infoboss

Data Foundations

What are they? and why do they matter when building an enterprise Artificial Intelligence capability for your business?

Data solutions. Data quality. Data compliance. Search & discovery.
Data Lakehouse. Custom data solution. Cloud migration. Data foundations. AI.

1 Introduction

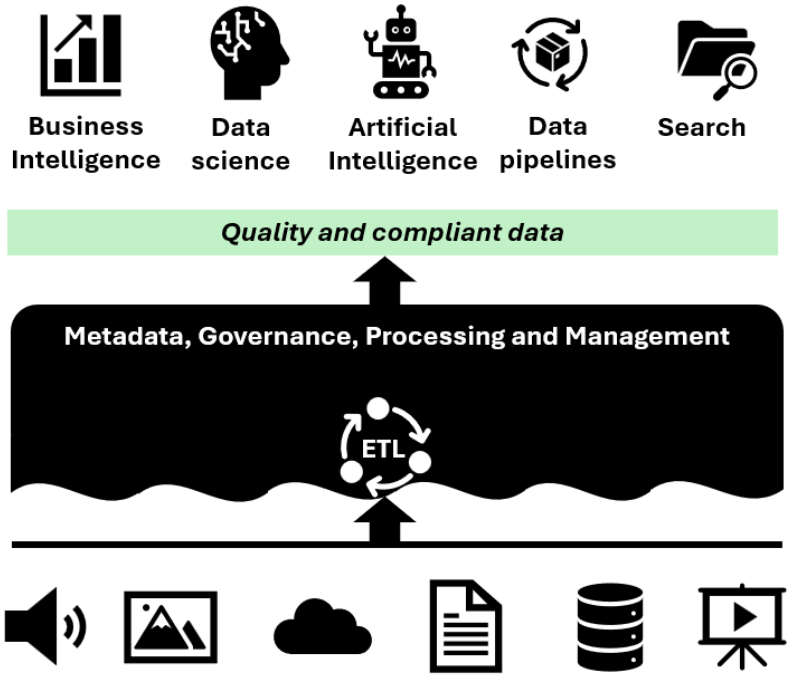
As you embark on any project that consumes and uses data - business intelligence, artificial intelligence, predictive analytics, cloud migration, digital transformation and so on, you will hear the term “data foundations”. Essentially the building blocks required to be in place to ensure fit for purpose data is availed to your initiative to ensure a successful outcome.

In this paper we discuss what’s involved in building your organisation’s data foundations and how infoboss can support you to achieve your data project objectives. We’ll discuss the topic in the context of Artificial Intelligence (AI) but the principles apply to any initiative that seeks to leverage value from your data assets.

2 What are data foundations?

“Data foundations" refers to the underlying data architecture, infrastructure, processes, and practices that are essential for building, training, and deploying AI models effectively. A strong data foundation such as infoboss ensures that an organisation can leverage its data assets in a way that is reliable, scalable, and aligned with business goals.

Here’s a breakdown of what "data foundations" encompass and how the infoboss data platform (shown) supports you in attaining them.



Structured, semi-structured and unstructured data

Infoboss Data Platform

2.1 Data infrastructure

This encompasses three topics for consideration (data storage & management, data pipelines and scalability).

