manual governance enforcement processes compromising the agility of your data lake?

Are business analysts able to effectively contribute their knowledge of data context to your data lake–through tags or other classifications?

Part Two

Build a Data Supply Chain Engine

The problems with manual and specialized processes

Speedy discovery of new business insights is a key benefit of data lake environments that are a foundation for data marketplaces. In a competitive environment where lines of business need speed, any processes that are not automated in a systematic fashion will delay the production of new insights. Without a high-speed production process, data assets can never be delivered to lines of business on time. As such, speedy processes are another critical tenet of data marketplaces.

More, antiquated hand-coding methods can often inhibit long-term maintainability of business logic. Hand-coded solutions built in very low-level languages pose a risk: if the language is no longer supported or the developers with the requisite knowledge leave the company, those solutions must be rewritten.

Even code generation solutions, that automate hand-coding, leave huge risk around supportability and maintainability. Once the artifacts of business logic are stored in specialized development paradigms, your business is forever dependent on those paradigms.

Beyond the lack of maintainability, hand-coded solutions pose a risk to auditability and governance. Most organizations are under internal and external mandates to track the modifications and usage of data. Without a logical view, operations executed by hand-coded solutions are difficult to track and monitor for auditing reasons.

Finally, there is a matter of practicality in relying on manual and specialized processes to handle the volume of data now required. With organizations facing intense growth in the amount of data they need to process, it is infeasible to expect a similar growth of resources to handle this data in response: organizations need to find an automated solution that can scale to support this ongoing data explosion.

Questions to ask yourself:

How much of your business logic is created and managed with hand-coded solutions requiring specialized coding skills? If your specialized developers leave the company, or the language falls out of use, how much risk and cost will it be to your initiative?

If you use code generation solutions, are they maintainable long term? Do they provide metadata transparency allowing for automation and business agility with changing needs?

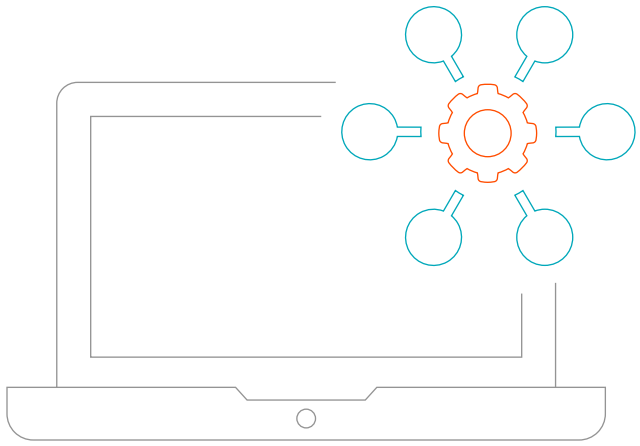
Are you equipped to scale a 30-50 percentage increase in data volume over the next five years?

Automate the ingestion and transformation of data

The most tactical aspect of any data lake environment is the automation of ingestion and transformation of data. Manual ingestion and transformation of data is a complex multi-step process that leads to unrepeatable and inconsistent results.

Successful organizations take advantage of pre-built connectors and high-speed ingestion platforms to load and transform datasets into the data lake. This enables data lakes to scale to increasing volumes of incoming data.

Automation also enables the fast iteration, flexibility and agility required to support changing business needs, because changes can be made to automated processes very quickly, without the risk of breaking down processes and impacting existing users.



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

As executives know well, problems that are not caught early cause larger issues later on. With data lakes, data quality errors that are not identified early can dramatically affect business insights due to inaccuracies or inconsistencies between different data assets. With the volume of data businesses must now manage and analyze, it is nearly impossible to spot data quality issues manually.

Artificial intelligence techniques that recommend and infer business rules are the answer: a method of automating data quality processes. Data lakes with rule-based data validation can automatically detect signals of incomplete and inconsistent data. By detecting these anomalies early, you can have a dramatic impact on the trustworthiness of business insights.

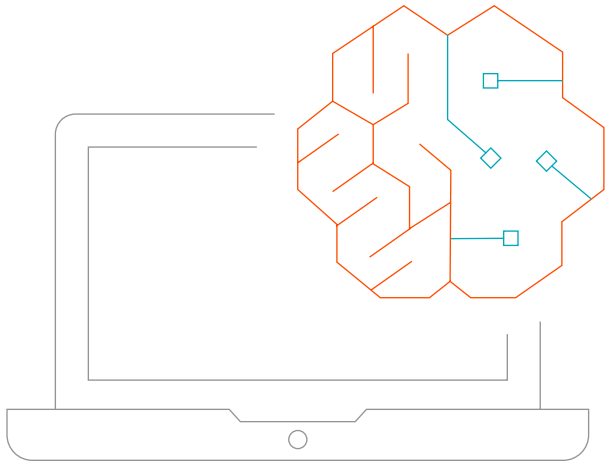
A system of rules must be used to profile and filter data as it is ingested and transformed in the data lake. When automated rules identify data that is outside threshold limits, these instances can be triaged and escalated for follow-up by data analysts and stewards. This type of rule-based data validation and data scoring focuses the limited time of team members, by highlighting the areas where data may have the greatest issues. Data quality scorecards and dashboards thus help drive visibility into, and understanding of, where manual effort should be focused.

