INFORMATION ASSET

WHITEPAPER

Data Governance in the lake

Abstract

Data Lakes were originally constructed to address advanced analytics by providing vast arrays of structured and unstructured data for data scientists and their peers to mine. More recently, however, they are being used to address many other use cases such as traditional Business Intelligence and ad hoc reporting. While they were originally thought of as storage platforms for 'Big Data' i.e. unstructured and semi-structured datasets, they are now being used to also store more traditional structured data. As companies move this data into the Data Lake, Data Governance becomes imperative to address needs such as understanding the meaning and lineage of the data, ensuring that the data is suitable for business purpose and obeys regulations such as GDPR, HIPAA and CCPA, and ensuring that restrictions to access and contractual terms are met. Data Governance in the Lake (DGITL) lays out some aspects of this Governance which are relevant whether companies have successful Enterprise Data Governance programs or not.

Data Lakes have been with us for some time now. They were originally named by James Dixon, founder and former CTO of Pentaho, who used the analogy of comparing a cleansed, packaged bottle of water to represent a data mart, as opposed to a large body of water in a more natural state with more water streaming in from various sources to describe the Data Lake. Looking at data repositories in these terms, the Data Lake would contain data in all forms: structured, semi-structured, unstructured and raw, cleansed and standardized and enriched and summarized.

Many of the early Data Lakes were deployed in-house on Hadoop infrastructure, and were largely used by data scientists and advanced analytics practitioners who would take advantage of the vast pool of information to be able to analyze and mine data in ways that had been virtually impossible before. Due to the unpredictable nature of the data, and the fact that it was virtually always accessed as part of a large dataset rather than an individual record, the thought was that the Lake did not lend itself to Data Governance in the same way that a Data Warehouse might. Most Data Lakes went largely ungoverned in the early stages.

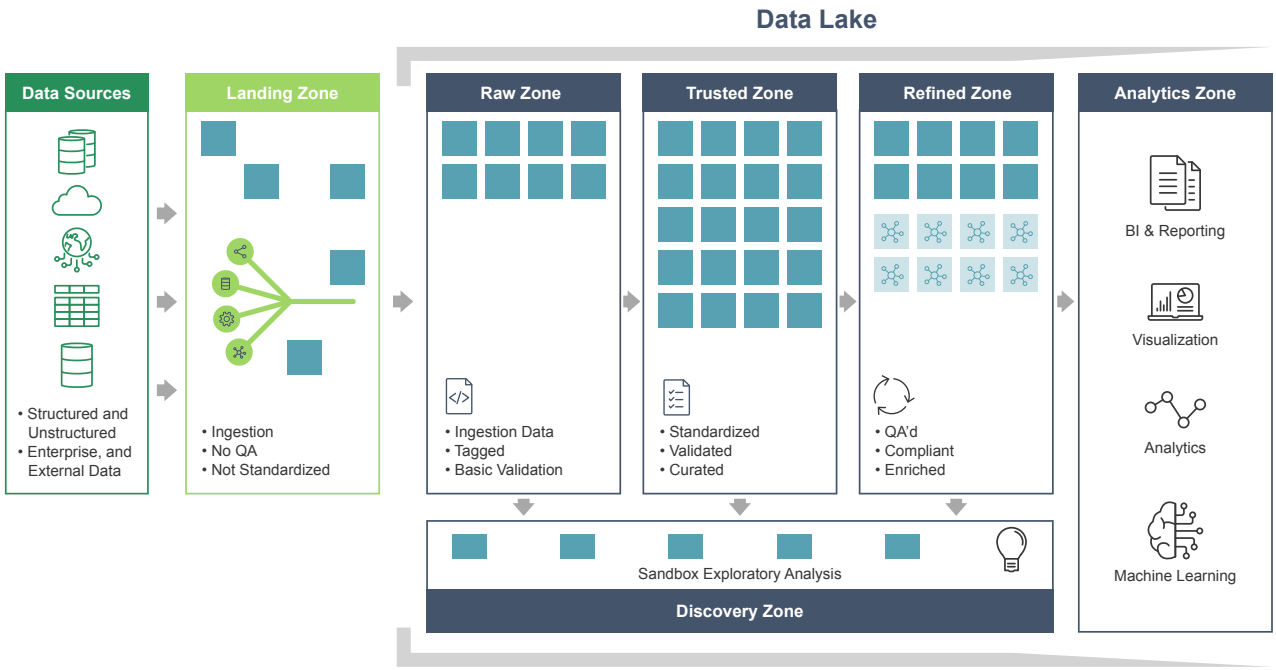
More recently, however, Data Lakes have begun to migrate to the Cloud to take advantage of the scalability of processing and storage, and the reduction in infrastructure costs. As they become more cloud-based, companies are beginning to take advantage of the added flexibility and reduced costs to move some of their more traditional structured data to the Data Lake. The intention is to make the data more visible and accessible to people who have a need to use it, but have had difficulty in accessing it due to the siloed nature of existing data repositories, and lack of any index of what is available. Some companies are beginning to move their entire Business Intelligence capability over to the Lake.

