Deep Science Publishing
https://doi.org/10.70593/978-93-49910-41-6_4

Chapter 4: Harnessing big data: Turning volume into value in financial services

4.1. Introduction to Big Data in Financial Services

Today, the creation of data is exponential, with more data created in the last two years than there has been in the entire previous history of the human race. Data is used and generated by every sector of the world today. Growth in electronic data creation and aggregation has exploded in the last few years with the rise of social media sites, blogs, Q&A, and content distribution through websites. The big shift from tradition to the new era of big data has been hailed by many experts across multiple industries. Retailers choosing not to embrace big data to realize their analytic potential could reluctantly cede around 60 billion in profit in the next five years. Financial services markets are going through a similar data explosion, with the demand for data to help drive business decisions exploding at a faster rate than the capability of many established internal data warehouses (Manyika et al., 2011; Chen et al., 2014; Gai et al., 2018).

To frame the boundaries of this report, it is useful to turn to literature on digital information in the field of finance and data analysis. We use the definition suggested by IDC: 'Big Data is a shorthand label that encompasses a combination of several major trends and changes occurring in the IT industry that aim at offering greater insight and better business decision-making analysis.' These trends include an increased volume and variety of data being generated, asset management, and business intelligence. Whether considered internally or more broadly as a part of the state of the global economy, financial services firms have seen two recent, decades-old market trends that suggest either a beginning or a huge transformation with rich payoff potential.

4.1.1. Overview of Big Data's Impact on the Financial Sector

Many sectors are experiencing the great potential that big data holds for improving decision-making and productivity levels and for addressing new or ongoing challenges. These sectors include health, transport, and others, but few have more to gain than the vast and diverse financial services industry. Indeed, with its increasing size and globalization, the financial services inventory of products and services has adapted to easily capture, store, and manage big data, as well as to quickly analyze big data resources. Today, every interaction with banking, insurance, equity and bond markets, or public and private pensions generates valuable data. Each of these interactions offers the opportunity to improve a financial institution's performance or to better meet customer needs with better products or to understand systemic risks. Essentially, the challenge is that of transforming massive amounts of available data, not all of which are reliable, into value (Wamba et al., 2015; Wang et al., 2016).

Efforts are required to standardize and classify data, undertake process reforms, and cope with data privacy issues to allow the financial sector to fully harness new information and communication technologies. Regulatory authorities must also invest in new and better tools to leverage big data whenever necessary. However, although much progress has been made and continues to occur at an ever-increasing pace, some value-diminishing attempts have gathered the most attention, prompting academics, regulatory authorities, and financial institutions to have a more balanced perspective and informed conversation regarding big data.

4.2. Understanding Big Data

Big data is most simply defined by three dimensions: Volume. Organizations collect data from a variety of sources, including business transactions, social media, and information from sensors or machine-to-machine data. In the past, storing it would have been a problem, but new technologies have abated this issue. Today, data warehouses and other solutions can store large volumes of structured data, and some big data systems can store and manage even more data much more cost-effectively. Velocity. Data streams into businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors, and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations. Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data, and financial transactions. For some organizations, this might mean tens of terabytes of data; for others, it may be hundreds of petabytes.

Big data is typically defined by volume, velocity, and variety, but for some organizations, these three Vs are not sufficient for capturing the foremost attributes of big data. For some organizations, big data is also about creating value. This could come

in the form of visibility, intelligence, and agility. These three Vs and three Is – and other similar ideas – can be collectively used to provide a helpful definition of big data. In addition to recognizing the sheer scale of data, this definition of big data also recognizes the real-time nature of much of the new high-frequency trading data, along with the digital exhaust that investors leave on social media and other platforms, such as collective intelligence data. This view of big data can also help firms focus on the most valuable data and determine where such data resides on the trust continuum, with its increasingly time-sensitive demands. Such clarity can also help firms distinguish themselves and take advantage of big data technologies that are increasingly available.

4.2.1. Definition and Characteristics

Big data opens up growth opportunities for financial services that were previously out of reach, and at the same time promises substantial dislocation. Big data refers to data sets whose size is beyond the ability of typical data software tools to capture, store, manage, and analyze. This difficulty arises from a sharp increase in volume, velocity, or variety of available data. More recently, we've seen data characterized not just by these three V's but sometimes by a fourth V, voracity. Big data is unstructured, has unknown value, and exists in volumes where large samples can become analytically irrelevant indicators of future behavior. It has been attributed a number of characteristics, not least of which is the capacity to allow more informed, faster, and potentially better decision-making.

