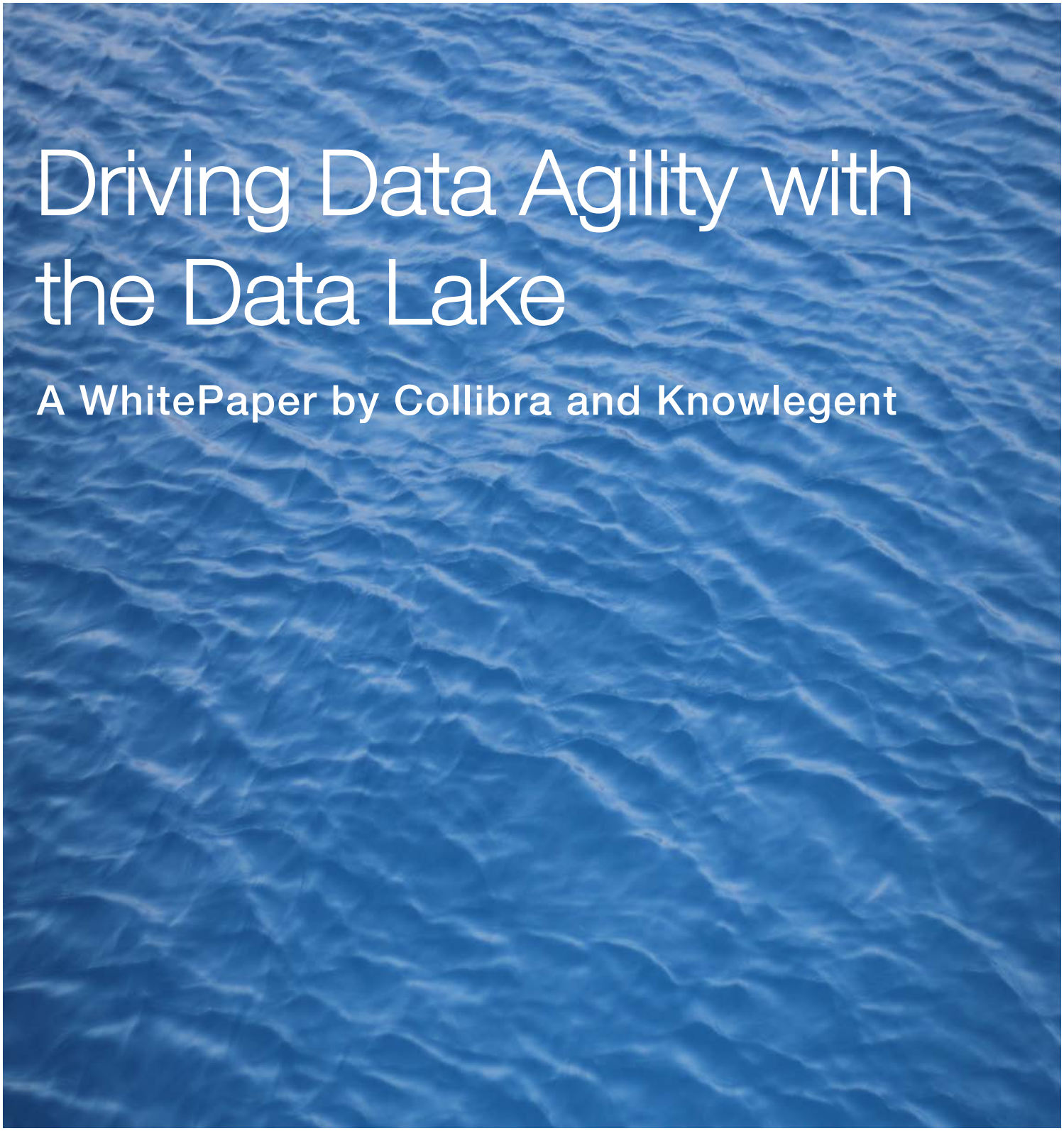
Collibra
DATA DRIVEN . BUSINESS . DRIVEN DATA

Knowledgent



Driving Data Agility with the Data Lake

A WhitePaper by Collibra and Knowlegent

Collibra Inc
25 Broadway, 9th Floor
New York, NY 10004 – USA
(t) +1 646 963 6513

Contact@collibra.com

Knowlegent
1177 Ave. of the Americas
10th Floor
New York, NY 10036
T 212-265-3300
info@knowlegent.com

Driving Data Agility with the Data Lake

Business users are continuously envisioning new and innovative ways to use data for operational reporting and advanced analytics. The data lake, a next-generation data storage and management solution, was developed to meet the ever-evolving needs of increasingly savvy users.

