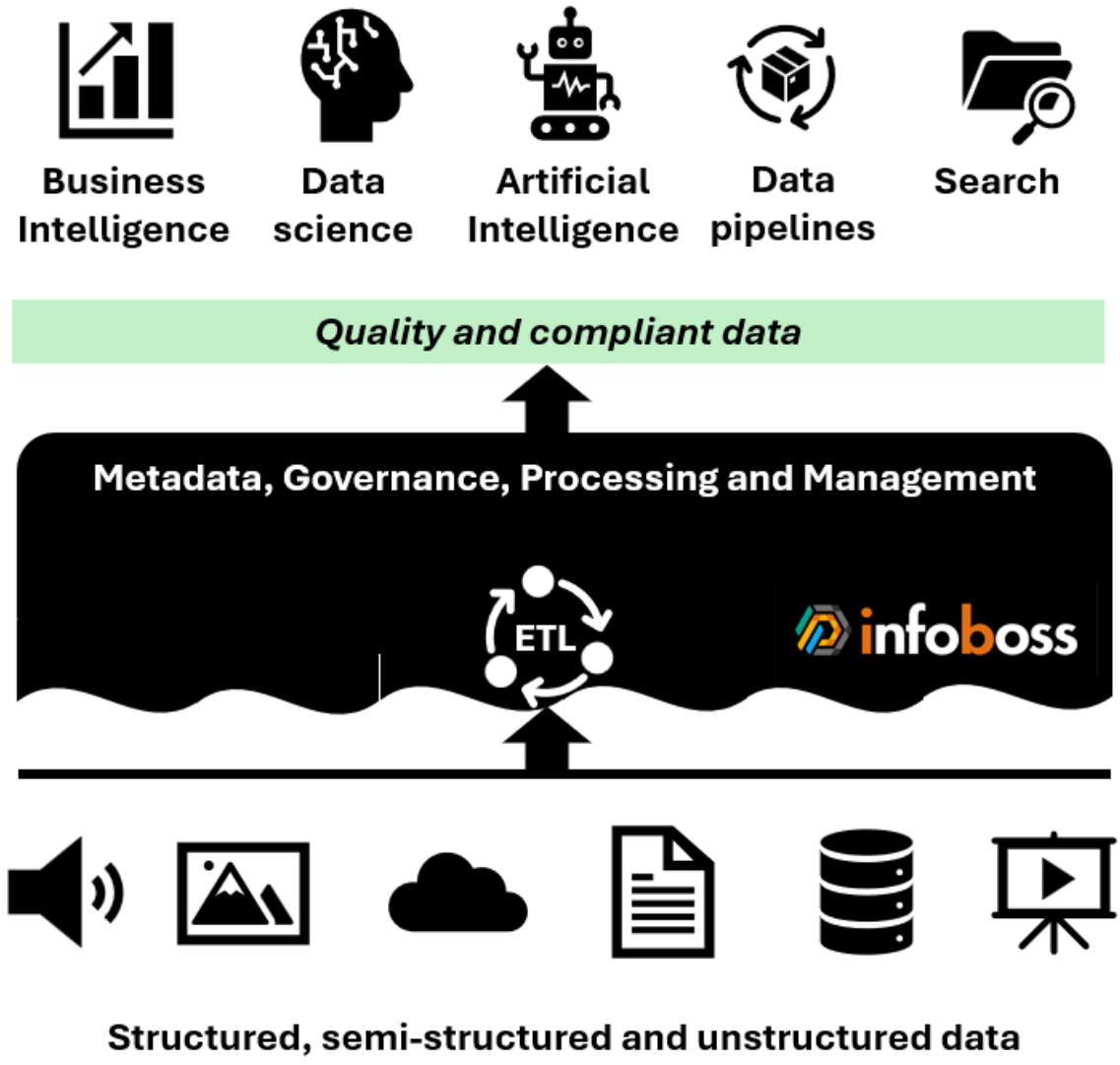
Data Lakehouse solution

An affordable, flexible, scalable and (most importantly) attainable Data Lakehouse solution