Fig 4 . 1 : Big Data Analytics in Finance

Big data is not a technology; it is a cultural movement driven by technology. It is based on the dual dispute that exponential advances in processing power do not impose limits on data capture, and that data is no longer the exclusive domain of IT departments. People and devices are both sources and subjects of the data, which is made up of both

structured data created by transactional systems and operational data derived from legacy systems, as well as unstructured data produced by either people or devices, or in both cases. Big data can reveal patterns and trends that have been invisible until now, and if acted upon, can allow organizations to develop new business processes and drive sustainable competitive advantage.

4.2.2. Types of Big Data

We define big data along three dimensions: volume, variety, and volatility. Of course, no single measurement threshold across all three dimensions will precisely delineate big data from not-so-big data for all relevant applications. In fact, no single threshold is likely to accomplish this for any of these dimensions for even a single application. Nonetheless, it is useful to proceed with the usage of the big data term by discussing, however inarticulately, some general benchmarks for each dimension. Volume is generally the most striking dimension of big data to those unfamiliar with it. Unlike traditional data, however, the meaning of big data is not its absolute size per se. Rather, big data is about deriving value from much bigger, more complex data. For a subset of research and policy problems in financial services, some of the most interesting and important of which are discussed in this context, an enterprise's data will reach the big level considerably earlier than the world's aggregate data will. The second dimension of big data, variety, recognizes that much of the new digital data comes in the form of text, images, sounds, and video. Traditional data processing software and statistical software do not effectively handle these data types. The good news is that traditional quantitative research and analysis approaches could benefit from the substantive information that is available within text, images, and the like. The performance of these approaches is expected to be enhanced by big data's volume dimension, as well as data may help to overcome some of the limitations that had previously stymied the constructive portrayal and treatment of between textures and image features. Rapidly advancing analytics, and big data itself, are the tools that will permit the finance industry, and financial regulators, to derive value from financial firms' and financial markets' unstructured data.

4.3. The Role of Big Data in Financial Services

Financial services have much to gain from both traditional and untapped sources of data. To date, much of the attention has been on digitized and software-generated payments, markets, client activities, and online interactions. These are likely to remain the most important organizational data sources for the near future, and we will pay careful attention to what could limit their value. Banks, investment and insurance firms, wealth and pension managers, and market infrastructure firms have a much larger universe of

spatial, statistical, and unstructured data to assist them in operations, product demand, risk monitoring and management, understanding clients and markets, product development, and innovating methods to tailor services and effectively meet statutory objectives and public responsibilities. Data and analytics form a closed loop. The more regularly analyzed a particular set of data, the greater the probability of uncovering actionable insights from that data. The more insights generated from a dataset, the greater the demand for new and additional data to test and refine those insights. This can lead to a correspondingly increasing stock of data, by a rapidly changing array of types. Data they wish to keep and use comes from a broad array of both internal and external sources, spatial, structured, and unstructured, or proprietary and non-proprietary. A small share of this data is newer data collected for specific business purposes, but the largest share of valuable data that financial services firms possess and wish to use is often in the form of behavioral, credit, payments, regulatory, and transaction histories.

4.3.1. Risk Management

Risk is a central concept in finance, predominantly entailing a financial dimension, but also other qualitative factors, such as guaranteed solvency, sustainability, and long-run stability, besides the financial return. Risk management is defined as a process of comprehensive planning for a change, often for an emerging condition or situation viewed as a potential for loss. Specifically in financial services, risk management involves the understanding, prediction, assessment, measurement, management, and controlling of risk to establish the trade-off between risk and return. Risk management provides a means to balance multiple stakeholder interests, a better-quality decision model, and enhances financial system efficiency. In terms of big data, risk management crops up as one of the most important areas of financial services. Risk management incorporating big data is not purely a chance, but a necessity to adapt to the integration of finance and the real economy and advance the reform of, hence the innovation in, the financial system. With big data, financial risk management can, for instance, take the effectiveness, efficiency, and accuracy of value at risk into account; improve adverse selection and moral hazard in aspects such as asymmetric information, incomplete contracts, imperfect agency, and commitment; estimate credit risk flowing from financial products such as options, forward rate agreements, and whipsaws; and forecast long-run firm solvency, considering structural and strategic risks.

4.3.2. Fraud Detection

Fraud detection has been a traditional application of big data in financial services, long predating the big data era and involving the efficient analysis of large data sets. The

challenge for banks here is to detect cases of fraud that are rare hidden events in a massive flow of normal transactions. While the data size for financial transactions increased substantially, the transaction throughput required for real-time detection powered by hardened business logic was a much more limited resource. In this environment, it was computationally impractical to analyze each financial transaction in detail and to ascertain a true anomaly score for the transaction, which is the transaction's likelihood of being suspicious. It was, however, possible to perform a first-line analysis of each transaction and to retain for further investigation a very small fraction of them, putting further investigative effort into this retained stream, either automatically by system analysis or interactively by security and risk specialists.

