Insight NWT An Insight company

Whitepaper

# Use of Data in a Digital World

## Introduction

The use of data with digital applications has been constantly evolving and is becoming more sophisticated, allowing businesses and individuals to harness the power of data for better decision making, improved user experiences, and more efficient operations. However, it is essential to strike a balance between utilising data effectively and ensuring data privacy and security are maintained to build trust with users.

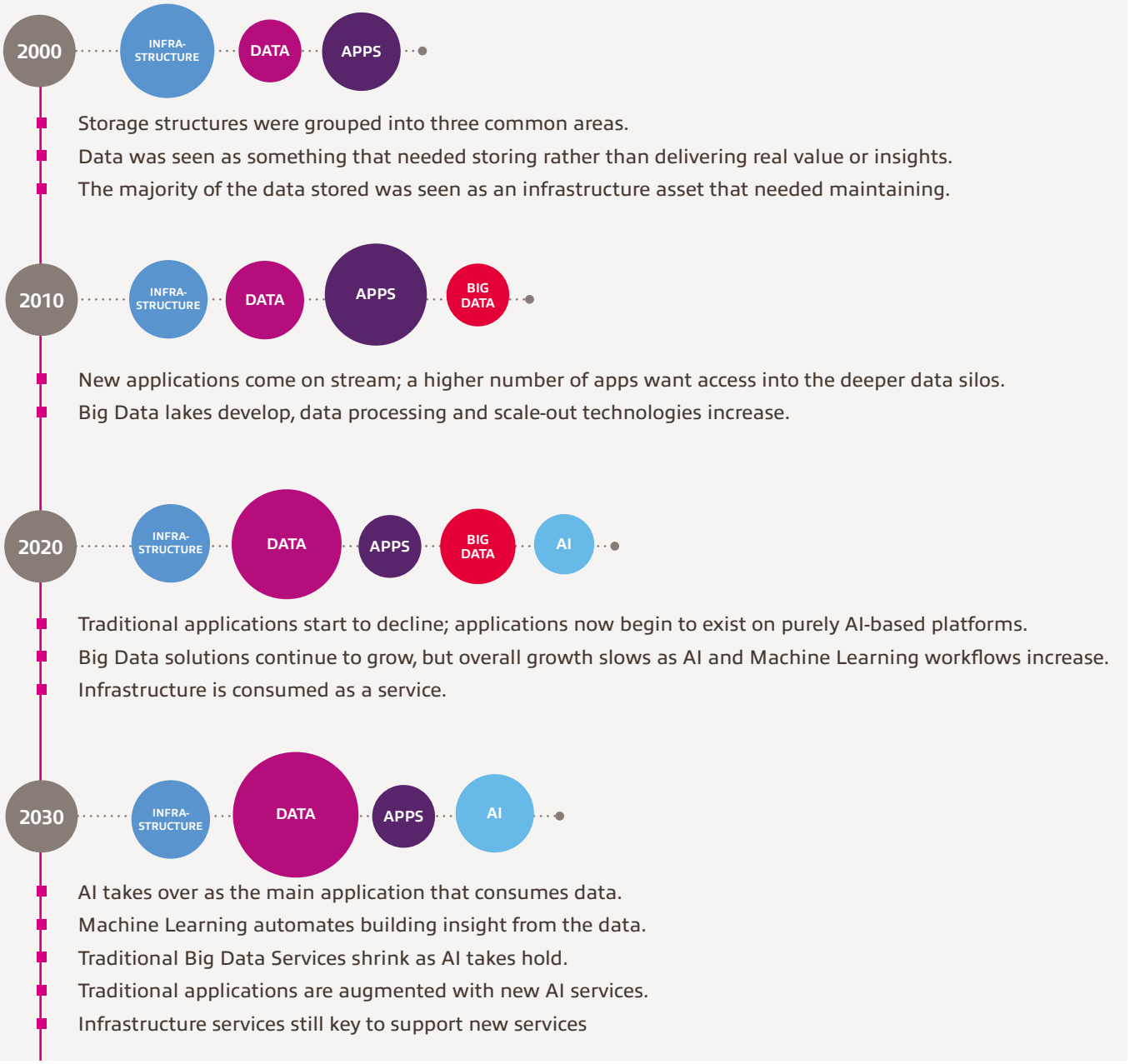
Even though 90% of the world's data was generated in the past 2 years, enterprise data continues to be managed in data silos, and liberating data from these systems is now the biggest obstacle to delivering data-driven outcomes. This is because enterprise data is fragmented in multiple systems, captive in vendor-owned applications that lack a rich API set for data access, and locked within in house legacy systems, with little or no knowhow of the underlying data models. Variably structured, or unstructured, in dozens of technologies and formats and often non-compliant. The end result is that over 80% of enterprise data remains inaccessible and unleveraged and is not driving business decisions nor is it being used to improve customer experiences or operational efficiencies.

