The evolution of the modern data platform

Summary

The history of the data warehouse dates back to the 60s/70s when the likes of Bill Inmon, and others, started discussing the concept of the data modelling practice. It looked to address problems such as data redundancy and duplication, data integrity and the associated costs. Historically, data was stored and duplicated across multiple decision points within an organisation, creating integrity issues on information shared. The data warehouse aimed to resolve these issues through consolidating and centralising data in a trusted holistic storage layer. As time moved on, different data warehouse modelling practices took shape, to address challenges around performance, time to delivery and scalability. From Ralph Kimball to Data Vault, different ways and means to store data became the data debate of the time. However, everything changed with the technology explosion which was the notion of big data. Technologies were now capable of ingesting disparate types of data incredibly fast and gaining answers from specific optimised queries on data quicker than holistic data modelling.

As time moves on, new data capabilities and architectural methods are developing as we speak. This point of view looks to walk through the data evolution, to explore where we are currently, with the modern data platform.

The data evolution 1 – relational database management systems

Data warehouses sought to address duplication of data, performance and reduce associated costs. Before data modelling was a concept to consider, organisations moved data from applications into siloed stores to gain access. This practice is still seen today, where business units that need to apply calculations on their data, pull it into individual stores, apply those calculations and store them locally for future use. Business domains like actuarial analysts, or finance accountants, still store data within Excel spreadsheets and Access databases, building macros to output calculated metric values which they distribute back to consumers of data. The speed at which those teams can produce results leads them down this path. However, these siloed data stores, with little to no governance, have historically caused a lot of challenges, from poor data quality, redundant data, duplication, little or no security, accessibility... the list goes on. By consolidating data within a central layer - the data warehouse - these challenges are addressed. Governance may be applied across the estate, regardless of which data modelling method is chosen, and trust is instilled.

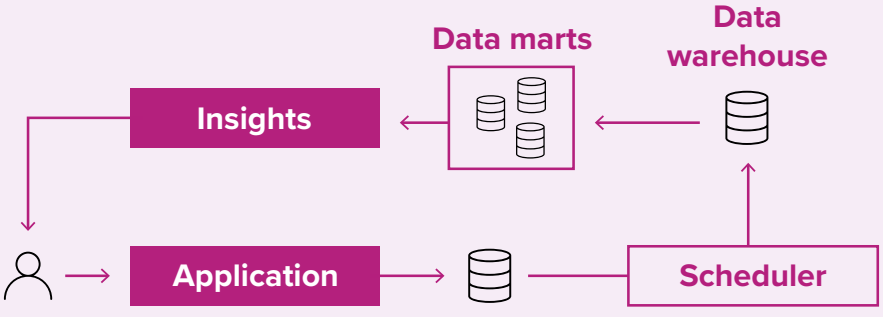
1

At a high level, data was scheduled out of the application database and moved into a predefined data model. That data model was designed with a relational database management system (RDBMS), and was schema-bound on writing data into it. This means, if an organisation decided to build its data warehouse in Kimball, the modelling structure would be with a "bottom-up" approach, and domain level dimensional data marts would be built, which would holistically form a central data warehouse through shared dimensions. If that organisation decided to build its data warehouse "top-down", with the Inmon approach, the full data warehouse model would be designed first, usually with a highly normalised model (3NF), and data marts would be created off that holistic model - much like a Data Vault model, where dimensional marts could be developed off a highly normalised link / satellite / hub model. All these methods were designed with the optimisation capabilities of RDBMS, which include various indexes, statistics, materialised views and partitioning. Technology vendors focused on creating tooling which helped the breadth of capabilities needed by organisations, becoming "the jack of all trades, the master of none, but oftentimes better than a master of one."

IT needed to build a data warehouse model which would be seen as being subject-orientated, time-variant, non-volatile and integrated (Inmon, 1995). Regardless of which modelling method one chose, there was an integrated holistic data model which needed to be developed, and it took time. Focus was on the full centralised platform and not on the specific questions the end consumers were asking. In the time it took to develop the full model, data silos were being built, and shadow IT formed within domains. Should the data warehouse project fail, not only were data silos being built whilst waiting, the failed project had large cost implications which defeated the entire purpose of its being in the first place.

