Big data in the banking universe

Laurence Le Buzullier
Founding Partner, Arenium Consulting

[special issue of *Réalités Industrielles*, february 2019]

Abstract:
The web giants have revolutionised the use and storage of data: big data.
Nowadays, all data is saved in "data lakes". These enormous containers of structured and unstructured data (such as videos and text) replace the former data warehouses. The cloud enables access to such huge volumes of data, while also reducing the technology costs borne by companies.
However, the potential financial windfall of data cannot be harnessed without human intelligence. To create value from data, it must be made intelligible and must be properly exploited. Data is more valuable when it is high-quality, up-to-date, and when it circulates.
In banks, data governance must be handled at the highest echelons so that new technologies such as artificial intelligence can be used optimally. Data regulations are becoming more stringent, and sector players who use these regulations to their advantage will have a real competitive edge.

"What cannot be measured does not exist", according to famous 19th century physicist Max Planck. Researchers, mathematicians and physicists have always attempted to model the world using data and equations. A season can be defined in terms of temperature, precipitation and daylight hours, a storm is characterised by the velocity of its winds and by its duration, a star is characterised by its luminosity, mass and temperature, and so forth.

Data describes reality through a prism. On its own, data has little value. Individual data points are indisputable, but as each one is unique, they are hard to interpret unless there is a point of comparison. The potential value of data comes from circulating, accumulating, exchanging and crossing data points.

The data that describes the world is now "big data", due to the volume of data generated every day. Each connected object contributes even more to this mass of statistics that quantifies the world around us. We become a series of numbers that measure and analyse us. The number of steps we take each day, what we eat, how we sleep, our heartrate… it's all recorded, stored and analysed. Since the mid-2000s, the global web giants have revolutionised the use and storage of data.

Data also comprises the photos we post, the texts we write, or all our online activity in the broad sense: a mass of "unstructured" data that describe our lives. Data comprises not only series of numbers that are well-formatted and structured, but also "unstructured", and therefore hard to interpret, data.

This financial windfall requires human skill and intelligence. To create value, data must be made intelligible.

Data in the banking sector

In this digital world undergoing deep transformation, banks are also faced with the digital transformation of their business. They must now adapt to this technological revolution or be swept away. However, to be fully successful in this digital transformation, banks must incorporate a data-centric approach into their changing business. They must go from a primarily product-centric setup to a client-centric, data-focused one. For banks, data represents a genuine source of wealth that is often poorly understood and especially poorly used. If data cannot be processed correctly, it loses all its value and warps the decision-making process.

Coinciding with this inevitable technological change, banking regulations regarding data are constantly being strengthened. Reporting reliability, privacy protection, anti-money laundering and combating terrorist financing are all subjects of great interest to the authorities.

Thus, instead of viewing banking regulation as a constraint, it would be in banks' firm interest to consider it to be an actual opportunity to overhaul their entire setup and their data governance.

On 9 January 2013, the Basel Committee published BCBS 239, a set of principles whose goal is to bolster banks' capacity to produce regulatory reporting with reliable data. Moreover, the EU's GDPR (General Data Protection Regulation) – which was passed in April 2016 and came into force in May 2018 – aims to scale up protection of personal data, requiring companies to define data governance from the design phase of a new product or service. Lastly, banks are faced with growing obligations in terms of "Know Your Customer" (KYC) processes and in anti-money laundering and combating the financing of terrorism (AML/CFT).

Data security must also be strengthened. Clients must be informed within 72 hours of any data breaches involving their personal data; the reputational risk of such breaches is high for banks. Not only can a data breach cause actual harm to the bank via the disclosure of confidential information to third parties, but especially, it adversely affects the trust that clients and the authorities place in that bank.

Banks must view all these regulations as a whole, not as separate constraints. It is worthwhile for banks to incorporate regulatory compliance into an overall data-centric corporate project. Such a project should go beyond mere regulatory constraints and enable banks to steer their business more effectively, saving them time and giving them an opportunity to use promising new technologies such as artificial intelligence. Thus, it is now crucial for banks to roll out a genuine data-centric strategy.

Data governance in banks

A data governance programme entails organising all of a bank's procedures for collecting and using data. Therefore, the data governance programme should apply to the whole bank, not just at the level of one department or process. For the bank, the purpose is to comply with legal obligations and define an internal framework for optimising data use. Data governance must focus on aiding decision-making, and not simply on collecting the data.

Banks have often been set up in a silo structure, with business lines existing side by side and support functions all being separate from one another. However, data is cross-cutting by its very nature. It circulates from one activity to the next, while undergoing transformations. The full data lifecycle must be managed, along the entire production chain and across the various organisations. At the beginning of the production chain, data is input by an operator. It is transformed and used in marketing, risk and financial reports over its lifecycle, before reaching senior management at the end of the chain, where it is used as an aid in decision-making. And we mustn't forget the old saying: "Garbage in, garbage out!"

A company-wide data-centric strategy must be implemented at the highest echelons of banking management.

From the data warehouse to the data lake

Designed in the 1990s, the "enterprise data warehouse" brought the information from a company's various management systems together in a centralised architecture. It was modelled to provide the most detailed level of granularity and historical data over long periods. To assist the different business lines in using data, the data warehouse model also included "data marts", with a business line focus enabling functional issues to be addressed more specifically. The data stored in a data mart was generally pre-calculated and aggregated so that it could be more easily retrieved on a pre-defined or on-request basis within a reasonable timeframe for users.