- In today's data-driven world it is widespread to hear words like Master Data Management (MDM), Big Data, Machine Learning and Data lakes.
- Many organisations are researching different ways to deploy the most valuable data management processes and technology.
- The combination of competitive on-premises solutions, hybrid vendors offering private and public solutions along with pure public cloud providers means this market is complicated to understand and navigate correctly for the best for your organisation.
- With compliance regulations impacting the industry, e.g. GDPR, makes selecting a strategic solution and technology more complex.

# Data and Trends

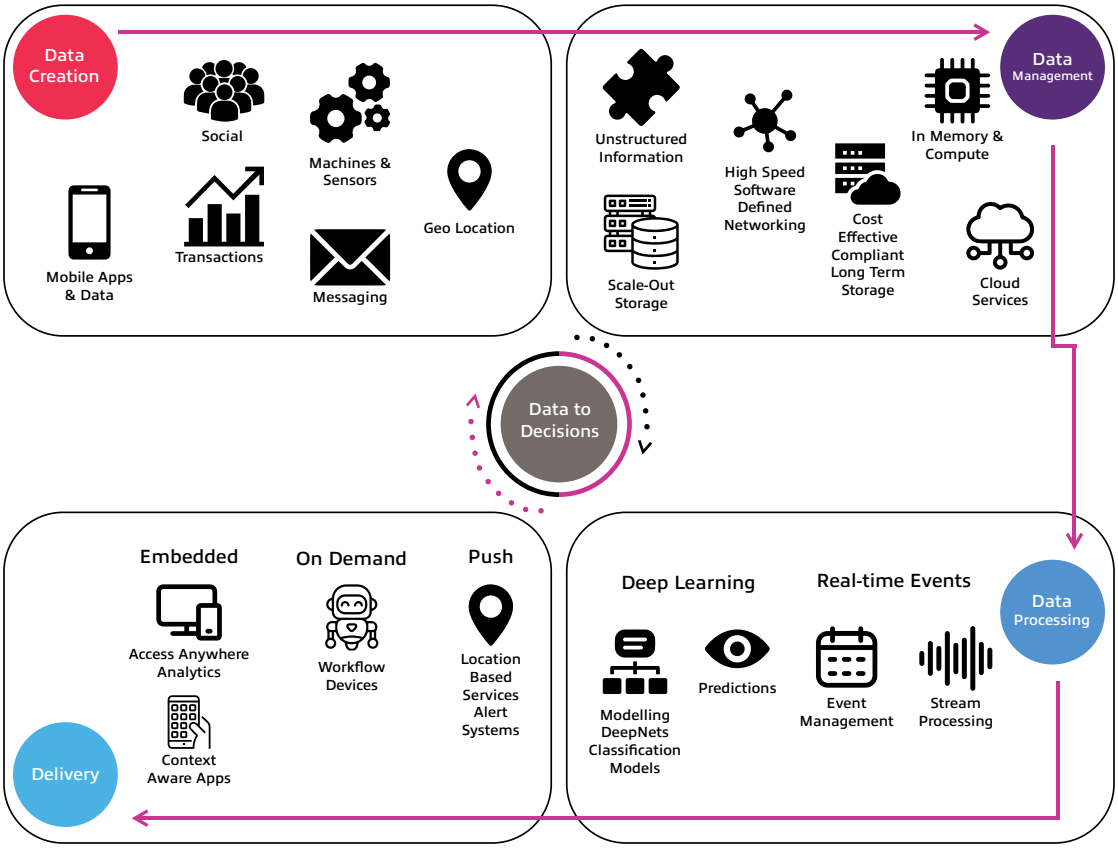
Some key trends and changes in the use of data with digital applications include the use of 'Big Data' where the proliferation of digital applications has resulted in the generation of massive amounts of structured and unstructured data. Data can now come from multiple sources including the Internet of Things (Io T) and data driven Marketing. Data is being used to enhance the Customer experience and improve personalisation. Real Time Analytics uses real-time data allowing quicker decision-making and better responsiveness to changing conditions. The one area that has generated most public attention is in respect of Artificial Intelligence (AI) and Machine Learning (ML) where AI and ML technologies are increasingly integrated into digital applications to analyse vast amounts of data and extract valuable insights. It is seen by many that the availability of data is the significant factor in driving the current interest in AI, as more data is generated and collected.