The addition of the other BI and data accessibility use cases to the Data Scientist one has had two distinct effects on many Data Lakes, one physical and one not:

- The Data Lake has physically changed by deploying multiple 'zones' of data that add the capability to cleanse, summarize and enrich the data to make it available for multiple uses
- The democratization of data and the need for multiple use cases that require that the data be fit for purpose and protected from misuse have meant that Data Governance now must play a much greater role in Data Lakes

Data Lake Structure

While this paper is not intended to be a deep discussion of the architecture of Data Lakes, it is necessary to look at how they are being deployed in order to talk about the various aspects of Data Governance that may be applicable. This section will provide a high level reference architecture of a Data Lake to facilitate the Data Governance discussion.



The depiction above is by no means a universal model. But it does illustrate the types of zones and the types of data that are being deployed in today's Data Lakes.

- **The Sources** are coming from both internal and external locations and consist of structured, semi structured and unstructured data.
- While not universally deployed, **the Landing zone** provides a location for data to be staged prior to ingestion into the Lake. This data is not persisted, it is only in this zone while it is needed.
- **The Raw zone** is the first true Data Lake zone. It consists of all the data that is ingested into the Data Lake in its native format. It is stored as-is, and provides lineage and exploration capability. Data in this zone is still largely used by data scientists for advanced analytics and data mining
- **The Trusted zone** is where the data is cleansed and standardized. Data Quality is measured and monitored in this zone. The data is secured for authorized access and may be certified as being fit for purpose.
- **The Refined zone** sees data being enriched, transformed and staged for specific business uses such as reports, models, etc.
- T**he Discovery zone** receives data from all of the zone for sandboxes and exploratory analysis traditionally done by data scientists.
- Finally, **the Analytics zone** allows the data to be used. It consists of reports, dashboards, ad hoc access capabilities, etc. It may also be used to provide data to internal or external users in forms such as flat files, spreadsheets, etc.

When looking at this structure using the same analogy as James Dixon, the entire data supply chain in all its forms from the natural state through cleansing and filtering to storage both on a large and small scale would now be in the Lake.

Data Democratization

Locating data in the Lake, free of the artificial restrictions that are inherent in siloed data repositories, provides companies many potential benefits in enabling access to data that is useful to them in their daily jobs. But there are some steps that must be taken in order to be able to take full, allowable advantage of these benefits:

- The data must be properly tagged, described, cataloged and certified so that potential users are able to easily find and understand the data.
- Legitimate restrictions such as company classifications, Government regulations (e.g. HIPAA, GDPR, CCPA), ethical considerations and contractual terms must be followed. There is a risk that making the data available to those who need it may also make it available to those who should not have it.

Both of these considerations can be addressed by developing a Data Governance in the Lake (DGITL) program.

