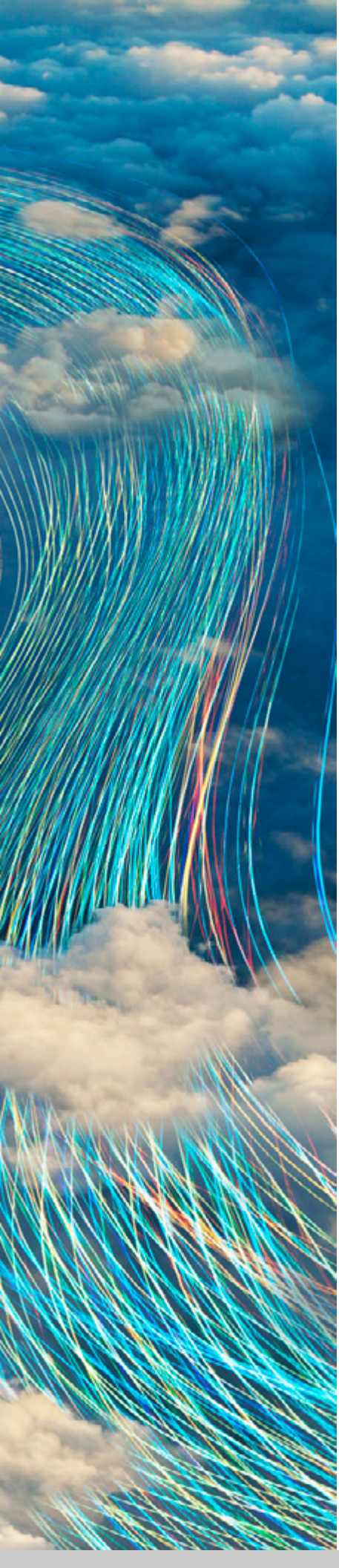
The Transformative Impact of the Data Marketplace

WHITE PAPER

early adopter research



CONTENTS

early adopter research

INTRODUCTION

To get value from all the data you have and all the data that is coming in, both now and in the future, a new concept is needed that incorporates what has been learned from data warehouses and data lakes and also transcends them by adding crucial new capabilities. We call this concept a data marketplace.

In some companies, the data lake has gone dry. In others, it has spilled over or become stagnant and polluted. It is the rare company that can claim victory in managing, analyzing, and operationalizing an exploding number of datasets from existing as well as new sources.

To get value from all the data you have and all the data that is coming in, both now and in the future, a new concept is needed that incorporates what has been learned from data warehouses and data lakes and also transcends them by adding crucial new capabilities. We call this concept a data marketplace. This paper explains:

1. The ways that data warehouses and data lakes are failing to solve the modern data infrastructure challenge.
2. What a data marketplace is and how it differs from previous approaches.
3. How data marketplaces ensure that data finds its way to those who need it and becomes an asset with increasing value.
4. What life is like when a data marketplace becomes operational.

Drowning in Data, Not in Value

Today, companies are confronted with managing more data than ever before. The advent of big data has led to an increase in the volume, velocity, and variety of data, which brings with it both opportunities and challenges. The number of new apps on-premise, in the cloud, and provided by SaaS vendors keeps growing, and so does the data from those apps. Add to this data from server logs, machine data, IoT devices, and other digital realms that provide new types of signals—if you can harvest them. New BI, AI, and ML technologies only generate business insights if they are fed with the right data. At the same time, they generate volumes of data themselves.

Businesses benefit from all this data only if they have data systems, architectures, and tools in place to track it, transform it, let people know it exists, and then deliver it when and where needed. But far too often, they do not. Data doesn't arrive in time to drive timely and effective decision making. More data doesn't in and of itself create more benefit. In fact, unidentifiable, unusable data is more of a source of cost than a source of value.

> "Organizations are realizing that simply putting lots of diverse data into Hadoop or a data lake won't magically create meaningful insights without further integration, transformation, enrichment, security, and governance. Delivering connected data across on-premises and cloud sources is not trivial, especially when it involves large data volumes, complex data models, and high speed of ingestion. Poorly integrated business data often leads to poor business decisions, reduces customer satisfaction and competitive advantage, and slows product innovation — ultimately limiting revenue, while adding little or no business value."
>
> NOEL YUHANNA, FORRESTER RESEARCH

The tools to date were either not designed for this era of big data or have failed to live up to their promise of being able to handle and operationalize it.

