Digital acceleration: How to build a robust enterprise data framework

Received (in revised form): 30th January, 2023

Pierre Kovacs
Senior Vice President, Endava, UK

Pierre Kovacs, Senior Vice President, Banking & Capital Markets, Endava, is based in London and has 25 years of experience in the industry. On a day to day basis he helps individuals to successfully build better solutions. Prior to joining Endava, from 2014 to 2021, Pierre was Chief Technology Innovation Officer and COO UK for a FinTech: Orchestrade Financial Systems. Pierre held several technology leadership positions in investment banks and asset management, such as Head of Risk, Finance & Operations Technology for Daiwa Capital Markets, VPFX Options at Credit Suisse, Senior Architect for Liability Driven Investment at Barclays Global Investors and BlackRock, and Senior Architect at Société Generale. Pierre holds a Diplome d'Ingenieur in computing science and finance from Paris's Engineering Graduate School, CY Tech.

ABSTRACT

This paper addresses data as a corporate asset that needs to be protected, but also shared and made accessible. The paper recognises data’s role as the bedrock of the analytics that drive business decisions and that should be gathered routinely during normal, transactional operations. The paper acknowledges that the volume and sources of data and its users are diverse, global and continue to multiply. This, in turn, identifies the need for flexibility to design and develop a scalable model for data management that will support future growth and risk management capabilities. This paper outlines how to build a data framework, recognises the scale and complexity of the problem and recommends courses of action. These recommendations include data scoping, awareness of context, the wider macro trends and pressures, goal definition and the need to be outcome led. The paper further identifies that although most firms are already at some stage of their digital journey, there is no real ‘end goal’; it is a constant process. To successfully build a robust enterprise data framework, it therefore becomes crucial to work diligently and understand the steps required. The paper discusses approaches, pitfalls and the constantly evolving solution landscape. The paper concludes that the journey to a robust data management framework is complex, but, through embracing digital acceleration, it is possible to gain the competitive advantage to which the framework aspires.

Keywords: data strategy, enterprise data framework, data programme, data management, risk management, digital acceleration

WHY IS DATA AND DATA MANAGEMENT VALUABLE?

Data management is often at the core of innovation. Consider the advances in vehicle satellite navigation systems. This involves a personalised human–tech collaboration that helps drivers navigate a busy road and safely reach their destination while factoring in current conditions. Real-time traffic information compares current data with historic patterns on a live feed that is distributed globally. Warning systems alert drivers to hazards such as traffic jams or breakdowns. Artificial intelligence (AI) is used to understand the driver’s context,
relationship to other cars, pedestrians, traffic lights, road conditions and more.

In just the same way, asset managers need data to achieve better outcomes for their investments. This means generating new trade ideas as market conditions change — or, even better — before they change; rebalancing portfolios at a minimum cost; protecting performance and managing risks against unexpected new stress scenarios; monitoring compliance with regulatory requirements or analysing the environmental social and corporate governance (ESG) performance of a portfolio.

After all, data may be prolific but, if it is not managed, then how can it be of value? Never neutral, data management is equally about opportunity and risk. The crucial information that supports the business functions and decision making is a corporate asset. It is not just a by-product of transactional operations but a first-class citizen for analytical purposes.

For any company, data itself is not the problem. The problem is being able to build and maintain a robust data management strategy that will provide a competitive edge and support future growth.

**WHY IS IT A CHALLENGE?**

The concept of data management is complex, which is one reason why it can provide a competitive edge if properly addressed. Equally, if left unmanaged, data is an asset that can quickly become a liability. An example of this is where portfolio managers use spreadsheets that potentially expose confidential data as a solution to the lack of a governed data platform.

Data is vast, diverse and global, compounded by the complexity of the data sources. Volumes will only continue to grow, and users will increasingly ask more complex queries of the data when looking for patterns and correlations. User profiles themselves are varied, dependent on roles and technology experience: portfolio managers, traders, risk managers, quant developers, operations teams, for example.