I_am_the@infoboss.co.uk



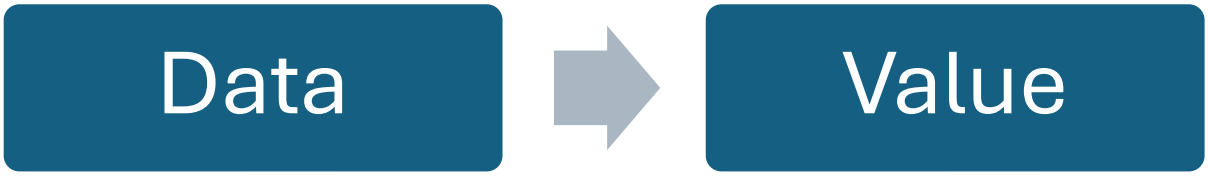
Contents

1 Introduction

Organisations are under immense and increasing pressure from both internal and external stakeholders to not only comply with regulation, but also take advantage of technology such as Business Intelligence, Predictive Analytics, Artificial Intelligence and more to improve outcomes for the business and its stakeholders. A widely accepted common denominator in achieving successful outcomes is *data*, not just any data, but trusted data, fit for purpose data, in short, data that is both high-quality and regulatory compliant.

Trusted data is fundamental within the modern organisation as an enabler for business transformation and supports its ability to be agile in adapting to change and being able to capitalise on opportunities afforded by new technology. Without it, the business will always be playing catchup and arguably lose its competitive advantage.

In short businesses want to maximise value from their data assets.



So how do they do this efficiently? Especially given the resources at their disposal? And how do they do it in a sustainable way?

1.1 Traditional approaches to attaining value from data…

Since the late 1980s the concept of the data warehouse was heralded as the way to maximise value from data. The concept is grounded in structured data, i.e. data held in databases, bringing it together into one place to provide views of the data (data marts) that the business can consume (via reporting tools) to inform decision making and support the delivery of business goals. However, as data estates began to grow and an increasing volume of unstructured data (emails, documents, text, video, images) began to enter the fray, data warehouses were found wanting as data growth and requirements for insights rapidly increased and the slow pace of data warehouse maintenance and enhancement led to data cottage industries evolving.

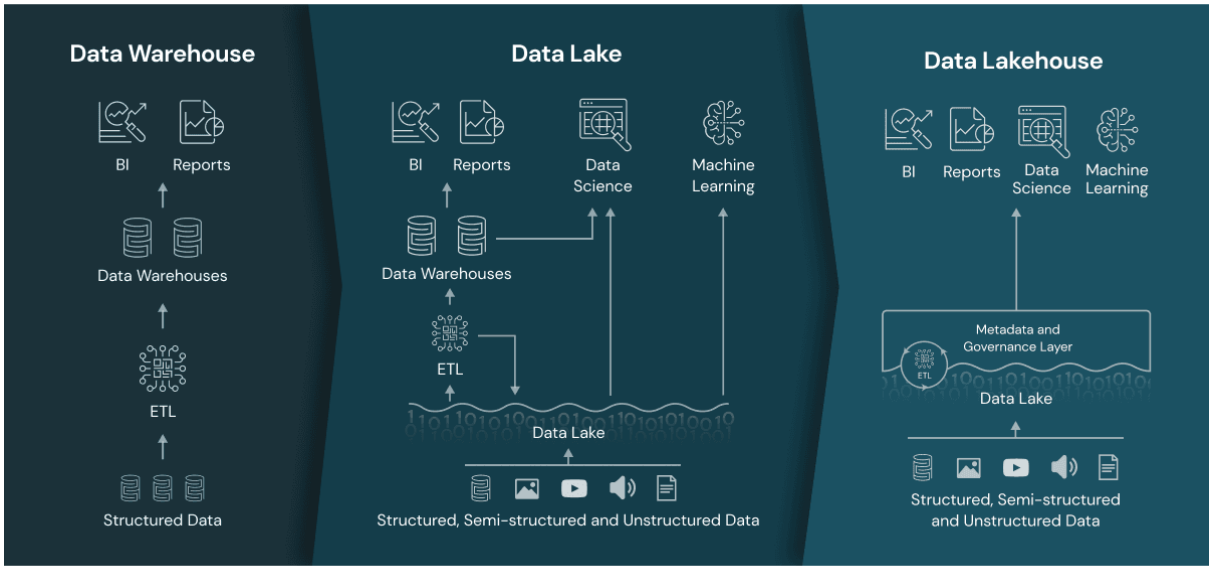
From 2011 a new concept emerged, the data lake. The data lake supported the storage (cheaply) of huge volumes of data that could potentially be analysed. However, the absence of data governance and control, the right resources and costly, poorly scoped projects resulted in significant investment in complex data architectures that are a challenge to build and maintain. Furthermore, the lack of tools to exploit the data stored in the data lakes has led to the emergence of the unflattering term “data swamp” to describe the outcome as data is stored and not utilised, effectively stagnating.