## The Evolution of Data Structures



To successfully deliver solutions in the future, organisations are going to need to learn more about the data and how to deal with it. A key factor as we move towards applications developed for the ML and AI environments is that the data is critical in defining the functionality. In AI applications, the data is encapsulated together with the code. So, real-world data is used in the development process for AI applications, and in an AI project, up to 80% of resources could be used on accessing, understanding, cleansing, preparing, and managing the data. Many organisations are still concerned about dealing with their portfolio of traditional applications and data sources, even if they are based on applications that use data as input and/or output, rather than part of the functional model.

Most organisations have produced a Data architecture which refers to the design and organisation of their data-related components within information systems or their organisation. It will encompass the rules, policies, models, standards, and technologies that govern how data is collected, stored, processed, managed, and used, and it is crucial for ensuring data quality, accessibility, security, and consistency across an organisation. But what is required within that architectural work is a data strategy that defines what data is needed and how to make it ready to use to meet the business objectives. It is about producing the plan to clean and make the data easily accessible




## The Future for Data

Many of the issues in accessing data is often traced back to the legacy siloed systems which remain critical to the organisation's business. The challenge is how to provide access to that data to improve the application development or the business decision making as provided through AI type applications. Building better data enabled solutions needs to be driven by multiple elements including the Data Architecture that enables the flow of data and the use of Data Products.

There are some stages to follow to determine what data you need, how to clean it and how to make it available. Firstly, identify the data needed from the work which has taken place on the Digital Roadmap (the Implementation Roadmap of Change). This work will consist of determining the specific data needs through identification of the Business Domains, Data Domains, Data Elements within them (name, address, etc) and the Data Products required. You will need to make prioritisation calls where necessary based on the Digital Roadmap. There will be duplication of Data Elements so decisions will have to be made on the most appropriate to deliver the use case for the Data Product. Secondly, once you have identified the data, you will need to judge its quality including consideration of many factors, e.g., accuracy, security, availability, consistency, uniqueness and coherence. Assessing and measuring data quality is a big task and many issues will arise. The good news is that many software products are now becoming available to help in this area including many AI/ML products. Finally, this will enable you to create the Data Roadmap which will help you to work through how you will deliver the priority set of data and when it will be the right quality for use.

Insight and NWT are firm advocates of the DataOps approach, a relatively new methodology emphasising the seamless collaboration and integration of various data domains within an organisation. By drawing inspiration from DevOps and Agile methodologies, DataOps aims to improve analytics' speed, quality and reliability.

## Key Principles of DataOps

**Agility:** By applying Agile principles, DataOps supports iterative development, continuous integration, and improvement. It enables teams to react quickly to changing business requirements.

**Collaboration:** It emphasises cross-functional collaboration between data scientists, analysts, engineers, and business stakeholders. This ensures that everyone has a shared understanding of the goals and responsibilities.

**Automation:** DataOps encourages the use of automation tools for tasks such as data extraction, transformation, loading (ETL), and testing. This not only increases efficiency but also reduces the risk of human error.

**Monitoring & Control:** Continuous monitoring of data pipelines and analytics workflows ensures that issues are detected and rectified early. Metrics and dashboards provide visibility and control over the entire data lifecycle.

**Data Quality & Security:** Ensuring data quality and security is fundamental to DataOps. By implementing rigorous quality checks and adhering to security standards, it builds trust in the data and its insights.

## Key Benefits of DataOps

**Faster Decision Making:** With streamlined processes and collaboration, data becomes more readily available, enabling quicker insights and decision making.

**Improved Data Quality:** Automation and quality checks minimise errors, leading to more reliable data.

**Scalability:** The DataOps approach supports scaling as per business needs, making it adaptable to various sizes and types of organisations.

