IBM Analytics

Taming the data dragon

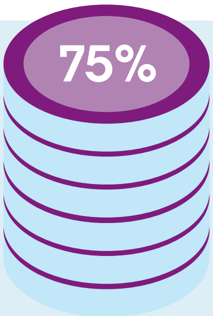
How to take data from anywhere, govern it everywhere and create value for everyone

Data has changed

Digital transformation of business has seen data volumes grow at an exponential rate. According to IDC, for every person on the planet, 1.7MB of data is created each second. Expansion of connectivity, provisioning of services into digital and mobile platforms, enablement of user-generated content within customer relationships – all are driving a relentless growth of the data points which an organisation needs to bring under control in order to process, manage, analyse and decision across this new environment.

Data volume growth alone is not new – it has been a perpetual feature of data management for decades. What makes the current era of big data more challenging is the increased variety of these sources. Internally generated data has expanded from conventional structured data on products, customers, sales and transactions to include notes created by customer-facing staff to texts and messages sent by customers, through to images and videos created within brand platforms both internally and externally, via the constant clickstream from every digital touchpoint.

New business intelligence and predictive analytics practices, such as the arrival of data science teams, are further expanding the range of data sets which needs to be discovered and analysed. Recognising the value of external data, IBM recently acquired The Weather Company – which operates the fourth most-used mobile app in the US – and is making its data accessible via the IBM Cloud. IBM partnerships with geodata provider Mapbox and consumer-data broker Acxiom mean these curated third-party data sets are now readily integrated with first-party customer data within IBM solutions.



By 2019, it is predicted that 75% of analytics solutions will incorporate ten or more exogenous data sources from second-party partners or third-party providers (source: Gartner)

Businesses are now generating more data than ever – and also consuming it at a greater rate in order to make effective decisions and deliver real-time customer support. For data management processes and technologies, that means significant changes to ensure these new demands are met.

IBM

A new kind of data infrastructure

Data warehouses are a well-established, stable and mature environment in which organisations can store business-critical data, ranging from transactions and financial information through to customer records, interactions and demographics. Whether created at an enterprise level or within each operating division, the data warehouse feeds core business processes, supports reporting and analytics, and allows the organisation to understand its current position via the different views created for each department.

This data infrastructure is not optimised for real-time, high-volume, unstructured data flows, however. Provisioning new data sources – especially third parties – can require a long lead time. Similarly, supporting data science, which typically involves the discovery and analysis of large data volumes in an exploratory way, can be too disruptive to the enterprise data warehouse which is optimised to support repeatable, continuous, well-defined business processes.

As a result, a new approach to data infrastructure has arisen which makes use of distributed storage and cloud-based or commodity hardware computing solutions, often running open-source applications. One of the appeals of these solutions, especially for data scientists, is the ability to create them for time-limited or purpose-constrained projects and then shut them down once that purpose has been realised.

A downside of these solutions is the risk of becoming "shadow IT" operations – technology projects sitting outside of the enterprise framework, including its controls, governance and oversight. The ability to analyse combined structured and unstructured data to build the value-driving insights or to operationalise models which are created in this experimental environment can also be limited.

80% of all data created is stored by enterprises, but only 0.5% is being analysed (source: MIT)

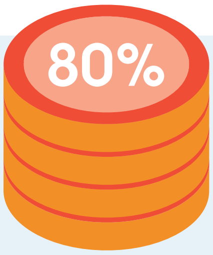
To resolve this, a new kind of data architecture is needed that blends the best of these two environments, preserving existing legacy systems while expanding the data footprint which can be accessed. Called a data lake, it accepts data flows from any source and brings them into a common platform for use. Data is stored in its raw, unrefined state and located, processed, refined and extracted as required.

IBM has progressed this a step further by cataloguing and classifying data on its way into the data lake, ensuring data is governed both while it is stored and at the point of use. This delivers important benefits to IT and data scientists since it allows for agility in responding to project needs, while keeping costs down, maintaining the resilience of business-critical data provisioning and preventing ungoverned data environments and usage from springing up. The result is a trusted data asset (open, monitored, maintained) rather than a data swamp (unbounded and unpredictable).

Business-use cases

Every business will have its own reasons for creating a data lake, reflecting its market position, technology adoption, maturity level and resources. A number of common drivers can be identified across deployments as follows:

Mobile apps – providing services and supporting transactions to customers via mobile devices challenges traditional data infrastructure, which is designed for batch processing rather than real-time decisioning and validation. With a data lake, contextual information (such as device ID and location) can be combined with structured data (account number, password) to drive mobile app-based services in a way that is secure, resilient, reliable and replicable.



Data scientists typically spend up to 80% of their time and effort on data preparation (source: Forrester Research)

Predictive analytics – decision-making based on propensity models, such as next best action or product recommendations, need to blend customer-specific data (purchasing history, preferences) with broader context (purchasing situation, product availability) and even external reference data (weather, holiday calendar) to deliver personalised, relevant communications with a high degree of accuracy.

