NTT

The Evolution of Data Management

A practitioner's perspective

Authored by Matthew Drayson and Amjad Bashir

Continue

Introduction

It goes without saying we're living in **unprecedented times.** Driven by extraordinary market disruption that we're experiencing right now and the digitization of business, **we're seeing the explosion in the volume, speed and variety of data.**

Most organizations are beginning to see – or have already recognized – that data is a critical ingredient. Yet, while many have invested in technologies to handle the torrent of data flooding in today, they still struggle to achieve the outcomes they want. Why? It all boils down to data management – having a suitable data architecture to meet the needs of the organization backed by a data driven business culture.

In this e-book, we review the changes we have seen in the way data is collected, organized and distributed. We also explore different data management architectures and their relevance to today's intelligent business, based on our experience as practitioners working with and helping organizations.

Finally, we provide some key considerations for designing a data fabric that will ensure that the right data – that's integrated, well governed and curated, of the right quality and with full provenance – delivers the outcomes your organization seeks.

The evolution of data management

The **advancement of technology in the past 60 years has changed every facet of our lives;** and the pace of change only seems to be accelerating. In the world of the internet, social media and IoT, the volume, speed and variety of data **continues to increase exponentially.**

Our own observations of this evolution from the perspective of the impact on data has led us to characterize it into three distinct ages – the analogue age, the age of the digital veneer and the age of digital orchestration. Today, many organizations still rely on legacy systems – or a blend of systems from different ages – and are at different stages in their data management journey.

The analogue age (1960s to 2007)

Automation was the primary outcome of this age. Because compute power was expensive and limited, carrying out tasks such as operational reporting had massive impacts on online performance. To combat this, reporting was batched up and run outside of business hours. Think mass transactions such as travel reservations, stock market trading and bank processing.

The invention of relational database technologies meant that data from these mainframe systems could now be copied into a replica that allowed business users to run their own reports on demand. Many of these technologies are still used in business today.

The problem? For the large part, data was – and still may be – captured and processed in data silos run by different business units. This makes it difficult to gain a single view of the customer, as the only data captured is transactional data. As the drive for an integrated view of the business and a single view of customers became a priority, the replication and integration of data into a single repository – the data warehouse – become popular.

> Getting an integrated cross-department view of the enterprise **used to be impossible.**

With the advent of the data warehouse and the rise of specialized Extract, Transform and Load (ETL) tools, enterprise descriptive analysis was possible for the first time. We could now 'observe' the business. Despite this, analysis of data is mostly constrained to the descriptive (what happened) and the diagnostic (why it happened) because the data is latent – often from the day before – rather than being available in real time.

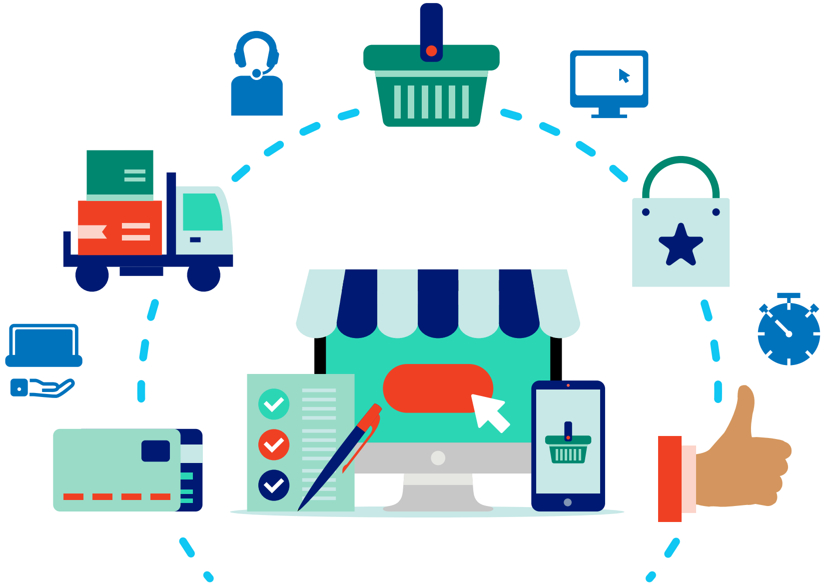
The digital veneer (2007 to 2017)

This era saw the arrival of many digitally native businesses whose systems were designed specifically for purpose. In contrast, older businesses with a huge investment in legacy systems did not have the luxury of building from the ground up. Instead, they were faced with implementing purpose-built e-commerce systems and connecting them to their existing transactional and reporting systems via APIs and microservices. They simply added a digital veneer.

As digital natives (those who've grown up in the internet age) and tech-savvy digital immigrants (those born before the internet age) have embraced the digital world, organizations had to scale their applications for millions of concurrent users. New sources of data – such as social media and IoT – led to the need to store and process unstructured data. It also introduced the need for distributed architectures that could support real-time analytics.