This white paper explores existing challenges with the enterprise data warehouse and other existing data management and analytic solutions. It describes the necessary features of the data lake architecture and the capabilities required to leverage a Data and Analytics as a Service (DAaaS) model. It also covers the characteristics of a successful Data lake implementation and critical considerations for designing a data lake.

Current Enterprise Data Warehouse Challenges

The evolution of business needs coupled with advances in data storage technology has made the inadequacies of current enterprise data warehousing (EDW) solutions become more apparent. The following challenges with today's data warehouses can impede usage and prevent users from maximizing their analytic capabilities:

Scale

The rigid and predefined nature of data flows within an EDW environment make it challenging to scale that environment. This scaling happens in a manner that is consistent with the 3Vs (volume, velocity, variety) of big data. While some EDW implementations are capable of handling the volume, the velocity is often a technical challenge that has high cost solutions (custom hardware, specialized databases, etc.). Variety is also difficult as each new bit of content requires new flows and end-to-end processes to be described up front. In addition, each new analysis or use of the data often requires a new data mart. In many cases this scaling is taking place outside the EDW environment, in spreadsheets or in memory exploration tools, because the EDW infrastructure is unable to handle it.

Timeliness

Introducing new content to the enterprise data warehouse can be a time-consuming and cumbersome process. When users need immediate access to data, even short processing delays can be frustrating and cause users to bypass the proper processes in favor of getting the data quickly themselves. Users also may waste valuable time and resources to pull the data from operational systems, store and manage it themselves, and then analyze it. Even data professionals struggle to find and manipulate the appropriate data.

Flexibility

Users not only lack on-demand access to any data they may need at any time, but also the ability to use the tools of their choice to analyze the data and derive critical insights. Additionally, current data warehousing solutions often store one type of data, while today's users need to be able to analyze and aggregate data across many different formats.

This flexibility must be one of the primary guiding principles of a new solution. There is so much variation in the organization in both the production, acquisition and consumption of data, that predefining the data constructs is nearly impossible, and certainly impractical. Instead, organizations must be able to identify the constructs, relationships and assets needed to accomplish a particular task. This implies that a lot of the metadata, policies, and rules about the data will be developed and applied on the fly, when needed.

This flexibility implies that the glossary of data must be highly flexible and easily modified, and that the

processes for defining new data and new relationships be automated and streamlined. Lineage of a particular business term in a report may include a whole host of complex assets including analytical models derived from other data, map/reduce jobs, visualization models, as well as the usual assortment of tools and antecedents. Having a flexible way to track the complete lineage is crucial for the development of advanced analytics. Onboarding data and creating new policies and relationships must be automated and structures or the perception of trustworthiness will suffer as the data becomes more complex and varied.

Quality

Often, the current data warehouse is viewed with suspicion. It is unclear where the data originated and how it has been acted upon. Users may not trust the data, and worry that data is missing or inaccurate,. A common response to this is to circumvent the warehouse in favor of getting the data themselves directly from other internal or external sources. This invariably leads to multiple, conflicting instances of the same data. Without a clear and easy understanding of the quality of data, users will lose confidence and create these conflicts.

The same can happen in a data lake environment. The primary solution for this is to create explicit visibility into the quality of data, and allow the users to decide whether it is suitable for the intended purpose.

Search-ability

With many current EDW solutions, users cannot rapidly and easily search for and find the data they need when they need it. Inability to find data also limits the users' ability to leverage and build on existing data analyses. Advanced analytics users require a data storage solution based on an IT "push" model (not driven by specific analytics projects). Unlike existing solutions, which are specific to one or a small family of use cases, what is needed is a storage solution that enables multiple, varied use cases across the enterprise.

What is needed is a simplified landscape where the data is consolidated and the formats are known.

This consolidation can introduce more complex relationships in the data assets, since the commonality of infrastructure and access methods provides the ability to reuse analytical models on different sets of data. While this was theoretically possible with traditional EDW, the likelihood that the data was accessible in similar ways made this a practical impossibility.

This in turn implies a need to create comprehensive lineage and governance that include assets that historically have been outside the data management infrastructure control, like models, visualizations, and data preparation, data processing and map/reduce jobs.