Exploit machine learning for data discovery and data stewardship

With data volumes growing so rapidly, one of the largest challenges for organizations is getting visibility over available data assets. While the act of building a data lake itself helps to centralize key data assets into a singular environment, there is still a question of deciding which assets to ingest into the data lake in the first place.

Much as web search engines crawl and index the web, automated data scanners should be used to proactively search and index new data assets throughout the enterprise. Machine-learning techniques should be used to identify correlations and similarities between different data assets and build a holistic view of data assets for data stewardship.

Moreover, this holistic view of data assets should be used to form an intelligent catalog of all data assets, and the inferred relationships between them. Data consumers, like business analysts, can then use the catalog to identify new assets that may be of interest to them.



Questions to ask yourself:

Have you automated the ingestion of data into your data lake?

Have you implemented business rules and a stewardship process to identify and mitigate data quality issues?

Are you taking full advantage of machine learning for data discovery and stewardship?

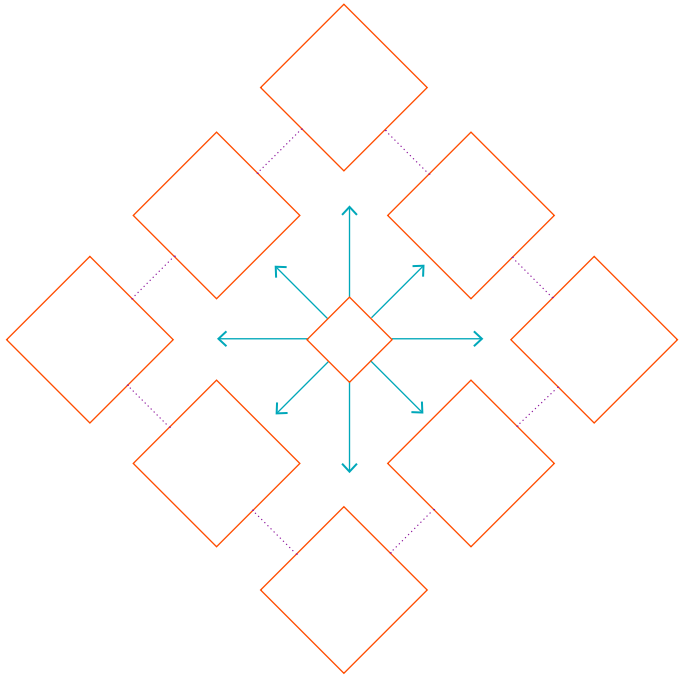
Part Three

Organize for Fast and Collaborative Success

The problems with siloed, decentralized teams

Organizations often face the challenge of working with IT and line of business (LOB) stakeholders across geographic and organizational boundaries. These organizational siloes can inhibit the benefits of data lake environments: one of the tactical goals of a data lake is to build a consistent single view of truth around data assets for multiple consumers to leverage. By making storage efficient, there is no longer a need for departmentalized data marts. A single inventory of data assets is another critical tenet of a data marketplace.

But the legacy of these departmental siloes, combined with a general tendency towards functional data hoarding, can limit the benefits of data lakes. Data lake management solutions can help facilitate collaboration and turn the wisdom of crowds into an asset, and not a liability.