This led to the invention of new data store architectures such as Hadoop file systems – often referred to as data lakes – and NoSQL databases. Analysis of data evolved from being retrospective (what happened) to being predictive (what might happen) and prescriptive (recommending an action based on a forecast).

This was made possible through the creation of Machine Learning and Artificial Intelligence – all of which is supported by cheaper compute and storage and, in many cases, cloud computing.



The ecommerce flow

The digitally orchestrated age (2017 onwards)

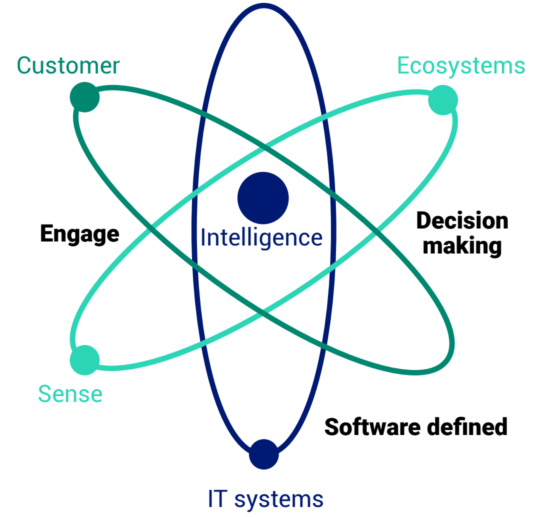
As organizations mature their e-commerce and other systems and start to develop or redefine their digital business strategies, we're beginning to see an increasing demand for a convergence of analytical and operational workloads. This is manifesting in the form of real-time automation and messaging driven by machine automation. The resulting solutions have marked the transition from the age of the digital veneer to the age of digital orchestration.

Today and in the future, business value will **be measured on actual monetization of data.**

What we're describing here is the creation of a data fabric with intelligence at its centre. A data fabric enables seamless processing, management, analysis, and storage of almost any amount of data from a multitude of siloed sources. The data fabric then enables applications and tools frictionless access to that data using an array of interfaces. It's also important to note that data fabrics leverage data in real time.

A well-designed data fabric directly connects software-defined IT systems with customers who engage through multiple channels, edge-based devices (the internet of things) and business partners (via APIs).

The intelligent business today is data-driven at the core, connected, digital and secure. Traditional businesses still stuck in the analogue or digital veneer ages will need a step change to embrace data and the challenge of its management. And the time to act is now.



The data fabric

Understanding the different data management architectural patterns

In this chapter, we explore some of the key data management architectures and highlight how their different capabilities can help you meet the needs of your business.

Data warehouses

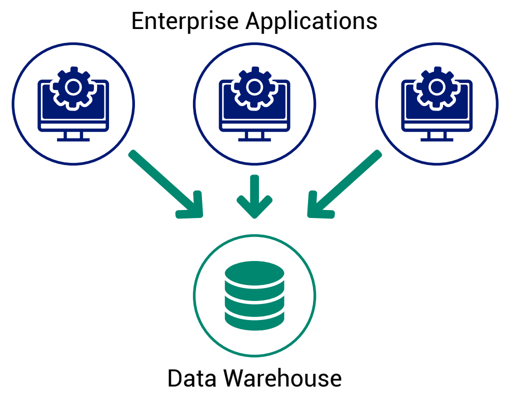
Data warehouses are the most well-known and common data management architecture. They are generally built on relational databases and have one or multiple of the following data modelling approaches:

- **Normalized** – used to address efficient storage using modelling constructs entity and relationship; the first modelling approach used by the father of data warehousing, Bill Inmon in the early-1980's.
- **Dimensional** – popularized by Ralph Kimball in the mid-1990's to address performance and ease of use, using the modelling constructs dimension and facts.
- **Data Vault** – a hybrid of the first two conceived by Dan Linstedt in the 1990s. It was released in 2000 as a public domain modelling method to address changing business requirements and improve load performance using the modelling constructs hub, links and satellites.