Fraud detection – cyber-criminals increasingly use valid customer credentials to access and take over accounts, gain entry into business systems or extract valuable, confidential information. The new generation of fraud detection not only checks these credentials against any reported theft or loss, but examines social and digital data which might reveal an inconsistent or unusual dimension, such as access via an unknown IP address or foreign location. Achieving this in real time across the broadest possible data set, ideally from within a governed environment, increases the likelihood of spotting fraud and reduces false positives.

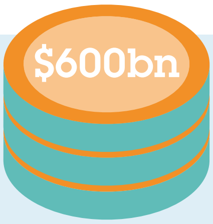
Data enhancement – in addition to the view of the customer generated by the business across all of its interactions and transactions, there is an advantage to be gained from introducing external data sources that give a perspective on the customer across their entire portfolio. Provisioning the data lake with these sources allows for dynamic changes both to the data and the sources themselves. IBM has acquired The Weather Channel and partnered with a number of third-party data owners, such as Mapbox and Acxiom, to pre-integrate validated and curated information into analytical and decisioning environments like the data lake.

Data science – new insights into customer behaviour, marketing opportunities, product features and service propositions increasingly arise from examining the largest possible data set, blending internal data with curated external sources, to spot significant patterns. If this data has been pre-assembled, catalogued and governed, the time required to discover and assemble it prior to exploration – which can take up to 80% of data scientists' time – is reduced and time-to-value is improved.

Data monetisation – an important new dimension of expanded data availability, especially for marketing, is realising additional value both through better customer engagement as a result of increased relevancy and also the potential to resell either the enhanced data, derived variables or targeted media opportunities.

Overcoming the obstacles

Data lakes can be transformational, enabling new processes and value-driving activities. At the same time, they can also be disruptive to existing practices since the technology associated with data lakes is still immature. In some functions, which view themselves as the owners of specific data types, there can be a reluctance to pool that data in order to create a bigger, more holistic resource. Top-level executive support is essential to create the mandate for dissolving these data silos.



Data quality problems cost US businesses $600 billion annually (source: Data Warehousing Institute)

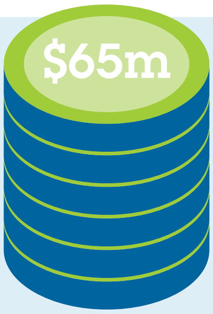
Governance needs to be applied to the data lake to ensure it becomes a trusted data source, rather than a formless landing area in which data is stored without consideration of its validity, value or shelf life (the rate at which it decays or changes and therefore its lifespan). A cross-functional team with representatives from lines of business, legal and compliance, IT and analytics can ensure future problems are avoided by drawing up sustainable policies and data definitions. Agreeing these definitions and deploying them through a metadata management layer maintains the consistency and usability of data added to the data lake over time. At first, this process can highlight significant differences across the enterprise, such as how a customer or sale is defined, what units are used to record stock levels, address or telephone number formatting, and so on. IBM supports this process via an information governance solution which brings consistent rules and controls to bear across all sources within the data lake. An important feature of this approach is the ability to ensure that the same standards and controls operate in both the core enterprise data warehouse and the expanded data lake, and also to govern across on-premises and cloud-based systems.

Access controls and monitoring also need to be applied to ensure there is visibility over the usage of this newly enhanced data asset. It can be tempting for some practitioners to overreach their permitted access, such as data scientists wishing to use full personal data sets when masked, pseudonymised data is perfectly adequate. Ensuring oversight in this way also helps the organisation to meet its compliance obligations.

Enjoying the benefits

Taming the data dragon should lead to significant benefits across the enterprise, from improved productivity to increased effectiveness in sales and marketing. Given the potential scale and complexity of a data lake project, it is important to ensure metrics are applied in as many areas as possible to capture the full picture of this return on investment. Benefits to look out for include:

Reduced data warehousing costs – within a data lake, many of the non-core data elements currently stored in the data warehouse can be offloaded to lower-cost distributed storage.



A 10% increase in data accessibility will result in more than $65 million of additional net income for a typical Fortune 1000 company (source: Baseline)

Reduced systems integration costs – rogue IT projects often mask the true cost of external consultants as these are being paid for from within departmental budgets, rather than the central IT budget.

Lower demand for external analytics consultancy – provisioning data science teams with data and technology resources can eliminate the reliance on external experts to provide insights and predictive analytics.

Higher productivity from analytics teams – a typical experience with a data lake is for accelerated output of models and insight, in turn driving better marketing and sales performance.

Fewer data quality costs – from duplicate records to undeliverable addresses, poor data quality carries hidden costs that the data lake should remove.

Next steps

IBM has deep domain experience in the realms of data infrastructure management, advanced analytics, data transformation, and quality management, governance and data protection. These can be brought to bear through a number of methods of engagement, such as proofs of technology, proofs of concept, data labs and on-site project delivery.

To learn more about the benefits and opportunities associated with successful data lakes, visit:

ibm.biz/data_lake

IBM United Kingdom
PO Box 41
North Harbour
Portsmouth
Hampshire
PO6 3AU

Produced in the United Kingdom
September 2016
All Rights Reserved

References in this publication to IBM products and services do not imply that IBM intends to make them available in all countries in which IBM operates.