It is now widely recognised that these data architectural approaches cannot on their own, address the requirements of the modern enterprise. This has led to the emergence of a new architectural approach that combines the best of data warehouses and data lakes to create a unified, scalable, and high-performance data management and consumption architecture that empowers the organisation to curate and publish quality and compliant data in a managed and controlled way. This approach is known as the Data Lakehouse.

1.2 What is a Data Lakehouse?

A Data Lakehouse supports the ingestion and use of diverse datasets, i.e. both structured and unstructured, meeting the needs of both business intelligence, artificial intelligence and data science workstreams. It leverages similar data structures from data warehouses and pairs it with the low-cost storage and flexibility of data lakes, enabling organisations to store and access big data quickly and more efficiently whilst allowing them to mitigate and manage data quality and compliance risks.

The illustration below provides a simple visual representation of the three architectural approaches.



A key component of the Data Lakehouse (not present in the warehouse or lake architecture), is the metadata and governance layer (sometimes called a semantic layer). This is a foundational pillar of success in the modern data architecture.

In short to achieve value from data it is now widely accepted that the metadata and governance layer, afforded by the Data Lakehouse architecture is the way to go, empowering your data teams to manage the flow of quality and compliant data to [data] exploitation technologies like business reporting, artificial intelligence, predictive modelling and more.

2 Considerations before you start your Data Lakehouse project

There are many potential technologies in play to deliver a Data Lakehouse (Amazon, Microsoft, Databricks, IBM are just a few). You'll discover as you research the various vendor solutions that they are non-trivial technologies requiring significant levels of expertise and budget to deliver a successful outcome. To illustrate this complexity point, please refer to Appendix A to see the reference architectures for two such vendors - Databricks and Microsoft.

There are several key considerations to bear in mind before committing on a journey with one of the mainstream solution vendors. CarlosData [Ed. nom de plume] writing for the Medium, says "Organisations venturing down this path [Data Lakehouse] must thoughtfully assess the costs, benefits, and potential obstacles before embarking on this transformative expedition."

https://medium.com/@cdgonzal/budget-nightmare-the-surprising-price-tag-of-data-lakehouses-231a02890d71

Implementing a Data Lakehouse architecture within your organisation will require expert resources that fully understand the numerous technologies in play. These resources will be required to not only build but maintain the solution for years to come. Ask yourself, do you have all the required resources? If not, can I afford to recruit them? Will we be able to retain them? If not, then you're probably going to require a 3rd party service provider to deliver and maintain your lakehouse – do you have the resources to manage this relationship and deliver a successful outcome?

In addition to skills, there is a significant up-front investment in infrastructure. Furthermore, the varying storage and compute costs through use of cloud platforms mean that cost budgeting is extremely difficult to fully understand without clear evidence as to the amount of resources required to service your solution demands.