## Data Architecture

In simplistic terms, Data Architecture is the system description of where data is stored, used, and how it gets there. Data Architecture will determine how data is stored, transformed, cleansed, analysed, and consumed. If the Data Architecture is incomplete, then you will be left with silos of data, especially in your legacy systems, rather than enabling effective data integration and utilisation across the organisation. It is possible to design a system that allows data from disparate and often outdated sources to be integrated, accessed and utilised efficiently.

While it is generally preferable to move towards a more unified and integrated data environment, there are situations where legacy data silos need to be maintained due to various constraints, such as budget limitations, technical challenges, or regulatory requirements. There are multiple options for designing a data architecture that accommodates legacy data silos, such as Data Integration, ETL (Extract Transform Load), Data Virtualisation, and data APIs.

Building on the principles of Data Architecture, in today's Distributed Data Architectures, APIs are the glue that hold services, applications and data together, API development has accelerated with frameworks such as Django, FastAPI, hyper-scaler platforms providing API gateways and serverless functions for scaling microservices and low-code tools and platforms. A relatively new approach to API development is GraphQL, developed by Facebook it has been gaining traction in large-scale data operations due to its many advantages.

## Benefits of GraphQL in Large-Scale Data Operations

**Flexible Queries:** Clients can request exactly what they need, and nothing more, reducing over-fetching or under fetching of data. Enables front-end and back-end teams to work more independently.

**Strongly Typed Schema:** The schema provides a contract that defines how clients can access the data, enabling robust validation. Facilitates API documentation and auto-generation of client libraries.

**Performance Optimisation:** Reduces the need for multiple roundtrips to fetch data by allowing complex queries in a single request. Enables server-side optimisations, like batching and caching.

**Real-time Data with Subscriptions:** Allows real-time updates via subscriptions, enabling live data synchronisation across platforms.

**Introspection:** Provides tools to query the schema, allowing for enhanced developer tooling and debugging capabilities.

## Case Studies

Several large organisations have benefitted from implementing GraphQL in their data operations:

- GitHub: Adopted GraphQL to provide a more flexible and efficient API for developers.
- Shopify: Utilised GraphQL to enhance mobile application performance and developer productivity.
- Airbnb: Employed GraphQL for optimising data fetching, improving user experience, and reducing development time.

## Data Domains

A Data Product is a ready-to-use set of data that the organisation can easily access and use. It is usually connected to a Data Domain, which is a grouping related to business domains (e.g., Sales, Customer Service). The Data Architecture will often use the same Data Domains. Organisations can define their data domains based on their specific needs and the nature of their data. The process of defining data domains involves identifying the different categories of data relevant to the organisation's operations and business processes. Once data domains are established, they become a fundamental building block for various data management practices, ensuring data consistency, reliability, and usability throughout the organisation.

One innovation that Insight and NWT are exploring is how to bring Generative AI to Data Domains to bring additional value to existing data that is contained within the Enterprise.

The explosion of interest in GPT technology has been mind-blowing, but while in a consumer world, the use of public base models is okay, in the Enterprise domain, it is not practical to use a Public LLM. To bring LLMs to private data domains, as we have described in previous sections, approaches and tools need to be developed to take Public Base models and train them on domain-specific data.

At the point of writing this paper, three different approaches are seen in the industry:

**Training an LLM from Scratch:** Creating a domain-specific model from scratch is not currently a common approach since it requires a massive amount of high-quality data and large-scale computing power in the form of specialised GPUs.

**Fine-Tuning an Existing LLM:** A second approach is to "fine-tune" an existing LLM to add specific domain-level content to the system that has already been trained on general knowledge. This approach involves adjusting some parameters of a base model and typically requires substantially less data.

**Prompt-tuning an Existing LLM:** Perhaps the most common approach to customising the content of an LLM for non-cloud vendor companies is to tune it through prompts. With this approach, the original model is kept frozen and is modified through prompts in the context window that  domain- specific knowledge. After prompt tuning, the model can answer questions related to that knowledge. This approach is the most computationally efficient of the three, and it does not require a vast amount of data to be trained on a new content domain.

**Use cases for LLM and data:**

- Search
- Summarise Data
- Extract and expand on existing data
- Sentiment Analysis
- LLMs can classify unstructured data with unmatched accuracy, flexibility and scalability.