Data warehouses were unable to store “unstructured” data such as videos, photos or raw text.

To overcome this difficulty, the “data lake” concept was introduced in 2014. According to James Dixon, the CTO of Pentaho[1] and the man who coined the term, in a data lake, “data flows from the streams (the source systems) to the lake. Users have access to the lake to examine, take samples or dive in”.

A data lake contains raw data, stored in formats that have not been significantly transformed. Each user should be able to dive in and find what he or she is interested in, for statistical analysis or reporting purposes.

However, the introduction of the data lake very rapidly gave rise to a problem: without governance, a lake of unstructured data is only accessible to the very small number of “initiated” users – the ones who know how to swim… Data that is dumped without a control system, without quality checks or governance principles is very hard to use.

The model of a static data lake with specialists accessing data to provide reports is already a thing of the past. Data must be integrated directly into banks’ processes, to guide clients and the bank in real time, to trigger alerts, to initiate corrective maintenance measures, etc.

The cloud

These innovations are made possible by the development of the cloud, which has coincided with the developments of big data and artificial intelligence. The cloud enables a company to develop secure, data-driven apps, and to stay at the cutting edge of technology. It is complicated for internal IT teams to upgrade existing services at the same pace because technology is changing so quickly. The cloud ensures the security of data and processing, while pooling technological developments. Faced with the cloud’s exploding popularity in the regulated financial sector, the EBA (European Banking Authority) published its “Final Report on Recommendations on Cloud Outsourcing” in late 2017. Each banking institution must assess the materiality of cloud outsourcing, must have the right to audit its provider, and must define adequate controls – as for any outsourced essential service.

In addition, new technologies bring about new risks. Outsourcing data and data processing to the cloud creates a risk for banks that can be described as systemic (cybercrime, interruptions to data availability, etc.). The cloud is an information system, so by definition, it is not fail proof. In France, ANSSI (National Information System Security Agency) has set up a security credential system for cloud service providers.

[1] Pentaho describes itself as “the first major supplier of decision-making solutions to offer functions for big data”.

Industrial platforms

Data can be regarded as a commodity, but it is a very special kind of commodity. It is not scarce (in fact, it is perpetually expanding), it is still available when it is used, and it can be used as many times as needed (and by any number of systems) without losing its value.

The value of data increases with its reliability, and thus with the number of occurrences that characterise it. Data is more valuable when it circulates, is updated and is crossed with other data.

Some industrial players have understood this principle very well and have decided to pool their data to achieve mutual gains in performance and reliability. For example, EasyJet has entrusted its predictive maintenance activity to Skywise, Airbus's data platform. Skywise collects data from thousands of aircraft with the aim of improving their operations.

In the banking sector, GCD (Global Credit Data, created in 2004), a global non-profit organisation of 52 banks (as of 2018), is working on pooling these banks' data on loan defaults and losses in order to enhance the performance of each member bank's internal credit risk models.

It would also be very beneficial for banks to pool KYC or fraud data to reduce processing costs and improve the performance of the systems used by all market participants. Several possibilities could be imagined, modelled on the "positive files"[2] used in most European countries. Data could be centralised by a trusted third party, such as the Banque de France, or by a private-sector entity. Or several competing systems could coexist, exchanging at least a certain level of data with one another.

Bank account aggregators may also have a role to play in the pooling of information. These new services let consumers use a single app for access to all their bank accounts at different banks.

The new Payment Services Directive PSD2 – which went into effect in January 2018 and must be fully applied in September 2019 – requires banks to provide certain third parties with access to bank account information via a secure communications channel. Such third parties include account aggregators (upon request from clients). The banking landscape is currently experiencing a major transformation and will continue to see changes on these topics in the years ahead.

[2] A "positive file" would list all the loans of French borrowers in all banking institutions. Such a file exists in most European countries. In France, payment incidents alone are listed in a "negative file".

Big data is nothing without intelligence

We often speak of big data by referring to the "4 Vs" or even the "5 Vs" that correspond to its key features: volume, variety, velocity, veracity (and value). But big data is nothing without intelligence. Data only has meaning if it is applied to analyse, make decisions or take action. "Analytics" can be defined in brief as all the techniques that "make data speak". These statistical techniques are not new, but the surge in computer processing capacity has enabled exponential growth in inferential statistics in the past few years. Credit scoring, for example, was developed in the US in the late 1960s and has been used by consumer credit specialists in France since the mid-1970s. Yet these days, artificial intelligence techniques not only make it possible to identify default risk at the time loans are approved, but also to detect suspicious fund movements in real time, to dialogue with clients, to carry out simple banking transactions, etc. Thus, although the foundations for artificial intelligence date back to the 1960s, computing power has enabled a shift from a "decision tree" to a "random forest", for instance.[3]

To conclude, we could even add a sixth "V" that is indispensable to big data: Visualisation. The term "dataviz" encompasses all data visualisation techniques. These techniques allow data to be presented in a visual way, in graphic form that is easier to read than long tables of figures. A good graphic representation must immediately highlight the key message, either allowing the analysis to go into greater depth, or enabling a conclusion to be drawn and possibly an action to be taken. With the digital era, the multiplicity of information and acceleration of the decision-making process, these storytelling techniques have become crucial. A story must be told to make sense of the information.

Data is a vital tool for banks. Sector players who use it successfully will derive a genuine competitive advantage.

[3] The Random Forest algorithm is a machine learning technique that constructs multiple decision trees (up to several hundred), trained on subsets of data that are slightly different each time.