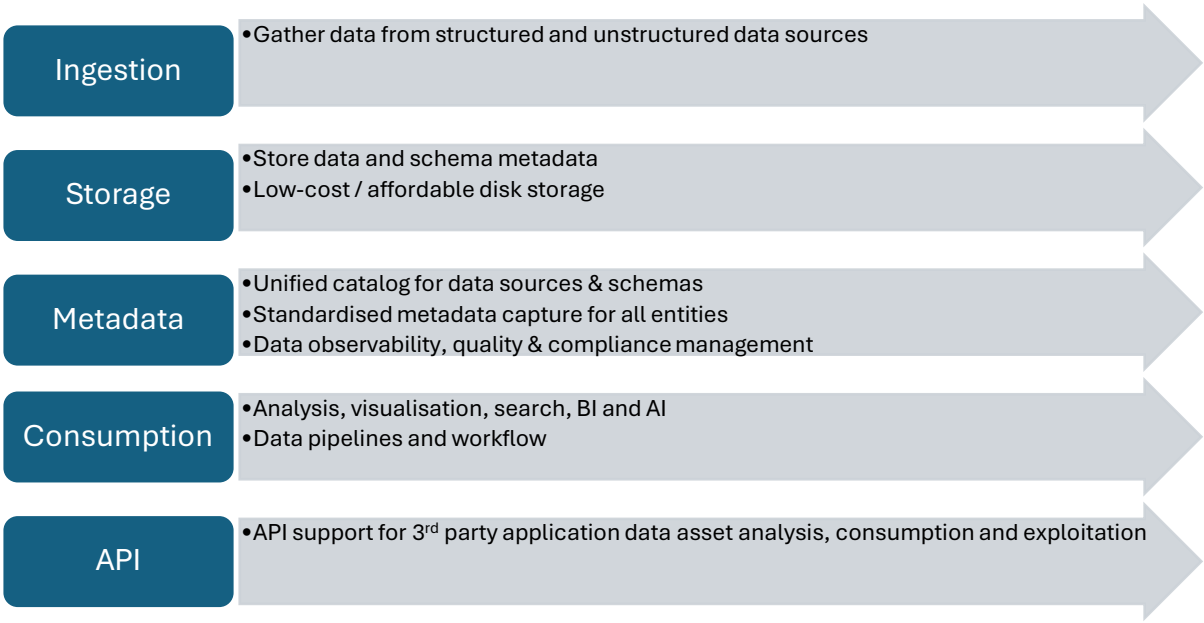
As CarlosData says, "Currently, the Data Lakehouse presents itself as a costly endeavour, with its benefits not yet surpassing the financial investment and effort dedicated to its implementation. This final assessment underscores the need for caution and thorough evaluation before embarking on a Data Lakehouse journey. While the concept holds promise, its real-world value is still evolving, and organisations must weigh the potential advantages against the considerable expenses involved."

Summarily, caution at this stage is well advised otherwise you may saddle your organisation with a high risk, high-cost project that has propensity to spiral out of control or worse still be unsustainable from a cost and capability perspective and only discovered at a later stage.

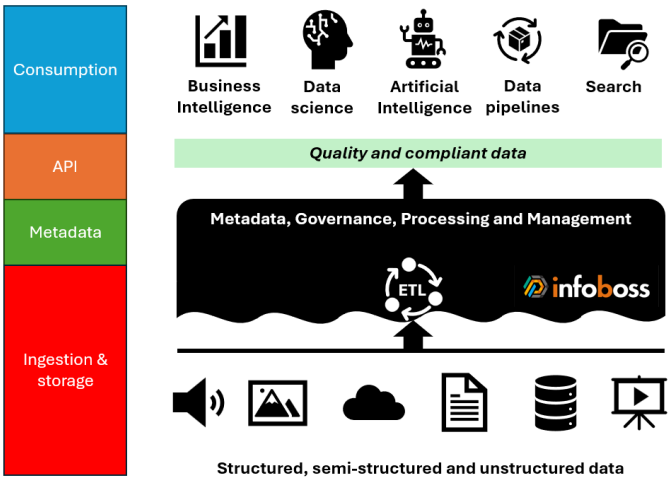
There are however innovative solutions that adopt a fresh approach to service the requirements and deliver benefits of a Data Lakehouse architecture but delivered via a single unified data platform, on low-cost infrastructure in a more cost effective way. Infoboss is one such Data Lakehouse solution...

