The next generation data platform

Martin Lidl
Director
Consulting
Deloitte

Tom Robinson
Senior Manager
Consulting
Deloitte

James Moorhouse
Manager
Consulting
Deloitte

Organizations are under pressure to deliver value through data-insight-driven solutions in a financial market that is fast-paced and in businesses that can be complex and constrained by financial regulations.

Over the last decade, there has been a trend in financial organizations to remove siloed data sets and introduce people and processes to tune their data assets into a state whereby they can deliver traditional use cases and make insightful decisions at pace worldwide.

Traditional business intelligence and analytical use cases are maturing and new advanced analytical use cases are emerging that were previously unachievable. For example, insurance businesses are now processing large volumes of sensor data to analyze driver behavior and provide tailored premiums. The analytical processing required for this is typically unachievable in a traditional data platform. Organizations are choosing to create siloed solutions for each use case rather than thinking strategically and building a platform that caters for multiple needs. Organizations now require a fusion of cutting-edge technology ecosystems and flexible delivery models to satisfy all use case types. Organizations now require a **next generation data platform**.

"The next generation data platform is a fusion of disruptive and state-of-the-art technology with flexible delivery methods and tight and effective business partnering."

Martin Lidl
Financial Services Director within Analytics & Information Management Consulting, UK

A data platform is a centralized system that combines scalable flexibility with distributed data storage and computational power for acquiring and analyzing large data sets to provide users with reliable and accurate data. On the surface, many data platforms are achieving expected outcomes for traditional reporting but are struggling to keep up with the demands of the business for new data sources, larger volumes of data, and support for advanced analytics and machine learning. Sound familiar? If you strip everything back, is there anything different from the same structured solution that has been applicable before, but on a much larger scale? This, when coupled with the advent of big data and cloud technologies that are just starting to achieve wider adoption, means that organizations now have the opportunity to transform their legacy "fixed capacity, fixed use case" data platforms into genuinely flexible and scalable solutions.

To build a next generation data platform, we must first determine the underlying needs of an organization, the fundamental capabilities that the platform must offer, and the use cases that it should be used to deliver. We can then assess their existing IT landscape to augment existing components that work and replace legacy tools that are not fit for purpose to deliver incremental value.

Organizations should never replace a solution with a like-for-like alternative before making this assessment, or they risk associating an overestimated cost benefit with a solution that delivers the same previous value. ▶

Figure 2: Old World versus New World

Old World		New World
Fixed capacity		Infinite scalability
Structured data only		Support for all data types
Traditional use cases only		Traditional and advanced analytical use cases
Long lead times on data acquisition		Rapid new data source onboarding
Frustrated data scientists		Enabled and empowered data scientists
Large point-in-time investments		Pay-as-you-go infrastructure
Data understood by power users only		Democratization of data through data catalogues

 "Organizations have been excited about big data technologies and the possibilities they can offer, but at the same time not strategic enough when implementing them. Technology should not be implemented for the sake of it. This requires strategic thinking about how your technology estate enables the business to deliver its vision."

**Martin Lidl
Financial Services Director within
Analytics & Information Management
Consulting, UK**

There are six key themes to be considered when implementing a next generation data platform. They are: **governance; use cases; infrastructure; privacy and security; tooling; and delivery approach**. These principles provide organizations with an approach to implement a top-down, all-inclusive strategy that can deliver tangible business value at pace across an organization.

A next generation data platform tackles these principles with key functionalities. The following outlines five stand-out technology features that address these:

Intuitive data management
An influx in data has meant financial institutions have had to implement data governance frameworks quickly or risk their data becoming unusable. Recent legislation, such as GDPR, has seen even more focus in this area, but rather than taking a holistic view of this problem, a "fast and dirty" approach has been applied to meet EU-imposed deadlines. This has left organizations in a position whereby they are compliant from a regulatory standpoint, but their data management capabilities are limiting.

Master data management and metadata management are often an afterthought when managing data, but if applied correctly they can act as valuable tools. They work in harmony to solve data management problems through:

- **Trust—**Recording information on data quality, lineage, accountability of data stewards, and coverage

- **Collaboration—**Intuitive user interfaces support communication and task sharing, rewarding user contributions

- **Discovery—**Providing a comprehensive view of data assets with summary statistics

- **Governance—**Supporting cross-functional workflows, complying with legal requirements, and providing appropriate access rights

Figure 3: Next generation data platform capabilities

Data exploration
Since the introduction of big data technologies, the approach toward data ingestion has been "more is better," with financial institutions on a mission to persist data in a single location and transform from a product-focused business to a customer-focused business through customer segmentation. Although this method has been successful in breaking down silos, it has led to data platforms containing sets of varied data types, thus making a data consumer's task of finding useful insights like trying to find a needle in a haystack.

To understand data and formulate value-driven use cases such as personalized marketing by analyzing customer spending behaviors, the following core concepts should be considered:

- **Analytical sandboxes—**User-controlled areas that are segregated from gold-standard production data. They provide a means to copy data into a secure experimental environment without having to rely on data engineers to curate data. Users can blend datasets

and create working pilots that, once completed, can be passed back to engineers to productionize and share with a wider audience.

- **Data science tools—**There has been a sharp incline in advanced analytical use cases, particularly within fraud detection. The varying data formats and use cases mean there is no longer a single "one-size-fits-all" solution. To ensure data scientists are not impeded, they require an extensive range of tools that can be used together to detect unusual customer spending habits using predictive analytics.

- **Self-service dashboards—**A next generation platform should reduce the dependency on IT by offering self-service dashboards to business users and empower them to build their own analytical reports. Embedding analytical capabilities into the business requires multiple layers of collaboration, consensus, education, and change management—something that can be achieved using Deloitte's "Insight-Driven Organization" framework[1].