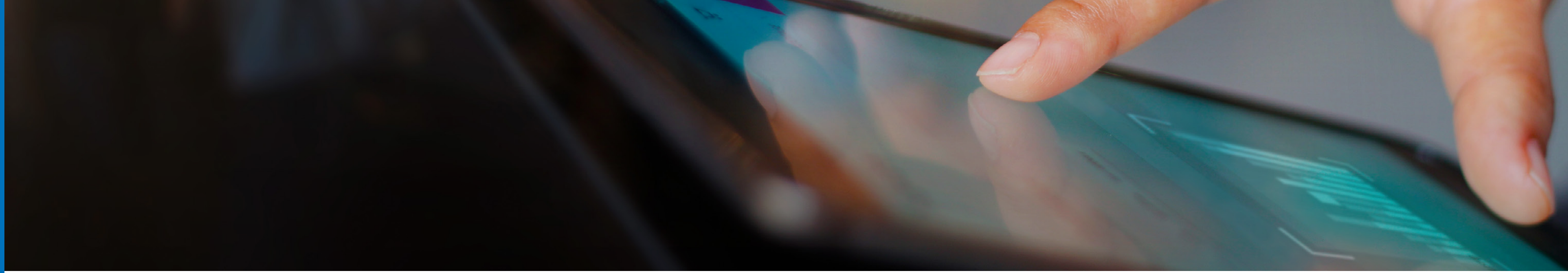
The outputs from data warehouses are often consumed by business intelligence tools to enable self-serve diagnostic analysis and visual dashboards.

This has put data in the hands of more citizen users and decision makers, enabling them to respond to changes in the business and external forces faster than before.

The downside of data warehouses is often that the same information is captured in multiple schemas and cubes leading to duplication of data and the possible inconsistency of results depending on how the ETL programs have been built.



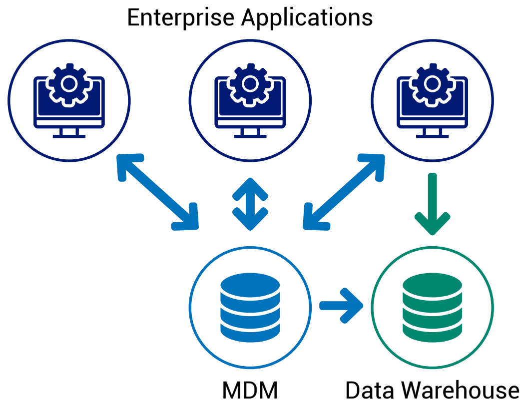
Data Warehouse



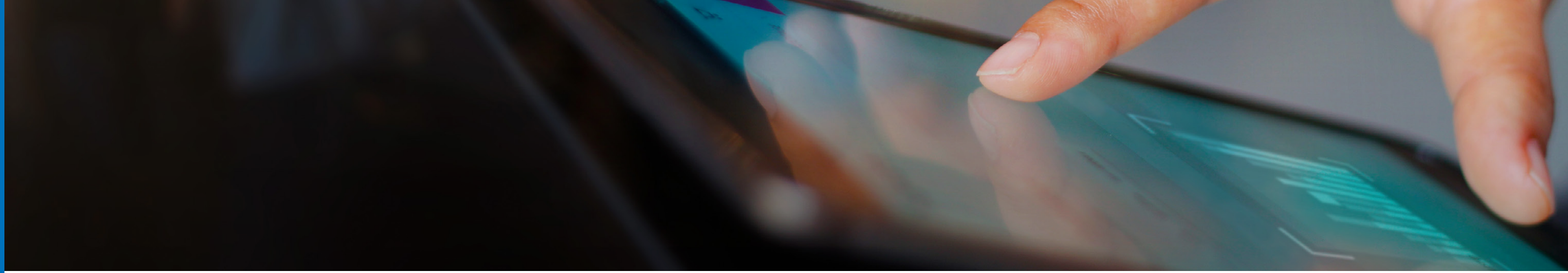
Master Data Management (MDM)

As technology has evolved, most organizations moved from single monolithic mainframe business systems to diverse, purpose-built applications. In many cases, each system is designed for the needs of just one department or business unit. As a result, it's not uncommon for entities such as a 'customer' to be recorded in multiple systems and, in many instances, for these details to be keyed in manually.

The proliferation of these departmental applications and the duplication of key entities led to the requirement for what became the first kind of a data hub – the master data management (MDM) platform. MDM solutions are used to support master data management by removing duplicates, standardizing data (mass maintaining), and incorporating rules to stop incorrect data from entering the system. The end result is an authoritative source of master data.



Master Data Management



An MDM platform aims to collect, aggregate, match, consolidate, quality-assure, persist and distribute data throughout an entire organization to ensure this data is maintained, controlled and applied consistently and accurately.

For example, an MDM platform enables a single view of an existing customer who's previously purchased products from a bricks and mortar store and is a loyalty member, who then goes online to purchase a product from the same company's e-commerce system. Instead of creating a brand-new customer record, first the e-commerce system will attempt to match the customer's details with any existing records held in the MDM system. If a match is found, it'll flag that record as the same customer and provide any missing or correct details to the e-commerce system. From there, the MDM system will create a link of the system id created in the e-commerce system with any other ids from other systems pertaining to that customer record.