Data Governance

Many companies have tried to address Data Governance over the years, with varying degrees of success. Lack of understanding and commitment, unwillingness to deal with the rigors of Data Governance and unwillingness to invest the amount of money that it would require are amongst the most commonly cited reasons for lack of success. But with the rise of the Data Lake, there is another opportunity to govern at least some of the data that a company owns without going through the rigors of governing it all. And if for no other reason than the ability to adhere to regulations and contractual obligations, there is a business imperative to govern the data. So the question arises, what type of Data Governance program will maximize the usage and efficiency of the Lake without necessarily incurring the cost and effort of a full enterprise program. If companies are willing to adopt a program aimed specifically at the Lake, we believe the answer is yes. Note that this paper is making the assumption that a full program does not exist across the company, although Data Governance may have been adopted for some subset of the data (e.g. Master Data). But even in cases where a robust enterprise program exists, it is still worthwhile looking at the Lake and its data usage patterns and ensuring that the concepts in the remainder of this paper are covered.

When we look at the way that typical Data Lakes are being used today, we see a difference not only in the types of data that are in the Lake when compared to an enterprise Data Warehouse, but also of the typical usage patterns. Data Warehouses were intended to contain the authoritative source of data for the enterprise, while in many cases Data Lakes are being used to contain ‘slices’ of data intended for different uses. Of course, the Raw Zone discussed above is intended to hold ‘all’ the data, but as the data is moving into other zones, there is a desire to make it fit for different usages. This provides us an opportunity to govern data at point of use rather than for the enterprise. The main advantage of that is that when governing data at point of use, you only need the agreement of the provider of the data and the user of the data on what the data means and when it should be supplied, while if governing at the enterprise level, you need the agreement of everyone with a stake in the data. Other advantages (and one disadvantage) of governing at point of use are outlined on the next page.

Governing at Point of Use	Governing at the Enterprise Level
Governing data for a business purpose/usage	Governing data regardless of purpose
Manage cost through Program (Usage) budget	Unclear how data governance is funded
Easy to identify the data usage owner	Unclear who will own data at the enterprise level
Data and quality are defined and measured for a specific use	Data and quality requirements need to be agreed across the enterprise
Scope can be identified and managed	Scope is amorphous
Data may be defined and managed at the enterprise level for a specific purpose (e.g. Master Data)	Functional users are 'forced' into definitions and quality rules they do not want
Manageable ongoing data governance	Data Governance activities grow exponentially as more data comes under governance
Enterprise view will need to be treated as a Usage if needed	Enterprise view constructed and available for all

It is important to note the last row in the table. While governing data at the Usage level makes the size, scale and investment more manageable, it does not produce an enterprise view of data, nor does it assign Ownership to a single person. This makes it imperative that other aspects of Data Governance such as cataloging the data and its meaning and usage rules to be of increased importance so that the data is locatable and understandable.

Cataloging the data is emphasized by another aspect of the Data Lake. The imperative to ensure that making the data easy to find and use for those who need it must be balanced by the equally important imperative that it must not be available to those who should not have it, or used for purposes that it should not be used for. This has been relatively easy in the past since datasets were typically built for systems to access, and the data was not available to those who did not have access to the systems. Data Lakes will still have security tools that prevent people from accessing the Lake or specific zones of the lake if they do not have the credentials, but having credentials to access the lake does not mean that all data can be accessed. For example, external data may have rules that limit access e.g.

- Purchased prescription data may only be allowed to be used for medical research
- Data from suppliers such as Lexis-Nexus may only be allowed to be used for specific purposes

Managing these types of restrictions becomes much more difficult when the data is in a multi-purpose dataset like the Data Lake. Setting minimum standards for metadata cataloging becomes important to ensure that any company, regulatory or legal restrictions on the data are known at all points where the data is deployed.