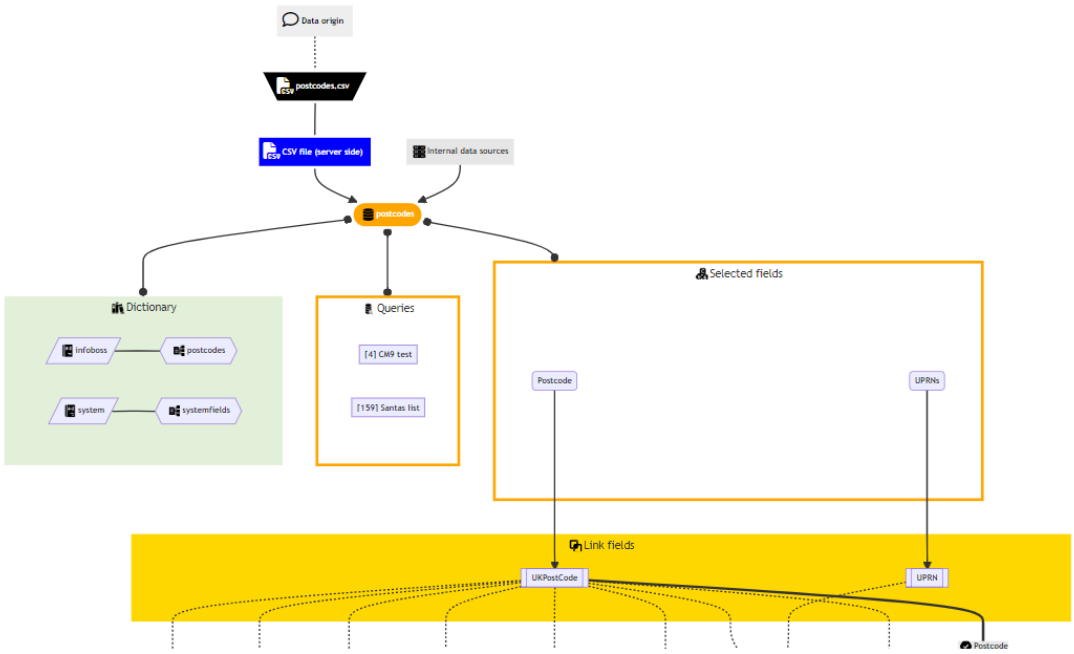
- Data storage and management: **Infoboss includes the systems and technologies for storing and managing large volumes of data, such as databases, data warehouses and unstructured data stores. Infoboss is designed to handle structured, semi-structured, and unstructured data, providing flexibility for various types of AI workloads.**
- Data pipelines: These are processes that extract, transform, and load (ETL) data from different sources into a central repository.
  Infoboss has efficient, automated pipelines to ensure that data is up-to-date, clean, and readily available for AI models. Infoboss is viewed by many of its customers as a single source of the truth. It facilitates the establishment, control and monitoring of the business data quality & compliance rules to effectively deliver fit for purpose data to consuming technologies.
- Scalable Computing Resources: AI models, especially those involving deep learning, often require significant computational power.
  Infoboss includes access to cloud or on-premise compute resources that can scale according to the data and processing needs. The software can be deployed on low-cost infrastructure and scaled to suit the requirements of your project. For example, the whole solution can be deployed on a single Windows Virtual Server machine in the cloud or on-premise.

2.2 Data quality, compliance and integrity

Perhaps the topic most highlighted when discussing data foundations as it's the one with arguably the greatest level of return on investment (ROI). According to Gartner, 85% of enterprise AI projects fail. Although there are many reasons why AI projects fail, one of the key reasons is being able to access and use fit for purpose enterprise data for training AI models.

- Data Cleaning and Preprocessing: For AI models to deliver accurate predictions, the input data needs to be free from errors, missing values, and inconsistencies.
  Infoboss has been built to empower businesses with the processes for identifying, root cause analysis of data issues and facilitate cleaning and preprocessing data processes to ensure high quality and compliance of your data.
- Data Governance: This includes policies and procedures that manage the availability, usability, integrity, and security of the data. Good data governance ensures that data used for AI is consistent, compliant with regulations (like GDPR or CCPA), and accessible to the right stakeholders.
  Infoboss is widely used in regulated industries to support organisational compliance efforts. The features within the product are designed to support best practice data governance. In the context of AI, the ability to monitor exactly what data has been used to train AI models and assess and understand its quality and compliance greatly helps management and control of the data being used.
- Data Lineage and Traceability: Understanding where the data comes from, how it has been transformed, and how it flows through the system is crucial for building trustworthy AI models.

Infoboss ensures that there is a clear lineage, making it easier to trace the origin of data used in AI training and validation. Data lineage from original source to model and processes undertaken on it can be observed using the Data Wiki feature of infoboss (example shown). This includes an interactive click and drill through interface to see and explore levels of detail.



2.3 Data Accessibility and Availability

- Unified data access: **Infoboss ensures that data from various sources—such as databases, cloud services, IoT devices, and third-party APIs—can be easily accessed and integrated. This helps AI models access the diverse data they need for richer insights. Indeed, infoboss provides the means of not only ingesting various data sources but also the means of curating new, enriched data sets by augmenting and combining data from different data sets to make them not only easier to manage but also provide greater functionality.**
- APIs and data services: These enable easy access to data for different teams and systems. APIs (Application Programming Interfaces) allow AI models to interact with data sources programmatically, making data access faster and more efficient. **Infoboss has a range of APIs to enable applications to interrogate data that has been curated within the platform and leverage the AI models for data querying and knowledge centric queries.**
- Real (or near real)-time data processing: For AI applications that require up-to-date data, such as real-time recommendations or anomaly detection. **Infoboss ingests data using its scheduling service. Data updates and refreshes can be scheduled to within a minute.**