One technique for the further investigative effort comes from an open-source intrusion detection tool, which has been commercialized. While this tool focuses on network packets of all kinds, transaction values and details are quite comparable to network packet data, which gets analyzed by numerous specialists worldwide who use the tool on global information traffic. A recent extension is applied on the basis of a sibling of this tool, which does anomaly detection on the packet-based data already inline, i.e., during the data flow, in real time, so that suspicious packets may be immediately categorized and discarded as usual. For a closer inspection and investigation of why the packet content is unusual, all of the suspicious packets are additionally stored in archive format for deeper offline analysis and re-inspection. Analogously, a select group of financial transactions can be inspected in depth after their additional contents have been stored on the basis of a detected overall risk score, to combine speed and effectiveness in real-time fraud detection. Other traditionally rare event detection applications in financial services for big data concern the interpretation of, e.g., news, statements, filings, or analyst reports in order to detect potential impending business or stock actions.

4.3.3. Customer Insights

In addition to identifying fraud, customer insights derived from analyzing vast volumes of transaction data can also drive product and service innovations, marketing insights, and customer service applications. Analyzing the customer insights would allow banks and financial services to build enhanced customer profiles, making data and insights about their customers and using their products and services more valuable. Banks and other financial services companies can use the details about the purchase habits of their customers and, combined with location and social network profiles, help their customers budget, understand their fees, and even suggest more appropriate financial products. Traditionally, banks collect most data about customer relationships but have not been able to translate this data into an analysis and modeling process that can be used to provide additional services that are very relevant to the company. Low-cost banking

functions have often been a key consumer support function, but only now can these really add value through the business intelligence that can be driven by big data. Big data can help banks accurately brand consumer spending and offer the customer high-demand financial services. The idea goes beyond viewing the analysis of large volumes of customer data from mobile devices as a great way to provide insight into consumer behavior. The selling point for consumers would not only be convenience but also comprehensive insights into consumer habits with the goal of setting up financial health.

4.4. Data Collection Methods

Similar to currency for an e-commerce service provider or competitive differentiation for an air express provider, transactional data is inherent to developing tailor-made products and assisting customer decision-making, as well as for maintaining comparisons on service value. Accordingly, data collection in this study relied, for the most part, on data mining, wherein it referred to data collection from numerous internal and external data sources to assist data analysts in assessing their primary data of interest. First, secondary data was collected within the scope of the so-called terabytes of transactional and other data produced on a daily basis. The sampled bank realized significant credit card transactions each day, with a large number of customers, making it conducive to the bank's achievement of its consumer service objective, transcending pure transactional operations to include the practice of financial services marketing itself.

All data were confidential and organized to ensure that the research could undertake the necessary steps to achieve privacy protection. Second, publicly available research was accessed on companies showing similarities in business strategies, particularly retail strategy and customer portfolio, which could assist in benchmarking the bank's performance against that of other global leaders in the retail financial services industry. Results from various companies formed part of the basis of the critical options made throughout the data analysis process, in particular with respect to the contribution to case study research recommendations and study limitations. Third, a partial list of service companies offering technology solutions was reviewed, as the particular analytical data techniques created specific cost constraints for exploring big data within a secure environment, impacting both data collection and analysis phases of the study.

4.4.1. Transactional Data

This category of data comes from systems of record that track transactions, process invoices and other documents, or otherwise record operational happenings. Retail bank deposits are one example; the history of all checks cashed, debit and credit card transactions, deposits, and withdrawals provides the raw material for analyzing the financial behavior of an individual or business entity. Brokerage account data is another type; all the trades, deposits, and cash outflows from an individual trading account tell the story of that account's trading and investing history. Utility billing records, mobile phone call detail records, and pay television subscriber data are other mass transaction data types that can be profitably analyzed using big data techniques. Combining transactional data sets can also yield insights and signals that could not be obtained by single set analysis. Many potential investment and marketing-related new insights become possible when brokerage or deposit transaction data is combined or merged with other extant data sets. For example, combining demographic information not typically found in brokerage or deposit account records might enable one to find and reach certain retail investors who would qualify for a special type of new financial product before competitors. The resulting marketing advantage would be temporary because of the possibility of new retail investors arising as generations age, but during the product's initial period of competitiveness, such data-based customer targeting could represent considerable competitive advantage.