Data’s diversity incorporates qualitative and quantitative data that is nominal, ordinal, discrete and continuous. Such is the nature of business that different data is needed for different assets. Cross-asset market data, trade data, static data, risk and performance data and unstructured data, each comes with specific requirements. Unstructured data, which is not arranged according to a preset data model or schema and cannot be stored in a traditional relational database, is common in the private market. Many business documents are unstructured and provide important insight, including financial statements and accounts held in non-standardised formats (such as PDFs).

Then there is the time sensitivity of data, where data loses its predictive value over time. Making decisions based on stale data is no better than guessing. For example, monitoring risks calculated with T-1 (yesterday’s) market data or T-1 positions is not acceptable in fast moving markets.

Data quality is a notorious problem. ‘Data is dirty’ by nature. If data quality is poor, time is wasted trying to reconcile, understand or critically assess the quality of the data as opposed to using it. At worse, when the confidence level is low, the data is simply ignored.

Most firms have global offices with different regional policies, regulations and cut-off times. Multiple systems hold data, including legacy systems from acquired businesses and third parties. The temptation is often to group everything in a single repository. However, it might not even be legal to combine data into one location, when regulations such as the General Data Protection Regulation (GDPR) are factored in.

All these factors need to be considered in the data programme decisions.
Insight/Recommendation 1: Decision makers should carry out a data audit to understand their data footprint. A common pitfall is to forget about ‘user developed applications’ that have become crucial for the business and are unknown to, or not managed by, the information technology (IT) team. This use case analysis helps to identify governance and risk associated with workaround practices, such as the use of spreadsheets that potentially expose confidential data. Although they may not be immediately apparent, workaround practices cannot be ignored. Not least because they have become established ways of working and it will be necessary to provide an alternative so that the business is not disrupted. Furthermore, unless recognised, they can become blocks in the critical path to migration and can define the length of the project. This tends to be forgotten when firms start a data management programme. It is ultimately change management, with the need for communication and bringing people along on the journey.

HOW TO BUILD THE RIGHT DATA PROGRAMME

State of play

Investment industry challenges are being driven by increasing cost pressure, an evolving regulatory complexity and pressure for more transparency and reporting. Unpacking these challenges against current deteriorating economic conditions adds an increased sense of uncertainty and urgency.

Most firms have some form of data solution in place, such as a data warehouse. But, in most cases, these solutions do not address all the challenges and risks, especially around data accessibility, security, quality and governance.

How can firms build the right data programme that responds to challenges while delivering the flexibility they need to scale? How can firms make the most of the latest innovation while also striving towards the best practices that are evolving in parallel?

Regulatory pressures

Growth in packaged ESG investment products is leading demand for high-level ESG data and underlying granular data to better understand risk. This demand is compounded by new disclosure requirements, such as the Sustainable Finance Disclosure Regulation (SFDR). Although these new regulations affect Europe, similar legislation is likely to be introduced in the US soon. The latest SFDR deadline of 1st January, 2023 required that firms disclose information at both a product and entity level, meaning that firms must understand the classification of their investments.

In Q3 of 2022, World Business Research (WBR) Insights surveyed 100 Heads of Investment Operations and similar at buy-side firms across Europe to find out about the challenges they face in developing a data solution for ESG integration. The results of the survey revealed that only 27 per cent of the cohort had a solution in place, whereas 44 per cent had mapped data across securities but required assistance with implementing reporting. Twenty-two per cent were still determining requirements or had yet to determine impact and 7 per cent planned to deal with the data challenge manually.

This is an example of the impact of major regulatory change on a firm’s data capabilities and demonstrates that data management remains a challenge. Firms will need to collect aggregate data and will be expected to already have good foundations for their data in place to effectively report on it. Once understood, firms will need to implement policies and procedures to procure the required information from deal sponsors, underlying companies, and third-party providers. As this is new legislation, there are no established best practices. This means firms will need to scale their processes using...
technology and continuously re-evaluate and optimise data, systems and procedures.

**Transparency**

Robust data management brings some transparency. At a foundation level, this means that data is discoverable, accessible, auditable, meets quality expectations and is processed according to policies.