2.4 Data security and privacy

The control of access, security and privacy considerations are paramount within your data foundation as follows:

- Data protection: Protecting sensitive data is critical, especially when dealing with personally identifiable information (PII) or financial data.
  Infoboss has a range of features to help identify data of various levels of classification and even undertake the classification process for you across all types of data.
- Access controls: Defining who can access what data is a key part of a data foundation.
  Infoboss utilises a role-based access model that ensures that data and associated AI models are only accessible to those that are allowed to see or interrogate them.
- Anonymisation and de-identification: For organisations that use customer data for training AI models, anonymisation techniques can help protect privacy while still providing valuable insights.
  Infoboss includes tools and processes for anonymising data before it is used for AI. PII and sensitive data can be tagged and automatically removed or anonymised prior to submission to models for training.

2.5 Data readiness for AI

The means of classifying, labelling and enriching data and its supporting metadata for use in AI models is an essential component for your data foundation:

- Data annotation and labelling: For supervised learning models, data must be labelled correctly.
  Infoboss includes methods and tools for annotating large datasets, ensuring that the AI models can learn accurately from the provided examples.
- Feature Engineering: This involves transforming raw data into meaningful features that can be used by AI models.
  Infoboss supports automated and manual processes for feature extraction and selection, ensuring that the models are built on the most relevant data.
- Historical Data Storage: AI models require access to historical data for training and validation.
  Infoboss ensures that past data is stored properly and is easily retrievable, which is essential for building predictive models.

2.6 Data Integration and Interoperability

- Connecting Diverse Data Sources: AI applications often need to draw data from various systems like CRM, ERP, social media, and IoT devices.
  Infoboss includes the ability to integrate these diverse data sources into a unified platform via out-of-the-box adaptors that can ingest and process the data from sources such as databases, CSV files, Office 365, Azure, AWS or on-premise shared drives and many more.
- Interoperability between tools: Different AI and analytics tools may require access to data in specific formats.
  Infoboss ensures that data can be easily transformed and made compatible with various machine learning libraries, platforms, and analytical tools. Indeed utilising

the APIs available tools such as Excel, Power BI, Python and more can directly ingest curated data from the infoboss platform.

3 Data strategy and alignment with AI goals

- Aligning data with business objectives: **Infoboss is built to support the strategic goals of the business. This means understanding what data is needed to achieve specific AI use cases, such as customer segmentation, predictive maintenance, or demand forecasting. Ensuring quality and compliant data is used.**
- Identifying key data assets: Not all data is equally valuable.
  Infoboss helps you to identify and leverage key data assets, the ones that are most likely to drive AI value, ensuring that efforts are concentrated on the most relevant information.

4 Why data foundations matter for AI...

A solid data foundation such as that provided by infoboss is crucial for the successful implementation of AI for several reasons:

- Ensures data quality: High-quality data is the backbone of any effective AI model. Without clean, accurate, and reliable data, even the most advanced AI models will struggle to deliver accurate results.
- Reduces compliance risks: By implementing data governance and privacy measures as part of the data foundation, organisations can ensure compliance with regulations and avoid legal risks when using data for AI.
- Speed up model development: With automated data pipelines, efficient data access, and a well-maintained data infrastructure, data scientists can spend more time building models and less time munging and wrangling data. The ability to build and test your enterprise AI models within the infoboss platform reduces risk and cost for your AI initiatives and helps to ensure a successful outcome for your project.
- Facilitates scalability: As data volumes grow and new AI models are introduced, infoboss ensures that the underlying infrastructure can scale to meet new demands without performance bottlenecks.
- Supports near real-time decision making: Near real-time data access is essential for many AI applications like recommendation engines, predictive maintenance, and fraud detection. Infoboss makes this possible by supporting near real-time data ingestion, processing and analysis.

5 Conclusion: Building a data-ready future

The phrase "data foundations" reflects the essential groundwork that needs to be laid to support AI capabilities effectively. It involves not just the technical infrastructure for storing and processing data, but also the processes that ensure data quality, compliance, accessibility, security, and alignment with business goals. A strong data foundation such as that provided by infoboss enables organisations to unlock the full potential of AI, allowing them to turn data into actionable insights, drive innovation, and maintain a competitive edge in the market.

infoboss.co.uk

info@infoboss.co.uk

0333 772 1963

linkedin.com/company/infoboss-limited