"The jack of all trades, the master of none, but oftentimes better than a master of one."

2

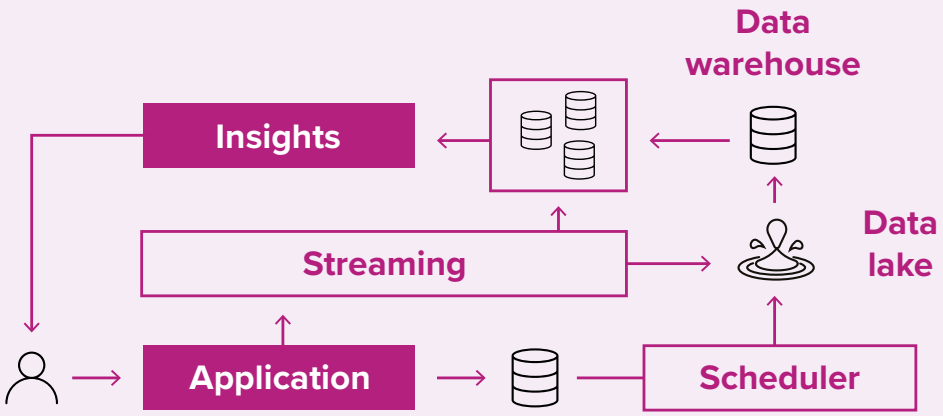


The data evolution 2 – the technology explosion

With the increase in data available to analyse, our capabilities in which to analyse that data needed to mature. Fast-moving data from social media and network feeds, web analytics and geospatial data, could not wait for a holistic data schema to be modelled before gaining insights. Data sizes have grown exponentially over time, and planning to create enterprise models which satisfy every data object is simply not feasible. This is where big data entered the room. The likes of unstructured storage facilities allowed data to be ingested quickly. We were able to ingest data at disparate speeds, in different formats, in massive sizes, without the worry that we needed to create that structure first - all our data in one place, at speed. End consumers wanted answers to questions fast, so the capability to ingest and answer those defined questions was built. The notion of 'data lakes' was born, and all things data landed in one location, without the need for holistic modelling. Predefined queries were written to query the data, to produce set answers, at blistering speed.

Initially, data lakes were seen to replace the data warehouse. No longer do the domains need to wait for data model schemas to be created as all the data is ingested without the need for one. Domains can merely define schema on reading the data. However, the optimisation techniques built by the RDBMS technology vendors, such as indexes and statistics, were now lost. So, speed in reading data was slower on questions which were not predefined. Ad-hoc analysis and joining multiple datasets became an issue as the big data platform was simply not defined to satisfy those types of requests. Organisations quickly learnt that a structured store with optimisation capabilities was still needed - a two-tiered architecture which included the best in both designs. However, instead of modelling all data, only data which may have to be used was holistically modelled in the persisted data warehouse tier. The more data which was modelled, the more data was added, and the further and more complex the schema became. In trying to bypass the need for the data warehouse, the platform had gone around in a circle and built data warehouses all over again. Domains ended up waiting for data to be extracted from the data lake and reprocessed into the data warehouse. Business domains ended up building their own silos yet again.