## Data Product

A Data Product, sometimes referred to as "data as a product" (although there is a difference), is a reusable data asset, built to deliver a trusted dataset, for a specific purpose. It manages data like a consumer product by delivering high-quality, ready-to-use data. It collects data from relevant data sources, processes it, ensures data quality, and makes it accessible and understandable to people and systems across the organisation who need it to meet specific needs. It can be used to meet different business challenges and an emerging use is as the core of a digital twin, replicating the operation of real-world assets and processes. In summary, Data Product is the result of applying product thinking into datasets, making sure they have a series of capabilities including discoverability, security, ability to be explored, understandability, trustworthiness, etc.

The use of Data Products is a completely different way in which organisations can begin to design, make data available and manage it. A data product makes a dataset easier to understand, easier to discover, and easier to access as a data asset. Data Products incorporate the means for each Business system to consume the data even though each system will be determining how the data is stored, processed and managed.

The benefits of this approach can be significant. Organisations that build high-quality data products see improved efficiency and collaboration and their product and data teams are generally better informed as to the value and end use of data. Teams that use data products spend less time searching for data, ensuring data quality, or building new data pipelines, and those time savings become significant when added up across your data ecosystem and lifecycle. Total cost of ownership, including technology, development and maintenance costs can decline by an estimated 30%. Additionally, data products speed time to insight because they can be reused and repurposed, increase trust in your organisations' data, and provide real-time data for in-the-moment decision-making.

Examples of data products include datasets, data streams, data feeds, or APIs; code or data models; analytics models; and dashboards. Within Insight we have used Dashboards to provide useful metrics using the digital products empowered by Tableau. In the era of moves to low/no development, the use of Data Products can empower even non-technical users to gain valuable insights via subsequent data analytics, as they make operational datasets easily accessible across the company.

What makes an operational data product so special, is that its dataset is always:

- Unified, and complete, for any business entity.
- Up-to-date, and enriched with operational intelligence.
- Protected, compliant with privacy regulations and properly governed.
- Accessible in real time via data services and a wide range of data delivery methods.

We believe that building Data Products is an essential step of the journey to a successful Digital Transformation. However, ensuring the Data Product will add business value and be relevant to its users, and the datasets are well understood is key. Identifying the priority Data Products and then constructing the roadmaps for when they deliver to support the overall Digital Roadmap will help you to set up the organisation which will be responsible for delivering them. Data Products will have dedicated pods within the Digital Factory and the skillsets required will be very specific to the data and purpose of the Data Product.

## Data Operating Model

Finally, consider the overall Operating Model including governance and how this will apply to managing data in the digital world. Do you need a Chief Data Office or Officer? Do you need an Ethics Board particularly in relation to AI? How will you recruit and retain the talent needed to drive this change? How will 'DataOps' work including both testing and operations?

Getting the governance piece right is key, particularly in complex data landscapes. Information governance uses a set of defined roles, processes and policies to help manage data assets and ensure their integrity, accuracy and security. Without these structures and controls, data assets lose much of their strategic value. Without effective information governance, no one can be certain about what data assets an organisation has, who controls them, what information they can provide and how they should be used.

There is a lot to consider but it is difficult to provide answers as every company will have a different starting point and may already have implemented major changes to their Operating Model as part of the Digital Transformation. Whatever the approach that you adopt, keep key stakeholders involved on how the data will be managed and make sure everyone involved understands their roles & responsibilities.

# Why Insight and NWT

Insight and NWT partner with organisations looking to transform their data into actionable insights for informed decision-making. Our expertise spans across various domains including Data Strategy, Architecture, and Governance, Data Platforms, Data Migration, Data Integration, Automation, Artificial Intelligence and Machine Learning.

Our services are designed to provide best-practice advice on system integration, engineering data (design and deploy), and technology operations, ensuring the best return on investment. We understand the complexities of data compatibility, interoperability between different technologies and platforms, scalability, performance and latency, as well as security and privacy.

We also recognise the legal and regulatory concerns that come with the use of advanced technologies such as AI and ML, particularly related to data privacy and security. We are well-versed in addressing these issues, ensuring compliance with regulations such as GDPR.

Moreover, we understand the ethical concerns that arise with the use of advanced technologies that rely on large amounts of data, including personal and sensitive information. We are committed to addressing these concerns carefully and responsibly.

In a nutshell, Insight and NWT are dedicated to helping organisations navigate the complex landscape of data management and digital transformation, ensuring they can leverage their data effectively while maintaining the highest standards of security, privacy and compliance.

For any questions, please call: 0344 846 3333

Option 1 - Public Sector | Option 2 - Sales



uk.insight.com