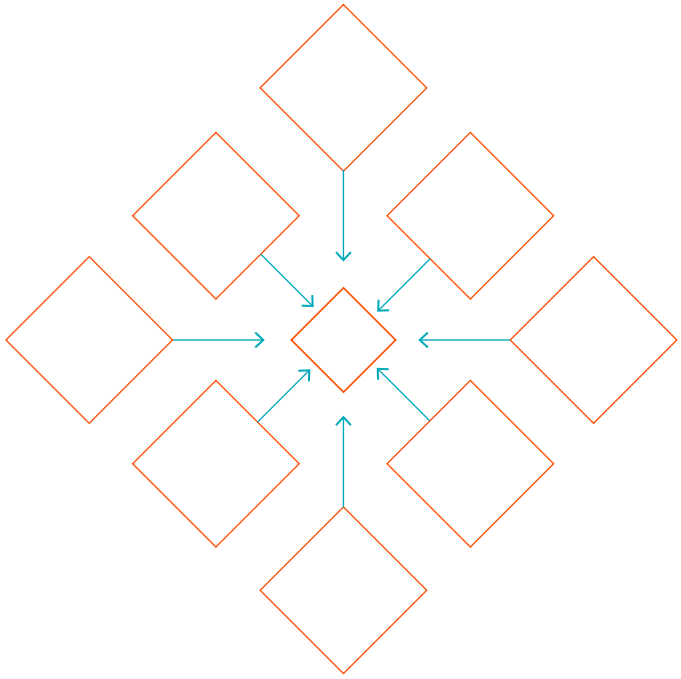


Design for centralization and collaboration

The legacy of antiquated models of data management still haunts many organizations. These antiquated models that were slow and manual forced organizations to hoard data in lines of business. Subsequently, departmental teams today, have started to 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 ensures that data lakes reflect single views of truth across the organization and minimize unnecessary duplication, which only increases governance risk and complexity.

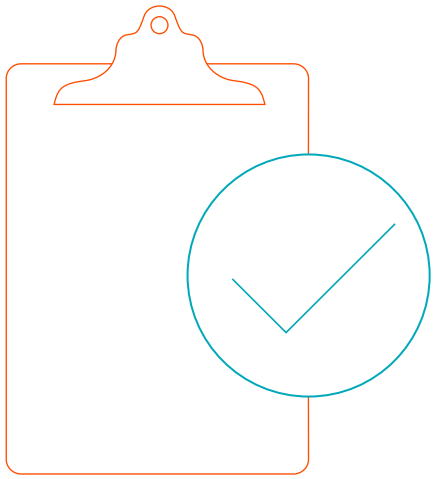
Furthermore, data lake management approaches that exploit data sharing, data tagging, and project workspaces can facilitate the needed collaboration for these environments. Data consumers should view one another as cohorts on analytical journeys where the work of one analyst in the data lake can be published and shared with other analysts to build upon.



Standardize the data management process and drive consistency in the architecture

Organizations often suffer from the curse of running into the same data management problems over and over. The absence of standardization can permanently damage data lake efforts as demands continue to increase, because environments are simply not built for scale: standardization and consistency are essential.

A standardized process and consistent architecture also ensures that your organization's resources are focused on innovation and analytics, and not on data management: the more that IT and LOB stakeholders are focused on data management, the less they are focused on driving the innovations that deliver the most valuable insights for your business.

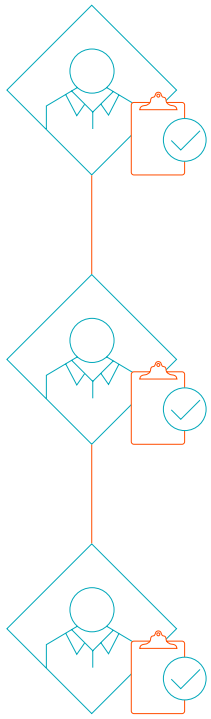


Establish taxonomies and classifications so all teams are aligned

One of the largest bottlenecks to speed, agility, and collaboration is the absence of a common language. If people across the organization do not recognize data assets consistently, it can create a siloed understanding of data that does not maximize enterprise-wide use. 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.

Curation of raw data into consistently parsed and prepared data dramatically reduces the overhead of data preparation by data consumers, such as data scientists. Standardized taxonomies and glossaries as part of a comprehensive metadata management program also ensure that everyone on the project team is speaking the same language. Simple exercises, facilitated by data catalogs, to establish what key data assets are and how they will be referenced can eliminate a lot of churn and frustration later on.

Standardized taxonomies can also radically simplify auditing and lineage tracking for compliance when data lake projects are handling sensitive data.



Questions to ask yourself:

Do your data lake users spend more time on data management (accessing, cleaning and transforming data to ensure data is fit for use), or on deriving value from insights (helping achieve business outcomes)?

Have you defined your metadata strategy for the data lake, ensuring you set up a standardized taxonomy and glossary of terms for data assets, and to ensure the necessary auditability and transparency is in place?

Are geographic and organizational boundaries inhibiting the structure and function of your data lake?

Conclusion

Data lakes offer a unique opportunity to deliver radically new business insights, quickly and efficiently. Exploiting best practices in solutions and processes will accelerate the delivery of these new insights. Moreover, by fostering collaboration in the process, data lakes are liberated from the basement of passive storage spaces into active data marketplaces for data consumers. The best practices outlined in this book can help avoid many common pitfalls and enable you to get your data lake environment started the right way.

Further Reading

Read the “Intelligent Data Lake Management” Exec Brief

Learn how Informatica can help address your data lake challenges and enable you to get more accurate and consistent insights.

READ MORE

About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and 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. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed 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.

Worldwide Headquarters
2100 Seaport Blvd, Redwood City, CA 94063, USA
Phone: 650.385.5000
Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871

informatica.com
linkedin.com/company/informatica
twitter.com/Informatica
facebook.com/informaticaLLC

CONTACT US

IN18-0718-3492

Informatica™