3



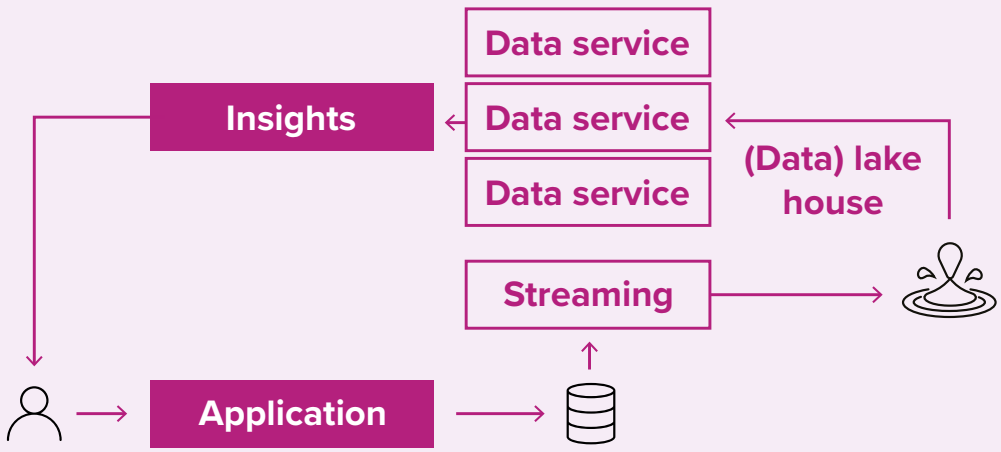
Enter the lakehouse platform

The lakehouse looks to close this gap and remove the traditional notion of the data warehouse by creating a metadata caching and optimisation layer (Zaharia et al., 2021) with the data lake which could be consumed by downstream domains. Issues where the RDBMS optimisation techniques, which would usually be used, are bypassed by modelling the data assets in a way to remove SQL-like functions. Modelling optimisation, or workarounds, include: removing grouping and ordering functions, removing joins (by creating large single tables), and duplicating data across data assets for the consuming domains. This reprocessing of data becomes a huge overhead when attributes are used across multiple datasets. If one attribute value changes, and that attribute is used across 100s or 1000s of datasets, each of those datasets will need to be reprocessed, thus increasing the risk of poor data integrity across those shared attributes. The net result is that domain consumers are taking their data services and creating their own data silos; yet again.

Up until this point, there is a growing theme across the evolution. The business domains are seen as consumers of the result of the IT transformation, rather than the key driver for change. Domains are the afterthought, as opposed to the starting point. How are domains using their data? What is their user journey? How are they collaborating with one another? How is data integrity and usage easily understood?

Ideally, we should start with the domain and build the solution from front to the back.

4



Analyst review: what do the analysts predict?

Changes in the data platform are moving as the need surfaces. With the rise of the Data Mesh, Zhamak Dehghani's architectural paradigm (Dehghani, 2022), aligns somewhat to what the analysts are predicting for the near future. McKinsey & Co. are seeing these changes in the market and highlight a number of capabilities that are coming over the next three years. They note that by 2025 interoperability between machine and humans is going to be far more seamless (McKinsey.com, 2022). Data is going to be embedded in nearly every aspect of their work. So, how does that affect the data platform evolution?

A few highlighted points they consider are:

- Flexibility on data stores is needed. The data platform needs to be modernised with the changing times
- Data needs to be treated like a product - a product being a collection of assets
- With the veracity of data moving in the market, maturity of data streaming must be considered
- Manual processes should be automated, including data management capabilities, such as: security, privacy and resiliency
- The CDO is no longer simply a governance keeper - needing to to realise value in the data within the organisation.

Four steps in order to become a modern data platform:

- Make use of road-tested blueprint architectures
- Create minimal viable products for deployment and scale
- Prepare the business domains for change through education, communication and collaboration
- Build agile data engineering teams by splitting platform teams (engineers, architects, modellers) and product teams (data scientists, analysts).