Data warehouses and data lakes each offer unique benefits that should be incorporated into the future of data storage and management, but on their own, they are not, and cannot be, the answer for 21st century enterprise data needs.

Data warehouses created a unified canonical model of a business that became the foundation for reporting, BI, and data discovery. Their enterprise-ready data management capabilities and strong attention to governance remain relevant, but they generally are not capable of delivering new data at speed or providing detailed views into big data.

Data lakes started out with the idea of driving business value of big data, the name given to the vast amount of data from new sources that could not be managed or analyzed using a data warehouse. The idea was then expanded by some into a vision for a super data warehouse that could become one repository to rule them all. Data lakes were implemented to allow us to store massive amounts of data with the goal of analyzing it in new ways to harvest valuable business signals. The problem is that companies that have implemented data lakes alongside their data warehouses generally lack clear business context of data in the lake, making it unusable at best and risky at worst.

WHAT WE LEARNED FROM DATA LAKES

Despite their failures, the experience gained in using data lakes has offered a variety of lessons.

The data lake was a step forward because it allowed:

- **Volume.** Storage of massive amounts of data at an affordable price.
- **Variety.** Storage of many different types of data.
- **Schema on read.** New paradigms for making use of data that allowed collection of data before it was fully analyzed. Data could be stored and provisioned before its actual business value was determined. This made it practical to have all enterprise data staged and ready for unknown business questions.
- **Support for many new types of ingestion and consumption patterns.** A single repository for ingestion and consumption of streaming data as well as structured, document, and columnar based access.

Many data lakes end up becoming places for hoarding data rather than offering a way to organize, gain insights, or derive business value from data.

Yet despite these benefits, the data lake vision has yet to be fully realized. Data lakes offer immense storage and the ability for experts to create new pipelines for certain types of analytics and data transformations. But companies spend resources deploying developer-focused open source products to get even limited value from the data lake.

Can One Repository Rule them All?

Data lakes promised that companies would finally have a big data repository to store everything, and from which analytics could be run. They would be "one repository to store it all."

But the data lake is failing to live up to this promise for a variety of reasons and companies have become increasingly disillusioned with them.[1]

Many data lakes end up becoming places for hoarding data rather than offering a way to organize, gain insights, or derive business value from data. They don't meet three essential requirements:

1. Become the single repository for all data
2. Provide actionable data on demand
3. Ensure enterprise-ready data management

Failure to meet these requirements is at the root of enterprise frustration with data lakes.

FROM DATA LAKE TO DATA MARKETPLACE

The ability to manage flows of messy big data, govern it, and enable broad business use demands that companies become adept at creating and managing flows of all their data. This must include the following:

Constantly incorporating new data. As new sources arrive, they must be evaluated and incorporated into the working set of data used by a company. Data can't just be tossed in; there must be an automated and scalable onboarding process.

1. Early Adopter Research has outlined a litany of reasons data lakes have failed in its "Saving the Data Lake" research mission.

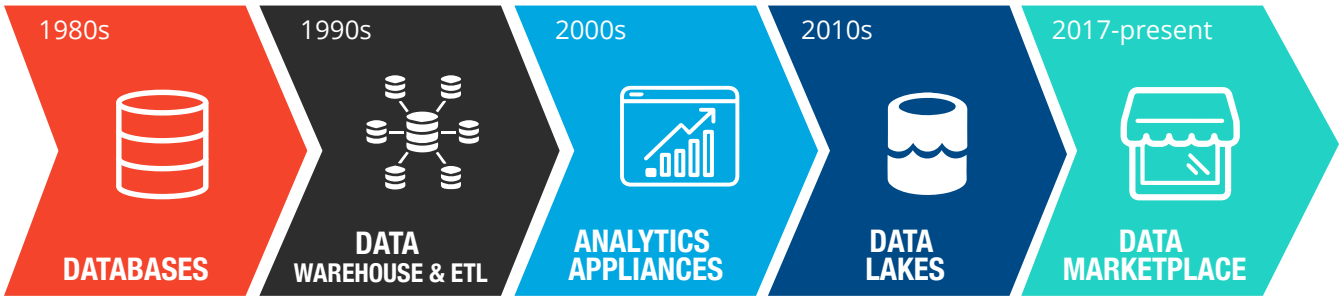
Creating a virtuous cycle of data usage. Once new data is collected, users need to be able to immediately access it and enhance it with business metadata. Automated analysis and profiling enriches the data with even more metadata. Users can encourage other users to access the data and put it to work. Over time, more data arrives, more metadata is added, and the number of people using it increases. This cycle becomes self-perpetuating to the profit of the business.