1. https://www2.deloitte.com/uk/en/pages/consulting/solutions/insight-driven-organisation.html

"Utilizing your Data Platform for Advanced Analytics, you will very rarely find the need for multi-year projects. Delivering value incrementally through targeted, self funding use cases is incredibly powerful and a very natural way to bring about change."

Martin Lidl
Financial Services Director within Analytics &
Information Management
Consulting, UK

- **Pre-built/trained machine learning models—**Good machine learning models can take weeks to configure, with model training limited to the volume of classified data that can be processed. A next generation platform promotes the sharing of proprietary models created in-house with the wider organization to reduce rework. Out of the box (OOTB) machine learning models follow the same path as OOTB data models that were introduced to financial services years ago.

- **Automated infrastructure—**As customer transactional data grows, technology dictates that platforms can be scaled to accommodate this. However, once a data platform reaches a certain size, even the simplest of tasks require a considerable amount of effort and time, eventually becoming unsustainable.   In order for next generation platforms to be truly future-proof, the following processes should be automated:

  - **Cloud deployment—**Financial services have tried to modernize their infrastructures by adopting cloud tools, but these can still take time to install and align to organization security policies and infrastructure. The introduction of cloud technologies allows next generation platforms to be implemented and configured before saving the set-up to an image, which can be reused at a later date. This provides IT with the control to automate the deployment of a data platform with varying size and power at will.

  - **Auto-scaling—**Scaling a data platform can be difficult depending on size and complexity. Traditional scaling involves monitoring the capacity of a cluster at set intervals and manually spinning a node up/down to adjust to changing user requirements. The problem with this approach is that users are unpredictable. If five data scientists were to train their machine learning models at the same time, a data platform would need to be dynamic enough to scale quickly and ensure performance remains consistent.

  - A next generation platform should monitor performance autonomously and automatically adjust capacity to meet changing user needs to maintain expected performance.

  - **Cloud optimization—**As next generation platforms implement auto-scaling features, you would be right to think that costs could soar due to poor performing queries over-utilizing resources and the platform reactively scaling up. To prevent this, cloud technologies coupled with a defined set of policies can work in tandem to ensure instance types, availability zones, and resources are optimized to manage a data platform at the lowest possible cost.

All-inclusive security
As data platforms grow ever more complex, the number of potential vulnerabilities increases exponentially. Changes in responsibilities and complex permissions can provide opportunities for unauthorized data access or even loss. If data owned by a bank were to be compromised, large fines would be imposed by regulators and their brand image would be destroyed.

Security should be at the forefront of any financial institution's platform, underpinned by a security philosophy that provides   guardrails—a security or governance framework that removes the opportunity for an organization to get it wrong or be compromised—to secure business assets without restricting innovation or pace.

Metadata-driven ingestion
There are traditionally five key methods that are largely adopted within finance to ingest data. Change data capture, streaming, direct database access, manual file transfer, and event-based messaging are all proven methods to acquire data, but they can be slow when trying to ingest a source. The process of building ingestion pipelines is difficult when IT landscapes are evolving, heterogeneous data source schemas are changing, and the quality of data is varied. Traditional data ingestion

"Next generation data platforms are an evolution of organizations' existing technology estate and delivering them without building up a mountain of technical debt requires strategic vision, a sound methodology, and flexible delivery methods."

Martin Lidl
Financial Services Director within Analytics
& Information Management Consulting, UK

frameworks are now dated and metadata-driven ingestion is becoming more popular. Metadata-driven ingestion is an approach to building ingestion pipelines by using a configurable set of attributes to define the common characteristics of data that determine extraction, transformation, and loading behavior across pipelines. This allows businesses to minimize dependency on data source technologies, reduce development and effort repetition, and therefore increase the speed of ingestion for new data sources.

Delivery approach
Solution implementations within financial services have been historically sluggish and detached from the business. A next generation platform requires a delivery approach that promotes agile ways of working with a DevOps paradigm from data ingestion to industrialization.

A next generation delivery approach provides an accelerated framework for implementation. This four-step approach can be used in different agile flavors for organizing tasks, releasing cadence, scheduling, and carrying out an iterative implementation:

- **Ingestion—**Managing and prioritizing data sources for acquisition, ingestion, and provisioning of data will help identify golden sources of data, gaps in

capabilities, and a controlled method of metadata ingestion for structured, unstructured, and streaming data.

- **Data controls—**Performing data analysis using tools available for discovery will help profile data and identify early data quality concerns. Combining this with associate data and linking to a business glossary can help pinpoint relationships between data sources and track data lineage. Finally, the classification of data will facilitate the early conception of data privacy and masking rules.

- **Industrialization—**To roll out repeatable, self-learning methods to operationalize proven methods or concepts, non-functional requirements relating to prioritized use cases must be collected, logical and physical data models for consumption layers must be designed, and development, testing and deployment into production must take place in an agile manner.

- **Analytical modeling—**Analytical modeling must define and validate a hypothesis to generate actionable insights aligned to organizations' strategic goals. To do this, business objectives must be collected to generate a hypothesis based on prioritized business benefit. Data is then prepared in parallel, analytical models are optimized and run to prove/disprove the defined theory.

Conclusion

In an environment of exponential growth in data volumes, data complexity, and business demand for deriving value from data, the challenges of combining and exploiting data can only be met with a strategy focused on delivering incremental value toward an end-state that provides a future-proof, holistic data capability for an organization. This requires a platform approach that implements an integrated next generation solution across the full breadth of data capabilities.

A next generation data platform is intended for innovative environments that are flexible enough to accommodate a fast-moving technology landscape with a proliferation of open source solutions to satisfy emerging use cases.