The common denominator for achieving data transparency is metadata. This provides better intelligence about data so firms can, in turn, be intelligent and automate operations. Data cataloguing for example, where metadata helps identify data’s location, its nature (description, relationships, semantic tree) its source, how to access it and how to process it with the applicable usage policies. Descriptions without context are ambiguous, as is naming. Terms like ‘entity’ or ‘user’, for example, can have many meanings. Consumers should be able to understand data lineage, data’s origins, with changes tracked and understood. This in turn feeds into auditability, assessing quality throughout the data life cycle to ensure accuracy and efficacy.

For example, a risk officer should be able to drill down into a report to zoom in on transactions or market data and have clarity on the data sources. This might include accounting book of record (ABOR), investment book of record (IBOR), and performance book of record (PBOR), fed by the third-party administrator and the name of the data finance provider.

Data policy applications, the rules and principles that provide the framework for governance, quality and architecture, can also be automated, to some extent, using metadata. This can help, for instance, in ensuring client data is properly handled as private information, or that market data is handled according to the licensing restrictions of the particular financial data provider.

Transparency also means that error management is less time consuming, and errors can be more easily investigated and data recovered. This is another aspect that tends to be underestimated: errors are inevitable. External dependencies and human operations are the main cause of mistakes. Robust data management does not claim to eliminate errors, but to detect them and remediate them in a timely manner.

The understanding of data quality itself is based on accuracy, consistency, completeness, if it is up-to-date. Sometimes, ungoverned data must be mixed with governed data to meet real-time requirements. For example, a position blotter can show a mix of validated/reconciled trades from the IBOR and freshly executed trades from the order management system (OMS) or market data from the master data repository mixed with a real-time feed.

Finally, data, like any product, has a life cycle. This is another key element of robustness. This life cycle must be managed, and the platform must be designed to evolve the shape of the data without breaking the existing processes or making data in a previous format inaccessible.

**Insight/Recommendation 2:** A good metadata architecture will truly unlock the full value of data. For these reasons, and to take data management beyond the reporting requirements, a data programme should include a ‘metadata management programme’.

**Self-service**

Users expect reports to be interactive and they expect to be able to generate new reports immediately. The data users are diverse and can include data scientists with strong technical expertise, as well as someone who has less technical experience. The demand for self-service data is a recognised trend, but accessing data from a variety of different sources remains complex.

Being able to self-serve data with little technical knowledge is transformational. It enables more direct mapping between what the user is trying to achieve, the goal or idea
and the kind of query asked directly of the data. There is a plethora of interactive data visualisation tools designed to help turn data into visualisations, such as Microsoft’s Power BI, Tableau and Qlik Sense. Data dashboards allow the interactive exploration of data by means of drop-down lists and tickboxes with no coding required.

To further facilitate self-service, it is necessary to understand the diversity of use cases, how to serve the data, what quality is expected and assign access permissions. Regulations, for instance, might forbid access, or a particular trading desk should not see certain information and may require a Chinese wall. User consultation and involvement is essential. If the focus is purely on gathering data and putting it in a central place, there is a risk that consumers will not know how to use or access it. Understanding of use cases might, for instance, identify the need to provide event notifications when new data is available, such as a new trade in the system, a market data tick or a change in definition of financial instruments.

In a burgeoning regulatory landscape, firms need to retain flexibility while continuously questioning practices. Technology and innovation can provide this flexibility while assisting in establishing best practice. This is the initial step before addressing data science functions and the advanced analytics that provide more value and competitive edge.

**Insight/Recommendation 3:** Decision makers must be aware that use cases must be detailed enough to facilitate their accommodation into the programme. Being use case driven and flexibly accommodating new use cases quickly helps to ensure that value is delivered. Conversely, aiming for a universal solution that is too rigid to evolve, will take longer to deliver value.

**Skills and organisational considerations**

When any firm starts to examine its data management approach, it often becomes apparent that a multi-disciplinary team is required. There are the expected roles of data engineer, data analysts or data architect. In addition, for scalability or cost reduction, the platform may need to be cloud-based. This means devising a cloud strategy and obtaining the necessary cloud skills, so potentially a cloud architect. The platform will also need to be tested, making testing experience a necessity. Processes might also need to be reviewed by a subject matter expert. Large programmes can necessitate an experienced enterprise architect and programme manager. As this might require an extensive mix of experience that might not be fully covered by the in-house IT team, it is normal to liberate these skills from a service provider.