Launching data products. As the use cases for integrated data become understood, the culmination of a data supply chain is a data product that is used in applications, in dashboards, or in other ways to promote awareness and effective action. The incremental addition of datasets over time means that the business is developing a more robust and thorough understanding of all its data.

Increasing the number of data practitioners. Many people must be able to participate in the process of ingesting, evaluating, and productizing data because so much data is arriving. Increasing the number of users also benefits the business because they each bring their own context that allows more data to have value. This, when shared, helps to create a broader understanding of enterprise data assets, thereby allowing more informed analytics.

Data marketplaces that provide these capabilities are thus the next evolution in data management and represent a scalable, complete solution moving forward. They are explicitly designed to address the unresolved problems of the data lake.

The data marketplace refines and extends the original intention of the data lake, enabling companies to know where all their data is and track its usage. By knowing what data is being used, companies can make more informed decisions about what data to expend resources on improving and how to organize data across the entire business. A data marketplace allows companies to answer more and better questions about their data and to compete on analytics.



The evolution of data repositories

> “451 Research believes that the need for data to be filtered, processed, treated and managed to make it suitable for multiple analytics use cases is critical to delivering value from the data lake. Data governance and self-service data preparation are key elements of functional data lakes and associated data marketplaces, with machine learning-driven insights and recommendations an increasingly important aspect of accelerating the generation of value from enterprise data.”
>
> MATT ASLETT
> 451 RESEARCH

Data Democratization: Power to the People

Data marketplaces enable data democratization, a core requirement of the future of data. The democratization of data means that users decide what data is important and what data they want to work with, without having to rely on IT intermediation. The data marketplace is thus aptly named: it’s a dynamic marketplace where users decide what has the most value. Companies can gauge which data is most popular based on usage and decide where to invest. Users can shop for data products within the marketplace and then join these products to other ones they’ve created to launch truly powerful data-driven projects.

CAPABILITIES OF A DATA MARKETPLACE

Data marketplaces bring together all the positive capabilities of data lakes and data warehouses into a single product that allows companies to meet all their data needs. Data marketplaces thus build on, expand, and improve the capabilities of data lakes and data warehouses by offering:

All data in one simplified view

Actionable data, on-demand

Enterprise-ready data management

All data in one simplified view that contains:

- Any and all data sources, that have gone through a validation and profiling process
- An assessment of the quality of each dataset, including every column
- A smart, integrated data catalog of technical and business metadata that organizes, documents and describes all data in the data collection
- Automated data management to simplify and scale delivery of analytics-ready data

Actionable data, on-demand. Instead of a static, isolated list, users search and shop for data with a navigable interface that caters to all technical skill levels. New data is rapidly incorporated so users always have the freshest data. Specific capabilities should include:

- Designation of "ready" data where quality is assured and sensitive fields are secured
- A guided shopping experience that finds and understands the best data for the user
- The ability to reuse and collaborate so analysts and users can easily share data across the organization
- Lineage tracking captures the origin, evolution, and meaning of data

Enterprise-ready data management. Unlike data lakes, data marketplaces are fully mature at the architectural level, meaning that security, governance, metadata, and data protection are all incorporated. Data marketplaces are highly scalable and agile, offering:

- End-to-end protection and control. Automated integration with enterprise security and governance frameworks
- Built-in scalability. Natively runs on big data platforms to deliver high performance and linear scalability across multiple configurations including on-premise, cloud, hybrid, and multi-cloud
- An open architecture. Delivers ROI quickly and integrates seamlessly with existing systems
- Connectivity with common data sources and analytics applications as well as support for legacy and complex data formats

A data marketplace not only unites the best capabilities of the data warehouse and data lake into one platform, but also enables additional features. Foremost is the increased agility it gives companies over their data, as data becomes easier to find and easier to understand because it's ready to use. The data isn't siloed, nor are users dependent on IT to gain access to it.