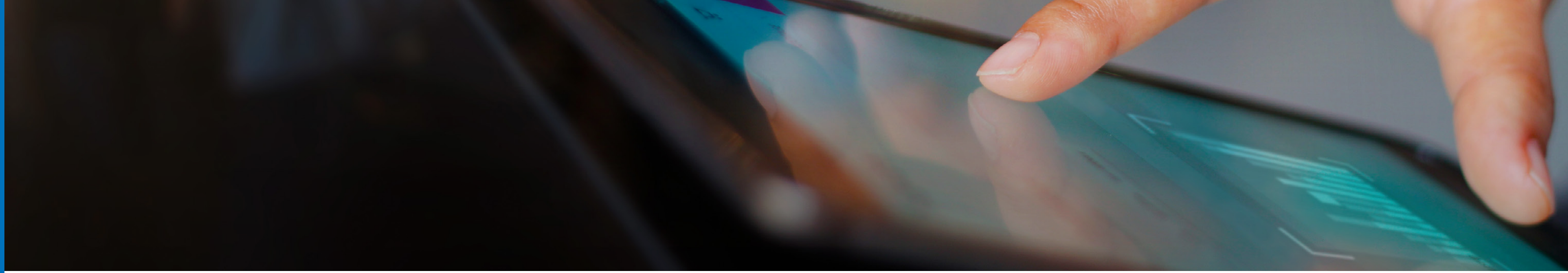
The outcome is a single and consistent view of the customer or any other key entity across the various systems in the organization where those entities are commonly shared.

There are lots of complexities around the matching and merging of master data. As a result, strong data governance and stewardship processes are required to design the survivorship rules and manage outlier records that matching rules cannot address.

This is mainly due to the fact that matching rules are defined on known data at the time of configuration. When new and different data is processed, it often leads to exceptions that cannot be automatically handled.

To address these challenges, organizations often need to build dedicated teams to manage the MDM rules and constantly evolve them. Machine learning and AI has enabled many of these platforms to become self-learning and adaptive; however, many organizations struggle to implement successful MDM platforms.

Strong data governance and stewardship processes are **required to design the survivorship rules and manage outlier records** that matching rules cannot address.

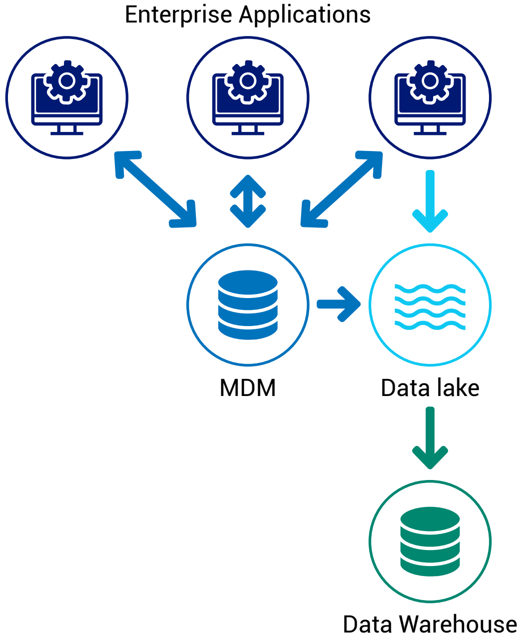


Big data and data lakes

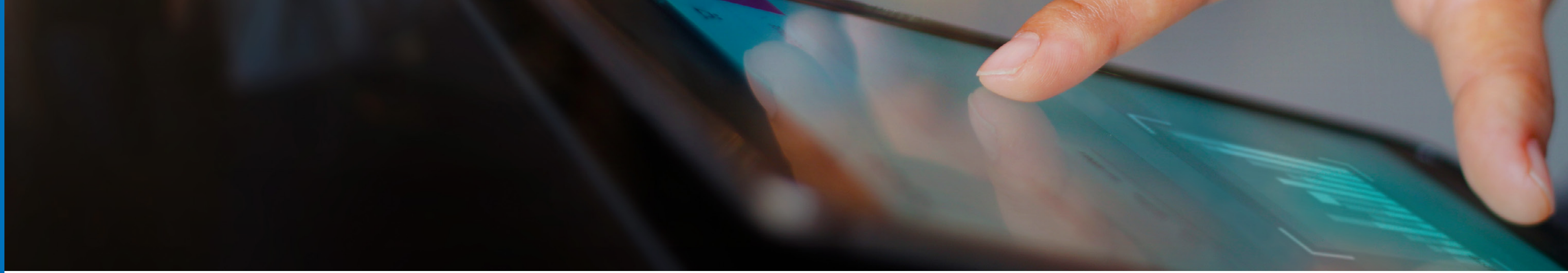
As organizations such as banks started to require scale to perform analysis of large transaction sets, appliance-based massive parallel processing (MPP) databases came into existence. However, these are predominately designed to manage structured data.