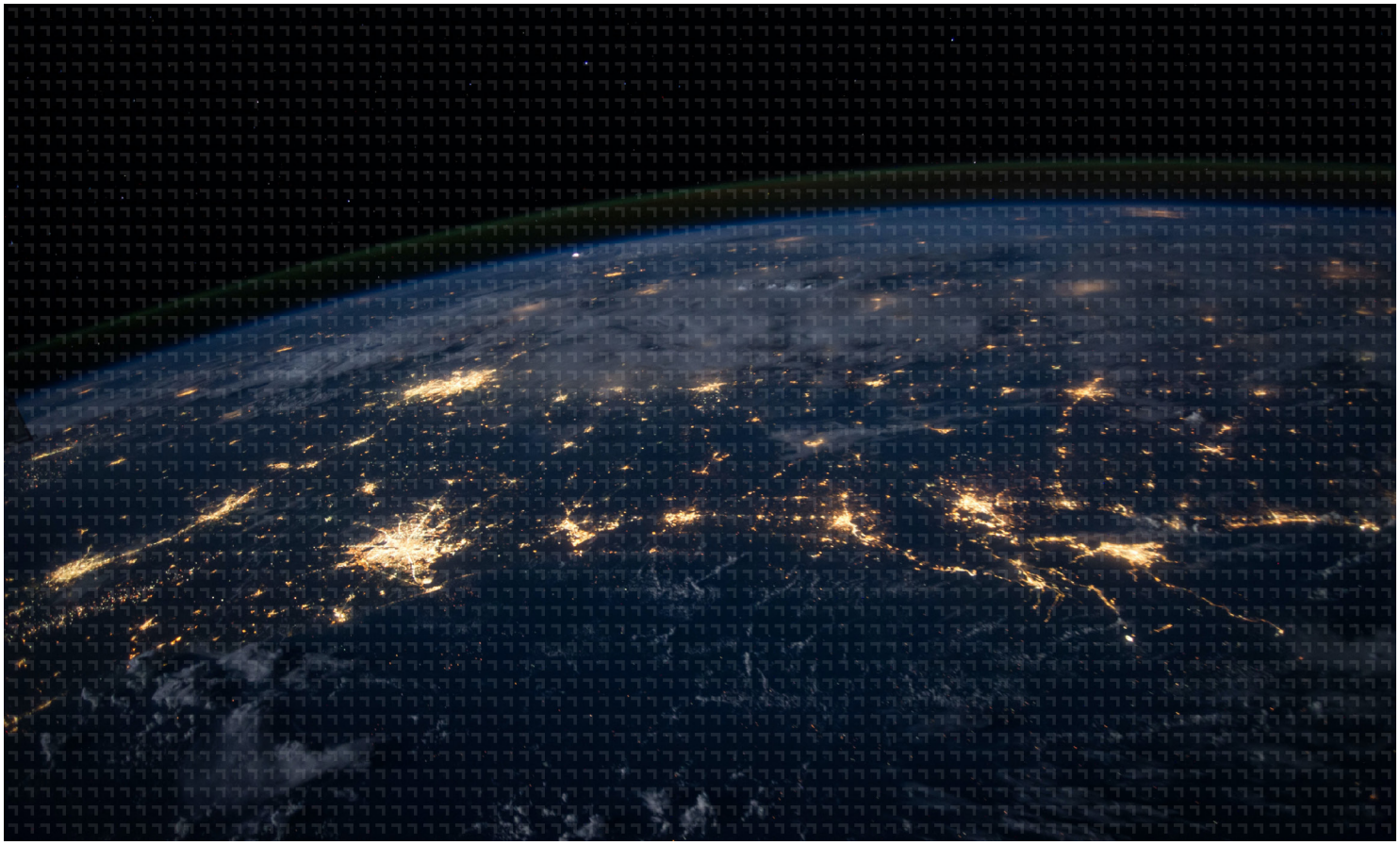
When reviewing the above, we can easily see the direction in which the analysts are seeing the evolution going.

The modern data platform: six key themes

The modern data platform considers the advances made up until this point and addresses the challenges identified. Each of the 4 stages of evolution have seen significant benefits in design, but in turn seem to either degrade past stages or build new challenges to those already solved.

The six key identified themes we should be considering when designing a blueprint modern data platform include:

- **Data as a product:** build and manage data product outputs which are a collection of data assets.
- **Data marketplace:** add a layer to trade data products. Adding this will clearly show integrity across data products, adding confidence to domains.
- **Domain level architecture:** focus more on building data solutions for each domain which will add value per domain. Different domains are more or less and as complicated as one another.
- **Front to back design:** leading on from domain level architecture, put the end consumer first. How are they going to use the data? Each domain will require data products which are going to be disparate to one another (with shared components), and each will have very different requirements in ad-hoc analysis versus optimised predefined questions. Work backwards from consumer requirements to domain architecture design.
- **Metadata management:** management of metadata produced through all these products is critical to track not only redundant data, but to govern the landscape.
- **Enhanced stream driven:** remove constant dependency on end-of-day processing when streams are available. Enhance the platform by supplying up-to-date data so that information value may be gained sooner.



How does this affect the existing design?