This new solution needs to support multiple reporting tools in a self-serve capacity to allow rapid ingestion of new datasets without extensive modeling, and to scale large datasets while delivering performance. It should support advanced analytics, like machine learning and text analytics, and allow users to cleanse and process the data iteratively and to track lineage of data for compliance. Users should be able to easily search and explore structured, unstructured, internal, and external data from multiple sources in one secure place. The quality and metadata associated with this information must be collected in a way that allows for continuous enhancement and improvement, as more uses for the information in the data lake are found. Policy definition and application is critical for insuring the quality and correct use of information.

The solution that fits all of these criteria is the data lake.

The Data Lake Blueprint

The data lake is a data-centered architecture featuring a repository capable of storing vast quantities of data in various formats. Data from webserver logs, data bases, social media, and third-party data is ingested into the data lake. Curation takes place through capturing metadata and lineage and making it available in the data catalog (Datapedia). Security policies, including entitlements, also are applied.

Data can flow into the data lake by either batch processing or real-time processing of streaming

data. Additionally, data itself is no longer restrained by initial schema decisions, and can be exploited more freely by the enterprise. Rising above this repository is a set of capabilities that allow IT to provide Data and Analytics as a Service (DAaaS), in a supply-demand model. IT takes the role of the data provider (supplier), while business users (data scientists, business analysts) are consumers.

The DAaaS model enables users to self-serve their data and analytic needs. Users browse the lake's data catalog (a Datapedia) to find and select the available data and fill a metaphorical "shopping cart" (effectively an analytics sandbox) with data to work with. Once access is provisioned, users can use the analytics tools of their choice to develop models and gain insights. Subsequently, users can publish analytical models or push refined or transformed data back into the data lake to share with the larger community.

Since a data lake is a consolidation of data, it must support many different types of uses. Most organizations begin to develop their data lake as a sandbox for predictive analytics. However, it should also support the ingestion of raw data, access to streaming data both internal and external, data curation, and ETL-like transformation qualities. Most data lakes have a two-way relationship with existing EDW solutions. Finally, it must be able to handle different domains of information with different oversight, policies and structures.

It is impossible to create a data lake that satisfies these requirements without having a robust architect, sound operational principles, and comprehensive governance capabilities.

Challenges Arising from the Data lake Solution

Even a flexible, next-generation solution like the data lake is subject to its own set of challenges. Although a large volume of data is available to users in the Data lake, problems can arise when this data is not carefully managed, including:

Lack of Data Governance

Without the structure and controls to manage and maintain the quality, consistency, and compliance of data, a data lake can rapidly devolve into a data swamp. One of the lake's principle advantages, the common store and access of data, is also one of its weaknesses. Historically, organizations have used distribution of data, and technical barriers as a substitute for policy controls and enforcement. Bringing this data together in the lake makes the need for policies and data sharing agreements acute. It also exposes other types of suboptimal behavior, such as information hiding and excessive manual manipulation. Furthermore, every time some new data is brought into the environment, it has the potential to combine with the data already present. This in turn creates the need for new policies and new quality checks. Without a robust governance environment that covers the information in the lake as well as its sources and use cases, it is impossible to leverage the data. Information is hidden, or incomplete, or of unsuitable quality because of missing governance.

Poor Accessibility

Although the data might be available, its value is limited if users are unable to find or understand the data. Again, this is the outcome of having highly automated data governance. Big data is a gigantic expansion of the 3 Vs of data. For this to function properly, there must be a robust infrastructure that is easily scaled to deal with the volume, a solid platform for managing the performance of the data and the velocity of change, and comprehensive governance for tracking the attributes that explode with data variety : quality, lineage, sharing, policies and responsibilities for the data. Without these three things, the data will not be accessible, because you will either not be able to find what you need, access it fast enough, or be able to find understand and exploit it.

To maximize the value of the data lake and avoid these challenges, organizations need to ensure that their data lake implementations address specific critical success factors.

Characteristics of a Successful Data lake Implementation

A data lake enables users to analyze the full variety and volume of stored data. This necessitates features and functionalities to secure and curate the data, and then to run analytics, visualization, and reporting on it. The characteristics of a successful data lake include:

- Robust infrastructure that supports the scale of data to be used by the organization
- Operational excellence that can handle the increasing rapidity of data processing and use
- Governance excellence that enables maximum flexibility without sacrificing quality, control, security or privacy.

These can be further broken down into some specific requirements that are top of mind for any data lake implementation:

Heterogeneous Tooling

Extracting maximum value out of the data lake requires customized management and integration that are currently unavailable from any single open-source platform or commercial product vendor. The cross-engine integration necessary for a successful Data lake requires multiple technology stacks that natively support structured, semi-structured, and unstructured data types.