But simply knowing what the restrictions are is not enough. There needs to be a facility to ensure that the restrictions are adhered to at all points along the data journey in the Lake. This can be managed by assigning accountability for data at every stop and for every usage, and mandating that as data gets passed from one usage to another, the accountability will also be passed. This requires 2 constructs:

- We need to assign accountability for the data at each place of rest. This may be accomplished by identifying the person responsible for the usage of the data and assigning them the accountability. We often refer to this role as the Usage Owner. This not the same role as the Data Owner in enterprise Governance programs. This person is only accountable for the data that they use. Most often, this person is responsible for a process or function that will use the data.
- A Data Sharing Agreement is then needed between the Usage Owner that is supplying the data and the Usage Owner that is requesting it. This agreement may exist at any point in the data lake, including as the data is being ingested. This Data Sharing agreement established three things:
  - The Source Usage Owner understands the requirements for use of the Requesting Usage Owner and agrees that the request is for an acceptable use
  - The Requesting Usage Owner understands any restrictions on the Usage of the data and agrees to be accountable for their implementation
  - The Requesting Usage Owner agrees to cascade the restrictions and requirements to any subsequent request for that data

This last point on the cascading requirements is important since it prevents all approval of any usage of the data from being referred to the original usage owner, and it provides the capability for Usage Owners at any point to enrich the data and still be able to supply it to others who might need it.

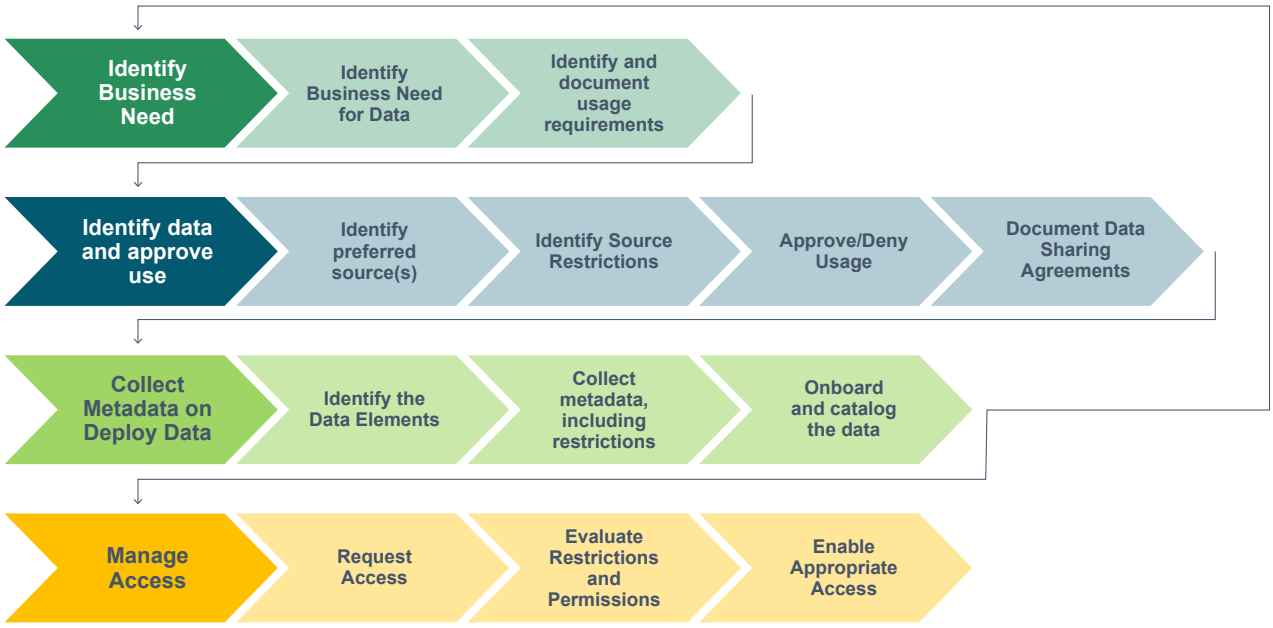
An added extension in the Data Sharing agreement is that it may bring into play other stakeholders around the restrictions such as the Privacy, Legal and Regulatory departments, and even Audit. Once agreed, this Data Sharing Agreement becomes part of the metadata that is stored with the data it covers.