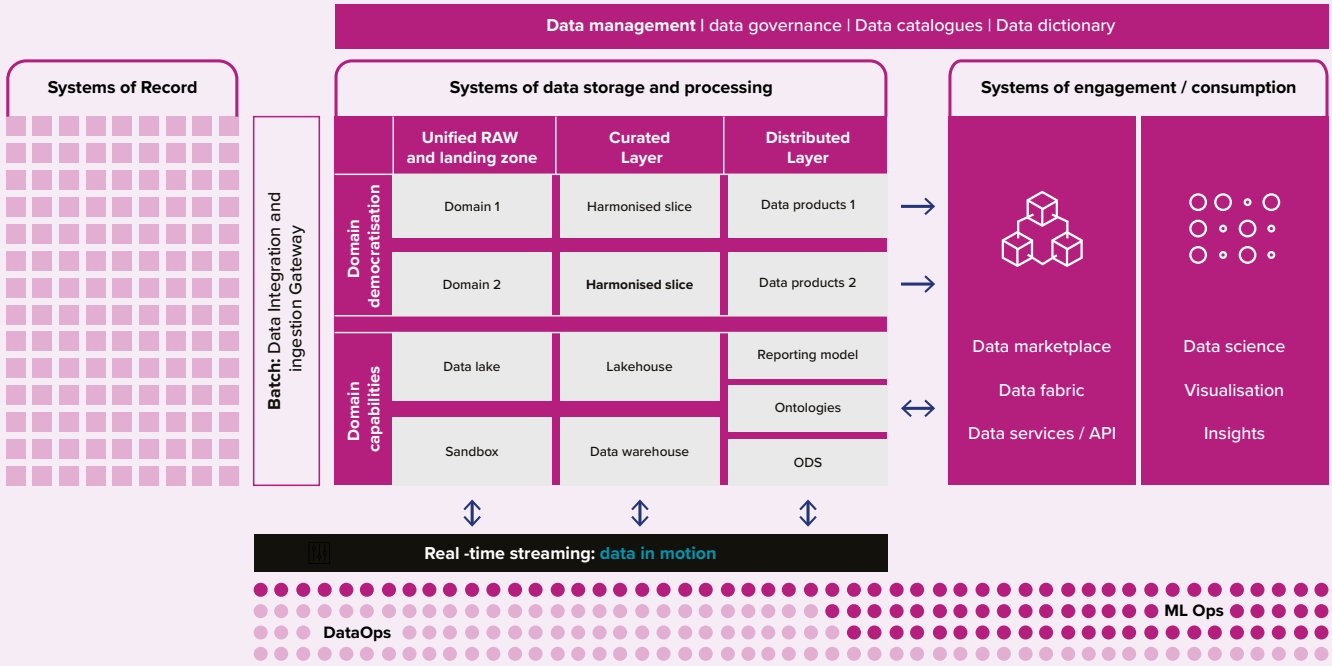
The modern data platform should consider the differences in domains. A mature platform, much like the idea of big data, can ingest multiple sources and distribute multiple products. The change in mindset is that business domains are different to one another. Trying to holistically treat all data the same leads to unnecessary complexity for a handful of shared components or dimensions across those domains. Focus on creating an architecture which democratises the estate, whilst supplying the capability to share information through an engagement layer.

The modern data platform blueprint architecture

The change is not to replace the lakehouse, or the data warehouse, or other legacy data capabilities. The biggest change is purely the realisation that the modern data platform needs to put the business domain first, and work from front to back, rather than designing technical solutions considering IT first. If that means democratising a massive centralised repository, which is causing domains to create shadow IT functions, then that is the design which should be designed.

Data lakes, data warehouses and lakehouses may all exist, but are split by domain, and managed as separate entities, as opposed to one holistic solution. Should there be a need for fast ingestion and schema on read operations, data lakes are still added, by domain. Conversely, should there be a fundamental need for ad-hoc analysis and slicing and dicing of data, structured reporting stores may still exist, by domain. Agnostic ontologies may exist across data products, which are produced by domains, to manage data catalogues and understanding of the data estate, and ease classification of critical data elements, whilst operational data stores, which seek to surface transactional data, may have a place within the estate too.

The fundamental point is to create these capabilities, per domain, and for them to be optimised as the domain level requirement becomes apparent, thereby creating a faster time-to-market, domain by domain.