	DATA WAREHOUSES	DATA LAKES	DATA MARKETPLACES
ALL DATA IN ONE SIMPLIFIED VIEW Integrated view of all data. Documented, accurate, complete	F • Conformed to answer specific business questions • Limited to BI and reporting data • Most enterprise data not accessible • No big data accessible	C • Can't easily include all enterprise data • New data not automatically verified, quality checked, or documented during ingest • As a result data in the lake is opaque, dirty, and hard to navigate	A • Includes all enterprise data from raw to ready • Smart data catalog including all technical, operational, and business metadata
ACTIONABLE DATA, ON-DEMAND Self-service. Fulfill requests for new data in hours	C • Only data in warehouse accessible • New data takes months • Reuse of limited data in single platform • Widespread reuse and collaboration impossible	C • Most people can't access data • Requires specialized programming and technical skills. • Technical users spend >70% finding and prepping datasets • Widespread reuse and collaboration impossible	A • Amazon-like shopping experience • Users find, understand, prepare datasets themselves • New data added in hours • Empowers widespread reuse and collaboration
ENTERPRISE-READY DATA MANAGEMENT Security, governance, metadata, and sensitive data protection. Support current and new IT platforms	A • Comprehensive, mature enterprise-ready data management capabilities	F • Lacks enterprise-ready data management capabilities • Attempt to fill gap by integrating open source products; additional work required	A • Data marketplaces provide comprehensive, mature enterprise-ready data management capabilities

Registering External Data

The data marketplace has the plumbing to allow all data, including data outside the data lake, to be evaluated, catalogued, searched, and accessed. This is called registering the data. This means including all data in the catalog regardless of where it sits in the enterprise, handling the arrival of new datasets in a seamless way and then being able to reuse those datasets in a cost-effective manner.

With data marketplaces, companies get a unified view of data in the data lake along with registered data in other systems - regardless of whether the data is inside or outside the enterprise, on-premises or in the cloud. This allows companies to take an agile approach to data management. Instead of loading all data into the data lake before using it, analysts and other stakeholders can search the catalog and work with data from any source. Business demand drives the maturation and growth of the marketplace.

This concept naturally extends beyond a single data lake and enterprise systems to multiple data lakes and cloud data sources. With a data marketplace, companies can use this expansive catalog to see what data they have — and what data they're missing. They can analyze the metadata before calling IT to provision data across systems. Ultimately, companies can monitor high-value activities associated with their data over time and use this information to optimize how they manage, store, improve, and operationalize their data.

Automatic Extraction of Metadata and Profiling Information

Data marketplaces allow companies to profile and automate extraction of metadata. With automatically collected marketplace metadata, you can build a repository that allows you to productize your data in new ways.

Automated analytics can enhance the data catalog by finding patterns in the data. For example, the statistical signature of a column or table can be used to recognize and tag sensitive data such as Personally Identifiable Information (PII). Generating metadata by analyzing metadata enables the system to become smarter over time. Automation is a scalable, enterprise-ready approach to managing growing data volumes and garnering maximum business value from data. These rules will vary from industry to industry based on industry needs.

Usage Data Creates a Feedback Loop

What data is business ready? One absolutely clear signal that data is ready is that it's being actively used. A data marketplace also collects usage data. You're thus able to manage your data in tiers of value, depending on how much it is being used.

You're thus able to better determine your data priorities, deliver meaningful business results as early as possible, and build data products quickly.

All of this metadata drives greater agility in the development process. With a marketplace-based catalog of data, you can identify opportunities and challenges and prioritize high value use cases to gain immediate value while at the same time making the data lake more and more usable.

The Data Marketplace and the Big Data Ecosystem

It is important to recognize that the data marketplace doesn't claim to solve every problem completely. Technologies have already emerged that provide support for the data lake, ranging from data fabrics to data catalogs to data prep and application development tools, but all of these good ideas do not add up to a product. The data marketplace is the product that unifies these data sources and tools together to deliver to companies the integrated solution they're seeking. But a data marketplace also provides the flexibility that, when needed or desired, it is possible to use any number of technologies

LIFE IN A DATA MARKETPLACE

For a data marketplace to work in practice, companies must establish an end-to-end process that is systematically applied to all their data. All data must be cataloged into the data lake so that profiling and other forms of analysis can create a full set of metadata that describes the data. Having all the data cataloged also allows analysis and exploitation of the data at whatever level of detail is required.

The full complement of metadata allows the creation of a comprehensive metadata catalog that supports expansive access and self-service. Users of all technical levels can see how others are using the same data and then put it to use as they see fit, without needing to rely on IT.