**Who owns the data management programme?**

The overall management perspective is similarly complex. The introduction of GDPR caused real trouble for many organisations at the time when few companies had appointed a chief data officer (CDO). The responsibility for GDPR would usually fall to the compliance officer, as it was regulatory. Although seemingly logical, this role did not understand the requirements from a technical aspect. Management roles now include the chief data & analytics officer (CDAO), who will work with the chief information officer (CIO) and a chief information security officer (CISO). They all have overlapping and conflicting goals. The CDAO sees data as prime and wants it to be accessible, the CIO thinks application first and sees data as a by-product and the security officer wants to lock data down. These conflicting priorities make frictions more likely.

**Insight/Recommendation 4:** Decision makers should define and communicate a clear outcome-driven roadmap that sets out model choices, key roles, responsibilities and skillsets. Do not underestimate the multi-disciplinary aspect of this undertaking.
HOW TO SELECT A TOOL STACK

Model selection

The responsibility of data management will depend on the model adopted by the firm. This might be vendor SaaS (software as a service), a centralised data platform or a distributed one. Each of these strategies have strengths and weaknesses. For example, going with a vendor might present issues around flexibility of data models and integration with other tools. However, a centralised data platform can create a bottleneck, and goes against the current microservice trend to organise by domain. At the same time, a data mesh, or data fabric, where responsibilities and coordination are defined across separate domain teams and their data products, is relatively less common and there are fewer reference implementations of this model, with potentially limited support options. Firms taking this approach may need to be prepared to lead the way.

A popular strategy for addressing data management is to start by establishing a near term solution as a foundation for more advanced analytics. More recently, an approach has been to focus on a ‘domain’ and to apply data as a product for this specific domain. It is different from the earlier broad and shallow approaches to managing data in one central place. This domain driven design (DDD) approach, instead involves data partition and gathering of in-depth data. Data is structured as data domains, has data owners and is properly catalogued so that interested consumers can access it. The team generating the data is the data owner and must make it accessible for data consumers. The generating teams need to satisfy several principles when building and managing their data products, such as data integrity, discoverability, self-description and interoperability. This increases consumer confidence in the data products, which, in turn, unlocks value. Teams across the organisation do not need to waste time trying to reconcile or understand what the data is; they can focus on using it.

If a firm moves from a central data repository to a distributed model, the advantage is that the data-producing departments will naturally know their data best. This makes it easier for them to derive benefits and develop new use cases. With this new distributed architecture, the data scientists and engineers become part of the data-producing team rather than go-betweens for data producers and consumers. They learn the domain knowledge to support team colleagues when preparing the data products. As the entire process is simplified and accelerated, it should lead to lower overall costs.

Whatever model is decided on will define responsibilities and coordination. However, the right technology is needed to enable it.

The tool stack

Firms at every phase of growth require a cohesive, integrated technology tool stack to assist and streamline operations. As strategic as the organisation and process it supports, the tool stack (which includes platforms, applications and outsourced services) should contribute data to the database for future decision making. The marketplace is inundated with technology and data management tools — everything from data governance to data quality to lineage tracking. However, tools are only as good as the practitioners using them and their understanding of which tools should be applied to which business case. As Grady Booch famously said, ‘A fool with a tool is still a fool’.

The tool stack is evolving ever more rapidly, across products and versions, and are often compatible with all the major cloud providers. The cloud providers offer out-of-the-box integrated end-to-end solutions including data ingestion, analytics, serving, monitoring etc.
With a data mesh approach, there is less focus on a universal tool stack, since the producer and the consumer can choose the stack that works best for them. The consumer would typically leverage the elasticity of the cloud to create a stack of choice to process the data and tear it down once done. Conversely, the producer can choose the tool stack, as long as it delivers the data in a suitable manner for the consumer’s use case.

In this way it is possible to leverage whatever new technology works best. This is highly relevant, considering there are still many innovations to be made. However, what is important is the data.

Before selecting tools, decision makers need to ask which tools are best placed to provide greater transparency of datasets and enable robust risk management. For example, asset managers might build their data management around the data warehouse (or even cloud data platform) of their portfolio management service (PMS) because it is perceived as the centre of gravity of their data. They may only realise later that they could not get the required flexibility for their data model, and the feasibility of the integration with the governance and monitoring tools was suspicious. These anti-patterns\(^6\) are a common response to a recurring problem that is usually ineffective.