Breaking down the blueprint

Considering the blueprint architecture, the modern data platform is split into three fundamental layers; systems of record (SR), systems of storage and processing (SSP), and systems of engagement (SE). To move data into the platform, there is a data integration gateway which merely checks schemas, if necessary. In the case of group consolidation of data, no schema is checked. But in situations, for example, in finance, where structure is critical, we can validate the structure of the data coming in.

The domain level data in any system of storage and processing (SSP) is split across three layers:

1. **The raw and landing data layer:** consolidates data, minor checks on data types and ranges. Minor preparation for curation.
2. **The curated data layer:** the bulk of the transformation happens here. Cleans data and builds trusted data for downstream slicing of data. No data products are created, but data is organised in a way which is easily harmonised for downstream consumption.
3. **The distributed managed data layer:** data is sliced into data products and ready for a multitude of consuming mechanisms in the SE.

Processing of the data between layers should be a combination of streaming and batch processing, depending on the domain requirement. So, instances such as producing a corporate balance sheet, will need end of day processing in batch, whilst instrument movements of market data will need to be streamed in throughout the day.

Data management and governance is conducted at the domain level, but group governance policies are applied where regulatory compliance is fundamental.

The SE layer creates a means to share, collaborate and generally consume data through trusted interfaces. Options such as a data fabric would enable the autonomous analysis of the metadata to allow auto-cataloguing of attributes, thereby streamlining data management capabilities. Further to this, the fabric would allow for a centralised layer to access data products. Built on top of that capability, the data marketplace can enhance this view in order to trade, rate, share and even recommend data products to additional consumers, based on their relationships to one another.

These data products, which could be one single asset if required, can then be fed into consumers, which would include machine learning models as model input. The output of that model could be fed straight back into the platform as “output-as-input” to enrich information value. With the SE layer, visualisation and business intelligence tooling would become agnostic, and domains will be able to use whichever tool suits them best. Viewing of the management reports would be achieved through the central marketplace, removing the various paths to disparate reporting technologies in the backend.

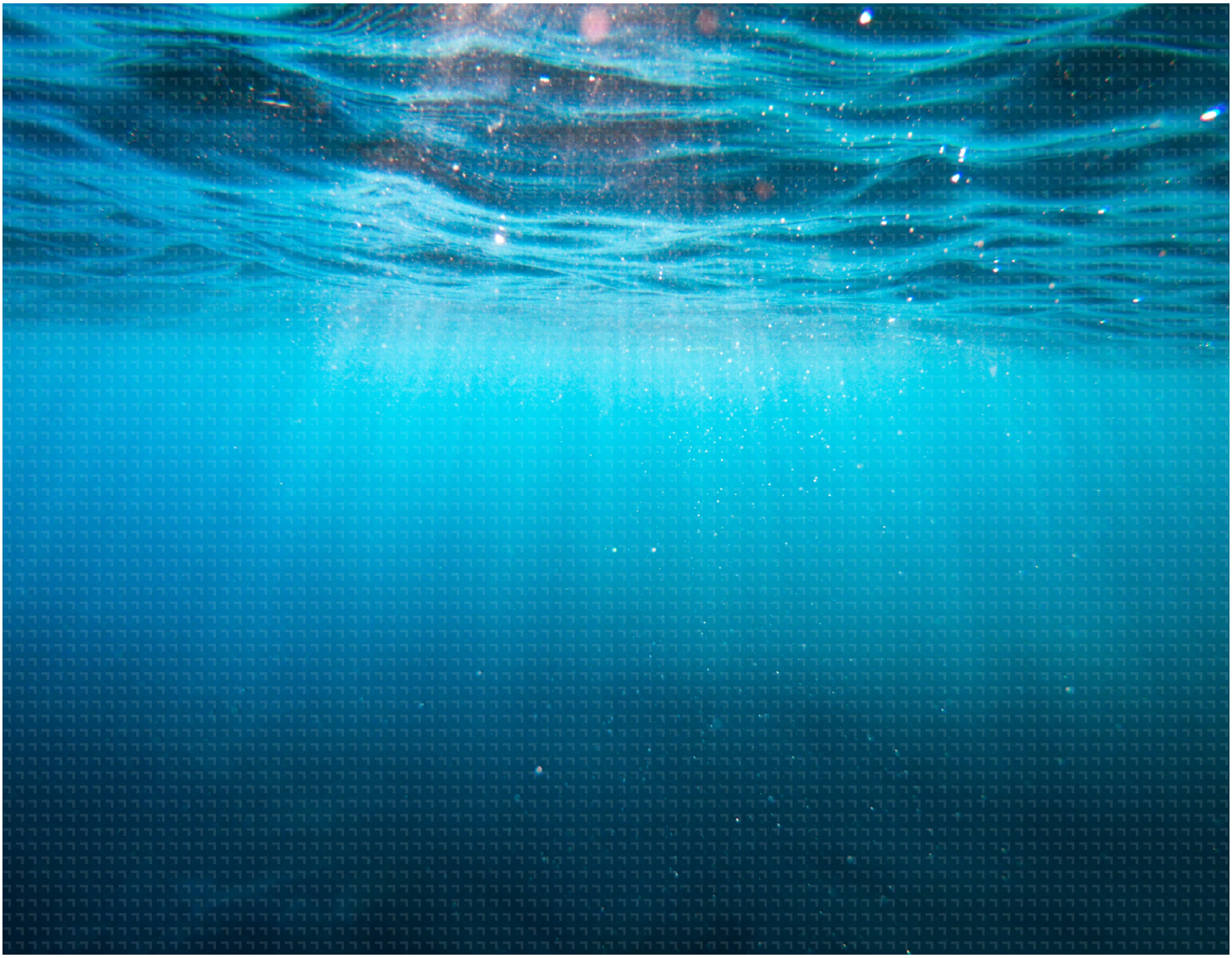
Focusing on autonomous operational management, data operations (DataOps) and machine learning operations (MLOps), should be included across the platform to automate metric validation, lineage and data transformation processing and leakage. In addition, this enables the retraining and redeployment of ML models, auto verification of output data and more.