When that is achieved, the only thing that remains is to manage the access to the data according to the restrictions. While managing access to data may not be thought of as strictly a Data Governance process, it does lean heavily on some of the controls put in place by the Data Governance program we have discussed. Ideally, there would be a technical solution to managing access but some types of restrictions are harder to manage than others, as shown in the table below.

Company Classification – Easy	
• Public Data can be see by everyone • Internal data can be seen by employees • Sensitive data can be seen by some employees • Highly sensitive data can be seen by few employees	Relatively easy to classify users and automate access requests
Government Regulation (e.g. HIPAA, GDPR) – Medium	
• Regulations specifically provide the rules for data access	Need to develop Framework
Legal/Contractual/Ethical – Hard	
There are many potential types of restrictions e.g. • Data cannot be used for commercial purposes • Data must not be used after a certain date • Data must be used in an ethical manner	Hard to anticipate all restrictions likely to be implemented at least partially manually

As can be seen, it is likely to be much easier to provide automated access management to data such as typical company classification levels than for other restrictions such as contractual terms, where there may be many different types. For this reason, the initial management of access may be manual, at least for some data. Management of use, in our scenario, will be the responsibility of the Usage Owner of the dataset that is being requested. Because of the Data Sharing agreement, this role has agreed to be accountable for the usage of the data at this point – another good reason not to have the data owned by a single person. In many cases, Access Management is being managed by a workflow engine, ensuring that all requests are being addressed.

In fact, this entire process can be enabled by developing a series of workflows to identify and ingest data, move it through the Lake and manage access against it. These flows can be administered manually or through a workflow engine. These high level workflows are depicted below.



This paper has focused on the accountability, ownership, documentation (via metadata) and ability to manage access against restricted data issues that the Data Lake is introducing. There are many other aspects of Data Governance that we have not focused on that remain important to the Data Lake. For example:

- Data Quality may be affected by data being ingested into the Lake. Apart from the fact that Data Quality will not be applied uniformly throughout the Lake (e.g. it will not be applied in the Raw zone), it may be different for different usages in the Lake. Some usages (regulatory reporting) may have much more strict requirements than others (analytics), and this may be reflected in the data that is being used for the usage. If there is an Enterprise usage defined, that is likely where enterprise data quality measuring and monitoring would take place.

- Likewise, there may be different requirements for metadata across zones and usages. Even if this is the case, however, it will be necessary to establish a minimum standard for metadata so that the ownership, meaning and restrictions can be applied.
- Lineage is another Data Governance aspect that may have different requirements for different usages. The collection of Lineage information should be enabled by the transportation into and through the Data Lake, but some usages may need it from its entry into the enterprise ecosystem, and that is not under the control of the Data Lake processes. Under these circumstances, extra effort will need to be expended to capture the full lineage.

These aspects, and others, should be taken into account as the Data Lake is being established and built out.

In summary, we have seen that the movement of data to the Lake for Business Intelligence use cases is helping companies address efficiency and cost challenges and is making data much easier to find and understand for business users. But with that move comes the need to ensure that the data is being used for legitimate business purposes, and is not being made available to unauthorized users. The Data Governance capability that we have discussed here can help ensure appropriate access. Due to its focus on Usage of data rather than all data, this program may be appropriate for companies that have not established enterprise data governance programs. For those who have been successful with enterprise data governance, this program may be used as a guide to enhance that program and try not to overload enterprise Data Owners with requests for access to the data for which they are responsible. Either way, the business reasons for moving data to the Lake will mean that Data Lakes will continue to be deployed, and the issues with managing data in a more democratized environment will have to be addressed.

About Information Asset

Information Asset is a consulting firm that delivers Data Governance, Privacy and Sensitive Data Management to diverse clients in multiple industries. We specialize in end-to-end business advisory services, implementation and technical solutions for the Data Governance Lifecycle including Consulting, Metadata Integration, Sensitive Data Management, Tool Evaluation, Product Implementation and Training.