Errors will happen as well as error handling. Their knock-on effects in the chain should not be underestimated. The impact could stop someone trading or might mean the right trading limits are not in place. This is especially true for exception handling, which can be overlooked but has proved to be notoriously time consuming. Errors are frequent even under ‘normal’ circumstances. For example, if someone has not received an ABOR update on time from their third-party administrator (TPA), how do they surface this to the business users with the implications for the rest of the process?

AI is starting to provide the heavy lifting in the data management workflow: for example, to automate data quality checks or in building observability.\(^7\)\(^8\)

Insight/Recommendation 5: A tool stack should be chosen according to the size and maturity of the organisation. A selection process starts by establishing a number of architectural principles, criteria and constraints. This could be a requirement for cloud agnostic architecture, using existing tools and providers that could mitigate the cost of migration or determining which overlapping tools should disappear. There is usually a trade-off between the effort to combine tools into the platform and a best-in-class approach. To determine requirements and specifications, it is important to engage the business in the decision making and be relatively outcome-driven. What is the problem the tools are being asked to solve? Being outcome-driven is required for demonstrating value and return on investment (ROI). If the end goal or benefit is unknown, how is it possible to measure success? Decision makers must look at all facets of the problem. This ensures the platform is geared towards use cases so that it is used while remaining flexible for future requirements, such as SFDR. AI automation can accelerate the creation, management, and use of digital assets. AI tools are particularly helpful when dealing with volume, where it would be close to impossible to do it manually, and for assuring quality management.

Connectivity

Traditionally third-party portfolio management systems, even financial data providers systems, were closed systems. Data was accessed via a user interface, a report extracted, and the system exited. This was replaced by the data warehouse, which usefully separated transactional and analytical data. This was a better way than the closed system, but each time it was slightly different for each vendor and requiring expertise in data extraction.

There has since been a breakthrough in how vendors share their data with the data
warehouse in the cloud. One of the most adopted of these platforms is Snowflake, which includes Blackrock, SimCorp, and global custodians such as BNY Mellon and State Street among its users. Snowflake’s ‘secure data share’ allows users to connect and communicate with vendors’ platforms easily and access data outside the boundaries of the firm. Snowflake’s Data Marketplace offers more efficiency and is a good example of cataloguing, which is exactly what firms should do internally. The lesson here is that connectivity is a very powerful enabler.

**Insight/Recommendation 6**: Decision makers must balance the strengths and weaknesses of models against business strategy and intended outcomes. In particular, the bottlenecks of a centralised model versus the complexity of federating a data mesh.

**HOW TO MEASURE THE IMPACT OF A DATA PROGRAMME**

How can the expected benefits from a data programme be quantified? How can its actual impact be measured, its costs be justified?

For enablers such as data management, it can be difficult to produce an ROI. Returning to the earlier analogy of navigation/traffic systems, it is difficult to say how many accidents, commute time, fuel or emissions good traffic management is saving. Ideally, the data management platform is developed around key use cases to which associated business value can be applied.

An easier gauge is to see if people are using the data. If they are accessing data often, then it is evidently of value. If intended end-users are not using the data, it is important to establish why. It could just be that they do not know it is available, or the infrastructure to distribute the data is not what was expected. There could be a technical barrier, meaning training is needed. Conversely, the user might be more confident about the technology, expecting more sophisticated delivery. Not accessing the data could be just a technicality because users do not have the right pipeline.

All of these reasons for non-usage loop back to the need for a proper data management programme in the first place. As is often said in DevOps, the value delivered is the key measure of success. In other words, if it is not in production or if it is not used, it does not deliver any value, however large the dataset. The data platform should be able to monitor its use.

On a similar theme, management information systems (MIS) have become more affordable as systems develop more observability. Asset managers can look introspectively at their operations and can reduce operational errors by detecting the commonalities between errors. For example, if booking errors happen at the same step in the process, or with the same financial instrument, there could be something wrong with this product or the user interface. Or perhaps an individual trader just needs training. It is operationally oriented and relatively easy to see the value. If it can be measured, it can be managed.