3 The infoboss Data Lakehouse platform

According to IBM, a Data Lakehouse typically consists of five layers: ingestion layer, storage layer, metadata layer, consumption layer, and API layer. These make up the architectural pattern of most Data Lakehouses.



The infoboss data platform adheres to this architectural approach as follows:



Infoboss provides the entire suite of functionality required for a Data Lakehouse architecture within a single, secure, scalable and performant software platform. Additionally, the platform includes embedded features for managing and observing data ensuring quality and compliant data is available to the consumption layer.

The infrastructure required for an infoboss solution can be as little as a single virtual machine (Azure or AWS) with 4 x CPU cores,16GB RAM and sufficient disk storage to hold the metadata. Alternatively, as budget or IT policies allow this can be scaled up to provide resilience and greater levels of performance as required by the business.

3.1 Ingestion and storage

Getting data into infoboss is straight-forward with out of the box adaptors available for most of the typical data sources observed within the modern data estate. These typically include:

Structured

- Databases
  - Oracle, My SQL, Microsoft SQL Server, DB2, Progress etc…
  - Any database where there is a JDBC driver available for it
- CSV files
  - Either server hosted or direct upload from local disk

Unstructured

- Office 365 tenants
  - SharePoint, OneDrive, Email
- Shared drives – file servers
- MAPI/POP3 mailboxes
- OpenText

Furthermore, as data is ingested all metadata is automatically extracted and stored in the infoboss platform, creating structured tabular views of the data. Data can be loaded as a one-off or to a timed schedule to suit the requirement. Unstructured text data (and scanned documents via OCR) can be dynamically processed on ingestion to identify and extract entities such as passport numbers, client and asset identifiers and more to augment and enrich the metadata.

Data is stored in the Elastic Search engine. This is a superpower of the product as it means almost instant query times for data selection even when analysing huge volumes (millions of rows) of data. Furthermore, the ability to search across unstructured text data greatly assists compliance and other consumption use-cases.

3.2 Metadata

Infoboss enables you to standardise on the metadata to be held against all entities within the system – data areas (or projects), data sources and queries/rules. From a quality and compliance perspective you could for example hold details as to the cost to fix the issue, the business impact cost, the risk pertaining to the issue identified, ownership, fix priority and more. For data sources you could hold metadata that helps to catalog and understand the source and its fields for governance and consumption purposes. The result is a collection of data products that are quality and compliance assured, well managed and easy to access and use within the consumption layer.

This layer also allows for the establishment of data quality and compliance rules that are to be applied against the data. These can be monitored through the software and empower data owners to take responsibility for the data, building a culture of data ownership and responsibility.

3.3 API

The API layer supports the consumption layer. The most popular API available for use is ODATA as this is supported by tools such as Excel, Power BI, Qlik, Tableau, Python, R and more. The platform also allows for direct SQL query via the infoboss JDBC driver.

3.4 Consumption

The consumption layer enables the business to maximise the value from the data products that have been curated and managed within the platform. The platform itself supports some basic querying, visualisation, search and discovery capabilities, but through the API layer it affords access to quality and compliant “trusted” data for reporting, business intelligence and data science tools.

Infoboss are also able to provide an out of the box artificial intelligence capability utilising the SCALONG AI platform. This platform supports use of large language models (LLMs) for natural language processing and solutions for knowledge management, chat bots, data analysis and many more potential AI use-cases, consuming quality and compliant data from your own data estate. The platform supports the training of the model and the prompt querying of the model to give AI generated answers to questions pertaining to your data.

In addition to the SCALONG AI extension described above infoboss are in discussions with other data consumption partners that can provide such solutions as predictive analytics and other AI model use-cases.