Domain Specificity

The data lake must be tailored to the specific industry. A Data lake customized for biomedical research would be significantly different from one tailored to financial services. The data lake requires a business-aware data-locating capability that enables business users to find, explore, understand, and trust the data. This search capability needs to provide an intuitive means for navigation, including key word, faceted, and graphical search. Under the covers, such a capability requires sophisticated business glossaries, within which business terminology can be mapped to the physical data. The tools used should enable independence from IT so that business users can obtain the data they need when they need it and can analyze it as necessary, without IT intervention.

Automated Metadata Management

The data lake concept relies on capturing a robust set of attributes for every piece of content within the lake. Attributes like data lineage, data quality, and usage history are vital to usability. One of the main characteristics of the data lake is that data which historically could not be combined due to technological barriers can now be brought together. This creates two interesting effects. One is that you need more metadata about objects, because some of that metadata was implied by the implementation (i.e. if it is in the support document store, then it must be either an answer to a customer question, a knowledge base article, or a release document) The second issue is that you need more information to create some of the novel combinations that yield the best insight.

Given the explosion in variety of information, the process of maintaining this metadata requires a highly-automated metadata extraction, capture, and tracking facility. Without a high-degree of automated and mandatory metadata management, a data lake will rapidly become a data swamp. This metadata is not limited to what can be found in the lake since the critical aspects of the metadata may depend on attributes of the sources of the data (usually in some traditional data management infrastructure, but could be from external sources), and also its usage (i.e. whether data is suitable can potentially be derived from the other uses of that data).

Configurable Ingestion Workflows

In a thriving data lake, new sources of external information will be continually discovered by business users. These new sources need to be rapidly on-boarded to avoid frustration and to realize immediate opportunities. A configuration-

driven, ingestion workflow mechanism can provide a high level of reuse, enabling easy, secure, and trackable content ingestion from new sources.

This is also critical for realizing the economic benefits of the data lake. Part of the implied value on the lake is that you eliminate many of the multi-source problems. However, finding the correct data becomes a needle in a haystack problem as more data is ingested. The business lineage of the data, not just its technical lineage, and the relationship of the data at rest in the lake to its sources and its uses is required to achieve understanding of the data. This must be coupled with usable views into the current quality of that data, and any other policies, agreements, or restrictions that govern its suitability for the intended purpose.

If this kind of information is not up to date and easily visible to data stewards, data scientists, and data users alike, the data lake will suffer and not deliver the full value. Users will not trust data if they do not have confidence that it comes form the right places and is interpreted in the right way.

Integration with the Existing Environment

The data lake needs to meld into and support the existing enterprise data management paradigms, tools, and methods. It needs a supervisor that integrates and manages, when required, existing data management tools, such as data profiling, data mastering, and cleansing, and data masking technologies. These tools insure that data which is inbound to the lake is as well known and described, and of the highest quality possible. The lake must also integrate with the uses of the data so that visualizations, wrangling, manual manipulations, Map/Reduce jobs, and analytical models are all linked to the data. Only in this way can the organization create predictions and insight of high quality with complete confidence.

Keeping all of these elements in mind is critical for the design of a successful Data lake.

Conclusion

Creating a data lake can bring tremendous value and flexibility to your organization, and provide your business teams with the information and insight needed to achieve their goals. Building a data lake requires good planning and the establishment of a solid architectural, organizational, and governance foundation. This model must be highly automated to handle the volume, variety and velocity of information inherent in the data lake. Organizations that take these steps maximize the value of their data lakes, and become truly data driven.

About Collibra

Collibra Corporation is the industry's only global data governance provider founded to address data management from the business stakeholder perspective. Delivered through a cloud-based or on-premise solution, Collibra is the trusted data authority that provides data stewardship, data governance for the enterprise business user. Collibra automates data governance and management processes by providing business-focused applications where collaboration and ease-of-use come first. Collibra's data governance platform embraces the new requirements of big data solutions, with automation, machine learning, and the flexibility to govern data assets from source to visualization. Find out more at http://www.collibra.com/

About Knowledgent

Collibra

Knowledgent

Knowledgent is an industry information consultancy that helps organizations transform their information into business results through data and analytics innovation. Our expertise seamlessly integrates industry experience, data analyst and scientist capabilities, and data architecture and engineering skills to uncover actionable insights. We have not only the technical knowledge to deliver game-changing solutions at all phases of development, but also the business acumen to evolve data initiatives from ideation to operationalization, ensuring that organizations realize the full value of their information. Find out more at http://www.knowledgent.com