With a data marketplace, you can create more effective ways to package and deliver data for better utilization. Crucially, by empowering self-service, the data marketplace puts managed data into the hands of business users, which is a key way to drive change and accelerate innovation by:

1. Lowering data management costs
2. Accelerating time to answers
3. Massively increasing the productivity of analytics teams and data consumers

Allowing more people to use data directly leads to gathering more data about usage, which in turn makes the marketplace more powerful.

The data marketplace incorporates feedback from the "customers" (the data consumers) to discover the most valuable datasets. With all their data linked through the catalog, companies can learn based on market feedback which data is most valuable and how it is being used, cleaned, and improved.

By tracking data as it moves from raw to ready in its lifecycle and using crowd-sourced feedback from data consumers, companies can begin to identify and categorize the value of their datasets. Even categorizing data in this way ensures that all data has some utility, whatever its ranking, because there is now organization where before there was just a data swamp.

From there, companies can invest in curating the most valuable data. The data that turns out to be the most useful can then receive the attention it deserves and be cleaned, modeled, integrated, and turned into an easy-to-use product.

Allowing more people to use data directly leads to gathering more data about usage, which in turn makes the marketplace more powerful.

In the end, the data marketplace allows delivery of data products as well as the raw forms of data that lead to their creation. Automated analytics can enhance the data catalog by exploring patterns. This can streamline organization of data, especially sensitive data, and the entire system becomes smarter over time.

Finally, the data marketplace does not limit the use or expansion of existing tools. Rather it unifies governance, big data, and security tools into a common, seamless data supply chain.

Once companies start using data marketplaces, they soon see the increased efficiency of the marketplace versus working with a data warehouse along with data lakes that are more like data swamps. Users can go in and put data to use, finding the data they need in minutes themselves versus putting in requests for IT that might take days, weeks, or months. Users can share data across the organization and thus transform data into higher and higher forms of value. The entire organization becomes more efficient and agile with its data, which is key as new data sources continue to pour in.

Here's how different parts of the organization benefit from the data marketplace.

Data marketplaces encourage users to share data across the organization and thus transform data into higher and higher forms of value.

Data Caretakers

The roles of IT staff, data stewards, and data engineers now have new dimensions of visibility and new processes for activities that have always been ad hoc. Caretakers now have the ability to:

- Make better decisions about how to allocate resources and the responsibility to act when the data marketplace reveals an opportunity
- Focus on making popular datasets more useful, cleaner, and distilled
- Run an agile process of creating and improving standardized datasets (data products)
- Better allocate and focus training and consulting resources based on the marketplace signaling which groups are making use of data
- Understand which third-party tools are most often used by different segments and have the most impact
- Easily say yes to adding new data to the marketplace as there is now a defined process for onboarding datasets with a full set of descriptive information that can be searched. The catalog can flag when a new data source already exists in the marketplace to avoid data sprawl. New data creates new value, not a swamp
- Register external data into the marketplace to vastly expand the power of the catalog to encourage and track usage of data not stored in the marketplace

Data Users

A data marketplace also changes the user experience significantly. Users gain the ability to:

- Search a comprehensive catalog that includes a rich set of metadata about each dataset, including profiling information
- Find external data that is registered in the marketplace
- Assemble their own datasets
- Build on and collaborate with datasets created by other users
- See which data is most popular based on marketplace signals
- Use a growing set of data products curated by the caretakers.

CONCLUSION

In an important sense, the data marketplace is simply a name for a new data management paradigm required to keep up with the growing amount of data that has become available. Companies can choose to ignore new data or manage it with older, bottlenecked processes. But that is equivalent to giving up.

Those who choose to find a way to be in control of this issue must face certain realities. Success depends on finding a way to:

- Allow new data to be captured, profiled, described, and be found by users
- Shape data into the most usable forms
- Enable all the most popular data to be found by or recommended to those who might find it valuable
- Get more people involved with this whole process

The data marketplace vision makes all of this happen. It is an excellent start on solving a problem of frightening complexity.

The data marketplace transforms the problem of managing the onslaught of new data to an opportunity to deliver more value to the business. If you don't feel that way when new data arrives, perhaps it is time to consider building a data marketplace.

This paper was written by Early Adopter Research and sponsored by Qlik

Connect with us

early adopter research