4 Summary

Organisations are adopting Data Lakehouse solutions to harness the combined strengths of data lakes and data warehouses, enabling them to handle a broader range of data types and analytical workloads while maintaining cost efficiency, scalability, and robust data governance. This unified approach helps organisations drive innovation, improve decision-making, and stay competitive in an increasingly data-driven world.

By implementing Data Lakehouse businesses expect to achieve several benefits and a rapid return on investment. They expect that:

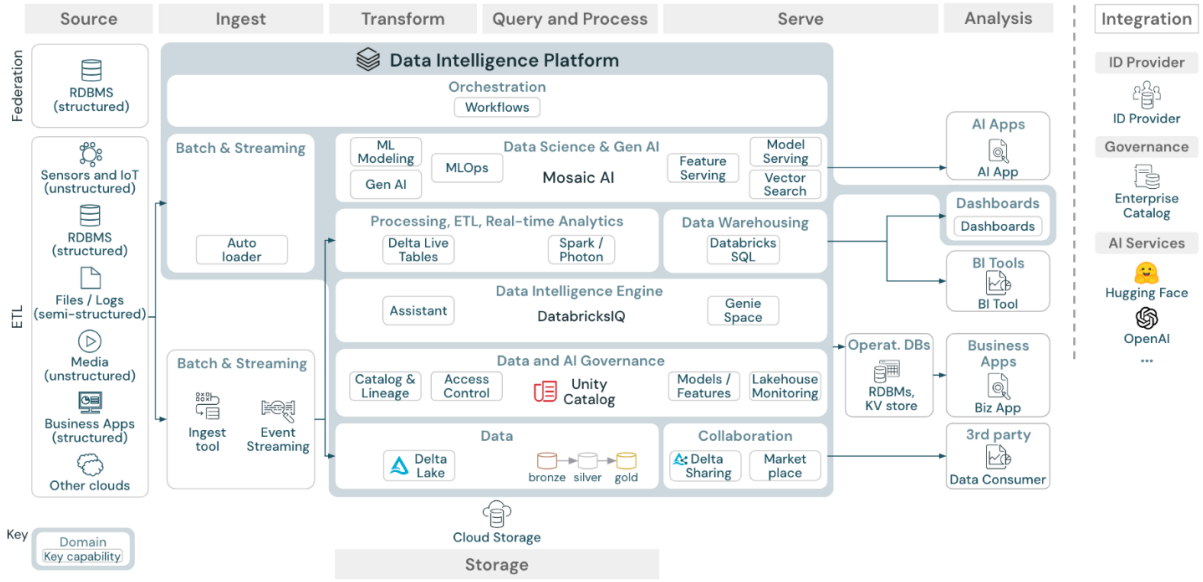
- Data from structured and unstructured data sources can be easily ingested, curated and managed; empowering their data consumers to maximise value from their data assets;
- Data served to them will be trusted, of high quality and meets their regulatory compliance requirements;
- The solution is flexible and will scale and adapt to changes in their environment in the future; and perhaps most importantly…
- There is some degree of certainty as to the build and the on-going infrastructure and resource costs to maintain and support it.

Data Lakehouse solutions such as those available from the mainstream vendors are highly complex, require significant investment and to take full advantage of them, require deep technical knowledge and expertise in the various tools they comprise. These resources are typically available in large organisations and are expensive. They are not so readily available in small to mid-size organisations or beyond their budgets and as a result this introduces significant risk to any Data Lakehouse project seeking to utilise these platforms.

An infoboss Data Lakehouse solution affords all of the benefits in one unified product. Indeed, one client suggested the difference between the mainstream platforms and infoboss is that “infoboss is a pre-made lakehouse whereas the others are like self-assembly kits”. Getting started is simple, all that is required is a virtual machine to install the software onto and you can immediately begin to ingest, curate and manage your data. Our monthly pricing means that you are able to trial the product before committing to a long term contract and our sector specific partners are available to provide the training and resources to help you maximise value from your data.

5 Appendix A – Data Lakehouse reference architectures

5.1 Databricks reference architecture



5.2 Microsoft reference architecture (for a small business)