The problem is, today we're seeing large volumes of data being generated in a variety of types at increasing rates of velocity – known as the 3Vs of big data. The data is being generated from all sorts of channels – websites, smart device apps, call centres, chat bots, virtual assistants, voice assistants and social media. Connected IoT devices are also driving the increase in big data. Much of this data is unstructured – including text and media files such as voice, images and video streams. This data is not compatible with the typical relational databases used for standard data warehouses.

Enter the data lake – essentially a distributed file system offering relatively cheap, large and scalable clustered storage, also known as Hadoop Distributed File System (HDFS). A HDFS is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications.



Data Lake



An HDFS is good for performing large and complex analytical workloads and new script-based languages, such as Python, have been created to perform advanced analytics.

Now, with the emergence of machine learning and AI, predictive and prescriptive analytics have now become the norm. We've also seen new types of databases that can handle unstructured data in different ways – NoSQL and graph databases are examples of these.

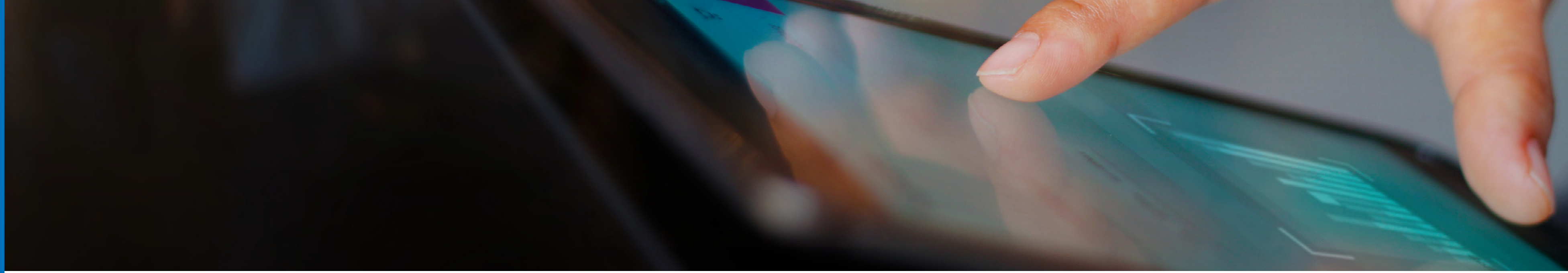
While big data technologies are not the silver bullet, they certainly have enabled a broader range of analysis, further complementing traditional data warehousing and enabling exploratory analysis of unknown questions on unknown data sets. This has led to a new challenge – organizations are now collecting data at a faster rate than they can catalogue, analyse and action.

Beware the swamp

In the early days of the digital veneer age, many organizations recognized the need to capture data, even if they didn't have the resources and budget to process the data and gain immediate insights. This resulted in data swamps – large data lakes filled with raw, uncurated and siloed data – that are underutilized and most likely unusable.

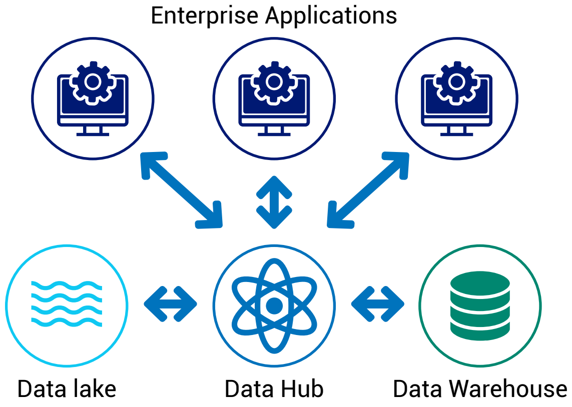
Gartner defines dark data as 'the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes.' It includes all data objects and types that have yet to be analysed for any business or competitive intelligence or aid in business decision making.

www.gartner.com/smarterwithgartner/how-to-tackle-dark-data/



The data hub

We've seen the recent rise in popularity of the data hub. This is a logical architecture that enables data sharing by connecting producers of data (applications, processes and teams) with consumers of data (other applications, processes and teams). Endpoints interact with the data hub, provisioning data into it, or receiving data from it, and the hub provides a point of mediation and governance, as well as visibility into how data is flowing across the enterprise.