However, looking at measures of activity or output such as data quality and improved decision making does not measure impact. A 2021 Gartner CDO Survey identified CDO objectives as including data quality, improved decision-making capability, increasing internal data sharing and process efficiency. These are all technical outcomes and measures of activity and output. They are not the resulting business outcome. Incidentally, only half of those surveyed even achieved these intermediate outcomes. The survey showed that CDOs and other such leaders are growing in influence and impact. Their success is characterised by an ability to provide measurable value to multiple business stakeholders, whereas unsuccessful CDOs focus on technology rather than people.

The true business outcomes might include improved customer experience, ROI delivery, cost optimisation, employee experience or data monetisation. Although the Gartner
Survey\(^1\) showed that one in ten data teams are taking data assets and turning them into realisable added value product or services, there is still a huge amount of work to be done on measuring effect and following through to make sure impacts occur. Once achieved, this will enable data consumers to recognise the true value of the data. This also aids the prerequisite mindset qualities of being curious and critical thinking that help ensure ongoing success.

**Insight/Recommendation 6:** Decision makers should ascertain the business value associated with use cases and use business outcomes as a means of monitoring and measuring impact that can be shared with stakeholders.

**CONCLUSION**

This paper has discussed the scope of building a data framework and identified considerations to aid in its construction. This has included the techniques for making data accessible, the awareness of context, trends and regulatory pressures and the need to consult and involve all data creators and users. Additionally, by being outcome driven, it is possible to define business value from the outset so that it can be effectively monitored and measured. With effective data management there is an opportunity to access alternative data sources that can feed into alternative models. This helps to provide the firm with competitive advantage.

There is no real end goal when building a robust data framework, as it is continuous. This reflects the changing technological landscape that has not yet reached maturity. Consequently, the data programme should be adjusted and improved as new models, technology and tools become available.

Decision makers must shift their thinking from purely digitising a process to instead adopting and leading the latest technological innovations. This innovation is crucial to an organisation’s success. Those with the vision and capability to embrace it will gain competitive advantage. As McKinsey and Company stated,\(^2\) technology investments are not simply one-time costs. Instead, they are becoming a continuous and meaningful portion of the cost of operating a successful asset manager.

There is ‘no one size fits all’ for a data framework and there is also a certain amount of risk involved in innovating. There is, unfortunately, no secret model to success. In John Kay and Mervyn King’s book,\(^3\) it was purported to be a mistake of thinking to apply physical models to problems faced in business and finance. It is important to ask the deeper questions and ensure you have options for dealing with uncertainty. The most useful models are those that do not attempt to quantify, but instead provide parables, insights into how the world works.

**REFERENCES AND NOTES**

3. Domain-driven design (DDD) is a major software design approach, focusing on modelling software to match a domain according to input from that domain’s experts. Under domain-driven design, the structure and language of software code (class names, class methods, class variables) should match the business domain.
4. The name Domain-Driven Design comes from the book ‘Domain-Driven Design: Tackling Complexity in the Heart of Software’, Eric Evans, 2003, that describes the approach through a catalogue of...
patterns. A community of practitioners has further developed the ideas. The approach is particularly suited to complex domains, where often disorderly and complex logic needs to be organised.

(5) Grady Booch is recognised internationally for his innovative work on software architecture, engineering and modelling and, with colleagues, developed Unified Modelling Language (UML). He is attributed to have said ‘A fool with a tool is still a fool’.

(6) Anti-patterns in software engineering are design patterns that may be commonly used but are ineffective and/or counterproductive in practice. Despite initially appearing to be an appropriate and effective response to a problem, anti-patterns typically have more bad consequences than beneficial results.


(8) An example of a solution that uses automated monitoring, automated root cause analysis, data lineage and data health insights to detect, resolve, and prevent data anomalies, Monte Carlo, available at https://www.montecarlodata.com (home page).

(9) Snowflake Marketplace enables users to discover, evaluate and purchase data, data services and applications from leading data and solution providers. Available at https://www.snowflake.com/en/data-cloud/marketplace/ (home page).


(11) Snowflake Marketplace, ref 9 above.