How to get there

There is no ‘golden hammer’ or end state for the modern data platform. There are capabilities which should be considered based on each domain’s requirement. Getting there merely requires looking at the first step as a minimal viable product (MVP), and not the end state, which could be seen as a very daunting task.

Consider a systematic approach to achieve the modern data platform, including; assess, transform, manage and monetise **(ATMM)** and define one minimal viable product (MVP).

- **Assess** the domain first and understand how it relates to the company-wide strategy. Focus on the MVP. Don’t boil the ocean.
- **Transform** that MVP, from technical design to people and change management.
- **Manage** and govern the MVP that is in place.
- **Monetise** and generate additional revenue / value streams from that MVP.



Summarising the platform

The modern data platform does not seek to define a golden hammer to replace all the different stages of legacy architectural designs. It ensures that we now put focus on the domain itself. We focus on adding business value and lead with the business strategy first, not the technology strategy. We use the multitude of mature data capabilities that exist in the market and blend them to build solutions for each domain. By taking this approach, we are able to deliver the modern data platform that focuses on the domain and delivers the highest amount of business value and return on investment.

Our specialists

David Tuppen
Head of Data, Analytics and AI

David Tuppen is Head of Data, Analytics and AI at GFT. He has been working in the data space for 20 years, from technical development through to business development. His area of expertise is enterprise data strategy, with a focus on hybrid data architecture patterns and applications. His career experience includes working at the Bank for International Settlements in Basel, Athene Holding in Bermuda, and Milliman Actuarial Consultancy in London.

References

Dehghani, Z (2022). Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly

Inmon, B. (1995). What is a Data Warehouse. Prism

McKinsey.com (2022, January 28) The data-driven enterprise of 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025

Zaharia, M., Ghodsi, A., Xin, R., Armbrust, M. (2021) Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In 11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021

About GFT

GFT is driving the digital transformation of the world's leading companies. With strong consulting and development skills across all aspects of pioneering technologies, GFT's clients gain faster access to new IT applications and business models. Founded in 1987, GFT employs over 10,000 people in more than 15 countries.

blog.gft.com
twitter.com/gft_en
linkedin.com/company/gft-group
facebook.com/GFTGroup
> gft.com

This report is supplied in good faith, based on information made available to GFT at the date of submission. It contains confidential information that must not be disclosed to third parties. Please note that GFT does not warrant the correctness or completion of the information contained. The client is solely responsible for the usage of the information and any decisions based upon it.