Data Hub

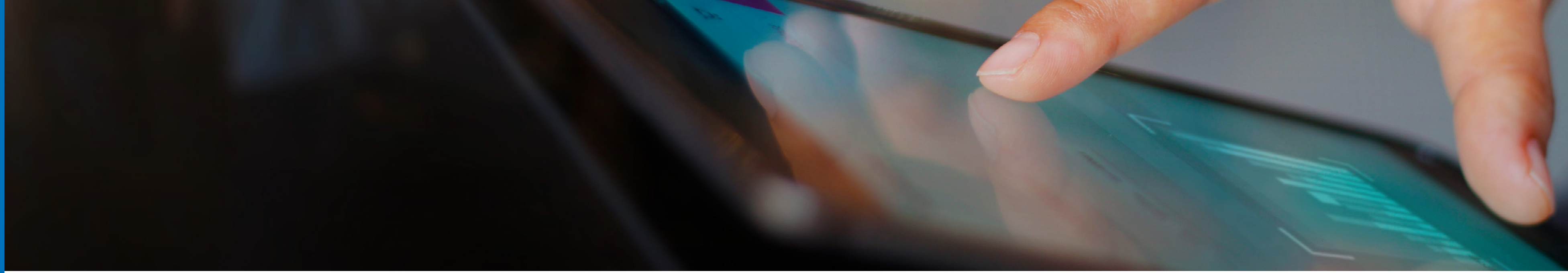
A data hub is not an analytics vehicle on its own but can be considered as part of an analytics ecosystem.

> It's important to understand that data hubs, data lakes and data warehouses **address different use cases and are not a replacement for each other.**

As described previously, an MDM solution is a form of data hub – a master data hub. There are also other kinds of data hubs to manage reference data, metadata and the integration of data for operational use cases.

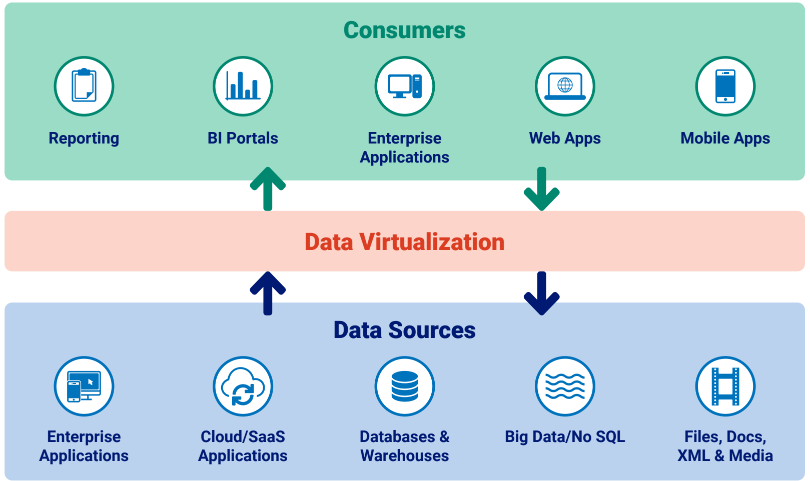
In other words, there's not 'one hub' for all use cases; there may be different hubs in the organization. MDM is a hub for governance and mediation of master data, whereas an integration hub is a 'data type' agnostic vehicle to do the same for all kinds of data. An operational hub, on the other hand, is used for real-time ingestion, harmonization and the indexing of structured and unstructured data to enable the sharing and searching of operational data.

As the industry evolves, we're seeing technology vendors cater for multiple data hub use cases within the one platform.



Data Virtualization

Virtualization is not a new concept in the IT industry. Nowadays, almost everything can be virtualized, including processors, storage and networks. Data virtualization is the process of offering data consumers an interface that hides the technical aspects of the stored data, such as its location, storage structure, API, access language or storage technology.



Data Virtualization

Data federation is a form of data virtualization where the data stored in a heterogeneous set of autonomous data stores is made accessible to data consumers as one integrated data store, by using on-demand data integration.

The biggest challenge we see with data virtualization and federation is persistence and integration. These technologies rely on data residing in the source systems and therefore must query and, at best, cache data to serve information. If complex integration and temporal use cases are required, then a data warehouse, data lake and/or data hub will be required.

That being said, virtualization may form part of an overall solution involving these other architectural solutions. In our experience, where we see virtualization tools coming into their own is as a semantic layer providing an integrated view across a number of data repositories. Because these tools have their own metadata repository, they offer data lineage capabilities and offer query performance across repositories by caching queries. In doing so, they provide a dictionary and map to business users. They enable an easier way to access, understand and navigate data across the different repositories used by the business.

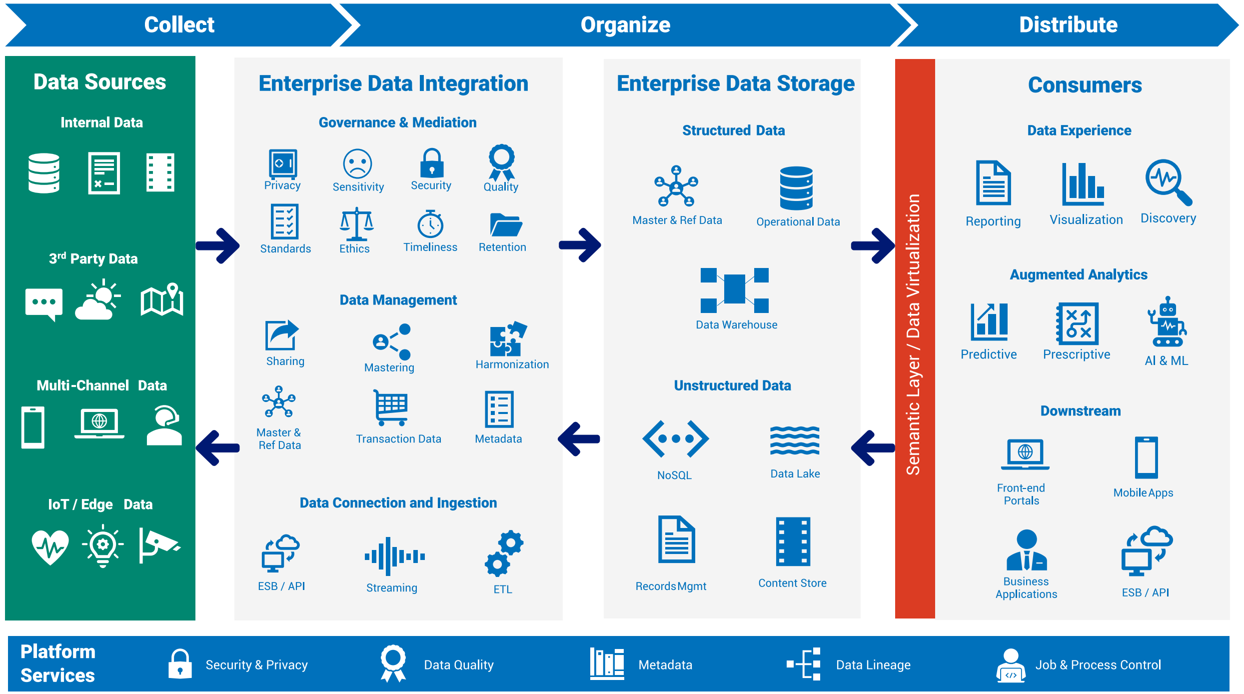
Choosing the right architectural pattern

We have now looked at the key architectural patterns for data management. Each of them provides solutions for different use cases, and sometimes a combination may be required. When looking at your current requirements for your data fabric, it's important to understand what architectural patterns will meet your organization's needs. To summarize, here are some key considerations:

	Usage	Performance	Data	End users
Data Warehouse	Use as the default choice for reporting, data analysis and business intelligence applications.	Good for performing historical trend analysis with temporal (point in time view of information) data that is persisted in the warehouse.	Use for combining data from one or more disparate sources into an integrated schema – data processing is schema on write (it must be transformed to fit the target structure of the data warehouse tables).	Most end users are business users.
Data Lake	Used to store a vast amount of raw data (structured or unstructured) in native format (a plus is that it can be stored inexpensively).	The data structure and requirements are not defined until the data is needed – processing of data for a lake uses schema on read (structured is applied to the data only when it is read).	The data does need not be explicitly harmonized, indexed, searchable or curated.	Most end users are data scientists or data modelers with business domain knowledge and who can interpret raw data structures.
Data Hub	Use where there is a need to share data between applications (both operational and analytical) in a governed, quality managed and secure way.	Persistence may be required but isn't necessarily mandatory. Instead, the modern data hub is a gateway through which data moves, virtually or physically. Data is often curated, indexed and searchable when persisted.	Data processing is a combination – can be schema on read like a data lake (in the case of an integration hub) or schema on write like a data warehouse (in the case of an MDM hub).	End users are generally other applications rather than humans.
Data Virtualization	Use where there's a need to provide an integrated view of data to support an application – for example, an external facing customer portal or reporting application.	Persistence is not required.	A semantic layer and possibly data lineage and query optimization are required across a number of different data repositories – i.e. a data warehouse, data lake and/or data hub.	Most end users are business users.

Designing the architecture

Through our experience of working with customers, we've developed a reference architecture to support these precise data management needs. This includes all the key components for a modern data management ecosystem, such as operational systems, data warehouses, MDM, data lakes, etc. This reference architecture works whether it be on-premises, in the cloud or a hybrid of the two. If designing your own data fabric, it's important to consider your current and future needs. And while it's important to approach this with an agile mindset, you also need to consider the long game.



NTT Ltd.'s data fabric reference architecture

Some thoughts before starting the journey

Data management isn't one of the most exciting or innovative activities, but it's vitally important for businesses seeking to harness the benefits of machine learning and advanced analytics. Many organizations still have their data in silos and face significant challenges with data governance and quality.

The old saying 'garbage in, garbage out' is appropriate here because if you have bad data, your analytics and insights will be flawed. This will, in turn, lead to misinformed, poor decisions, which can do significant damage to your organization. Conversely, when your data is managed properly, analytics and automation can absolutely transform the abilities and provide substantial benefit to your business.

Traditional data management practices haven't delivered agility, speed to delivery, or innovation – all of which are at the heart of many of today's organizational challenges. Delivering true data agility and integrated insight requires a fundamental shift in your technological approach. This means building data-centricity through the transformation of your data fabric.

> When your data is managed properly, **analytics and automation can absolutely transform the abilities** and provide substantial benefit to your business.

To ensure your organization gets started down the right path, you'll need to look at your data management practices. Among the key elements of strong data management is an understanding of:

- Where your data has come from, including within and outside your organization.
- Who has accessed and/or changed that data at various stages of its lifecycle.
- How your data can be used and in what context.
- When your data was collected and how changes are being managed.
- What your data has been used for in the past (and how it might be used in the future).

More importantly, it requires a deep and lasting cultural change. To create a truly data-centric culture, first, you must determine the value of, vision for, and relevance to strategy for data in your organization. Next, you should identify key personnel and executive sponsors to 'fly the data flag'. The end goal is to create a community of people who'll build, retain, and grow your organization's knowledge about data. Focus on building an open-minded, learning organization that's guided and aligned by agile and effective governance.

To truly become data-driven, businesses must also be committed to making effective use of their data, managing it through its lifecycle, and extracting value at every stage.

For more on being data centric – see the e-book:
Business agility: you're only as agile as your data.

Building a plan for success

The focus of digital transformation continues to shift dramatically, and today agile data management is in the spotlight. Intelligent businesses recognize it's not about extreme changes and restructuring the entire business through a one-off project. Instead, think of it as a behaviour: a way of working that's rooted in a culture where people are encouraged to think and act differently.

Dream big

As with many projects, you'll position yourself for greater success if you begin with a grand – albeit granularly articulated – vision of what you want your end-state to resemble. If you have not done so already, review your business strategy – paying particular attention to both the customer and employee experience. It's vital that you look at how data will be leveraged and managed.

Be sure to have a data management strategy and architectural blueprint – the target state you want for your organization. The next step is to define a roadmap to drive your transformational activities.

Start small

With a vision in place, it's important not to treat this as a large one-off project. Small and fast agile integrations win the ROI race. Win small victories quickly. Proceed with manageable, digestible steps.

Don't try and build out the entire architecture first; instead, focus on delivering measurable outcomes that will be of benefit to your customers and/or employees. This will ensure stakeholder buy-in across your organization and increase your chances for success over the short and long term.

We stress getting your data management right from the outset – and be sure you build in good governance. Building a machine learning model, dashboard or application on a foundation of uncurated data could lose you the trust of key stakeholders. This could be fatal to the success of your overall transformation program. However, ensuring good data integrity and access control will delight stakeholders and build confidence over time.

Think ahead and be agile

As you continuously evolve and adapt your plans over time, you'll refine them, so they get better and better. You may not end up where you thought you would when you started, but you'll likely end up in a great place. Changes may occur because new insights in the data drive new opportunities. There may be also be changes due to disruptive forces – for example: a new market competitor, a new business acquisition or a change in customer demand.

Based on our own experience, we recommend the following as you work through each data management initiative:

- Work with the business – engage with the stakeholders to ensure that you understand their context and needs so you can come up with valid recommendations and plans.
- Be realistic – ensure that any recommendations or plans have a good cultural, capability and investment fit.
- Keep what is good – ensure that the value in any existing capability is understood so that it can be used as a launching point.
- Communicate – maintain communication so you can proactively address needs as they evolve.

Conclusion

Conclusion

There has been much talk about digital transformation in recent years, however the word transformation infers abrupt change with an endpoint. While there'll be times when change needs to happen fast, successful organizations are in it for the long game and are constantly re-evaluating their business and their competition.

Many models and methodologies exist to support this. However, it's critical to ensure that any data maturity assessment is independent, technology agnostic and, preferably, uninfluenced by institutional politics. This ensures the result of the data maturity assessment is focused on delivering outcomes, rather just an output, and that it's aligned to the business' broader strategic direction.

From there, develop an organizational vision for an agile data evolution. This should be informed by industry trends, the latest practices and architectures (many of which are discussed in this e-book) and, of course, individual business needs.

> **Once the business understands where it is and its vision for the future,** stakeholders can plan and execute an agile data evolution – the first step on the road to becoming a truly agile business.

Matthew Drayson

Executive General Manager – Practices, Partners & Alliances

NTT Australia

Matthew.Drayson@global.ntt

linkedin.com/in/matthewdrayson

Amjad Bashir

Principal Consultant – Strategy and Architecture

NTT Australia

Amjad.Bashir@global.ntt

linkedin.com/in/amjad-bashir-3188098

NTT

Together we do great things