4.4.2. Social Media Data

Social media and the data it generates have become a significant driver of brand, product, competitive, and even market sentiment information across industry sectors. The capability to track how users interact with different products and services, companies, or brand mentions in real time is far quicker than traditional survey methodologies and public filing data, which is released at specific times by the vendor or governmental institutions. By tracking the frequency of specific terms a few months ago, an investor might have anticipated recent disappointing earnings reports and subsequent share price declines. Investors can also tap into markets dominated by peers of their choice, select and weight indicators, and track portfolio performance relative to their chosen benchmark. In principle, these ideas go back to the effect of the mass psychology of crowds, but now, data can be sliced and diced at every frequency, by various demographics, and can be seen in real time.

The use of social media data for investment management, therefore, seems a logical and exciting development. Not surprisingly, there is a growing interest in the potential value of alternative data from various online platforms. Research has demonstrated various correlations between social media data and financial markets. Prior work confirms that data from different social media platforms has explanatory power over trading volume

and future volatility, enabling market activity to be inferred within minutes of sentiment being generated. Other work distinguishes between various types of media for different short- and longer-term trading activities. In further work, it is found that social media data contains information that a large subset of traditional short-interval returns cannot predict. It has some predictive power for returns at early intervals and negative predictive power at larger intervals. Messaging board posts have a causative effect on stock price movement as well. Other research has shown that social media data can transcend financial markets.

Fig 4 . 2 : Social Media Data

4.4.3. Market Data

Ryan Spangler examines the widespread application of unclear or nonsensical limitations or exclusions in copyright law. To make an obvious example, a commercial company should not be able to copyright "market data" simply by having it appear in a proprietary form in their product. Nevertheless, they can, so long as they jump through the requisite hoops and somehow transform it from uncopyrightable fact into copyrightable, proprietary form. First, define what sort of real-time market data you are talking about. Are you receiving a direct live data feed from the exchange? Are you using a trading terminal that streams market data? Are you using a website or brokerage account that streams real-time quotes? Are you just watching CNBC? Data providers negotiate special fees and contracts with the stock exchanges in order for the right to distribute their data. The exchanges generate revenue from this market data by leasing it to individual investors and professional traders. Market data includes the information

that is being generated for all the financial instruments being traded. This includes all the quotes for the traded securities. Various rules apply to the data distribution based on various levels of access, including licensing fees, inclusion fees, membership fees, etc. Included also might be licenses from individual companies and disclosure rules due to the potential for market manipulation.

4.5. Data Processing Techniques

Different data processing techniques have been developed to make financial services data more amenable to data analysis, creation, and validation of assessment techniques. In many cases, these techniques are based on methodologies that have been honed in other industries. Nonetheless, there are some unique elements related to working with financial services data that have been the basis for the development of some of these techniques. This section introduces some of the most widely used tools in the big data toolkit, with a focus on tasks such as business data creation, data quality, predictors, and some basic relations of distributive analytics.

Creating new business data from existing data is a challenge, but it is essential in situations where useful data attributes do not exist; for example, predictive algorithms always perform better when more relevant attributes are available. It is also essential when data is lacking in some respect and can be simulated to facilitate prediction, estimation, or validation. Data quality is a crucial aspect of financial services analysis. For decision-making that is based on prediction, the prediction failure can be an overall concern only when essentially predicting firmly rooted dependent attributes like customer defaults. The introduction and exchange of new data in a business environment pose a technical challenge, as new data generation and benefit identification are decoupled in the big data approach. Big data is associated with low observability, while traditional data processing is based on data distribution. Big data has very limited traditional data distribution attributes, like data variance, and is more flexible regarding one-shot data introductions compared to traditional data acceptance and use. In this approach, the value is extracted from initial data and data processing procedures from its association with the selected set of determinative attributes, rather than the value delivered for estimation and prediction. Data processing's data reframing technique, for example, is a tool for creating synthetic data attributes that are dependent on earlier identified determinative or indivisible data structures.

4.5.1. Data Cleaning

The factors that are hiding the impact of dirty data are usually sunk in the cost of the process. Operational units have direct costs already embedded in the cost. Since processes usually attribute to history, the legacy of faulty process methodology has a long train. These include policy variety in the underwriting area that has been established over long periods and process programs that are full of tweaks and "special cases." A process methodology that is led with automated computerized rules and models, which is in turn supported by simple process steps, is less likely to produce bad data. A first step is to validate a process flow to ensure that no steps along the way result in bad data. Doing before and after data counts that ensure each step has the data elements in the correct range and provides the correct distributions offers a means to ensure no step produces big data that may contain no information. Data should not be accepted unless it results in the required answers. If an operation center is producing data that consistently falls outside of the prescribed range, it is not contributing to the operational process and should be adjusted. The task of post hoc data adjustment should be minimized if the operation is in control, but constant control charts can be created that run combining an underlying data processing process that shows if there are any tasks where adjustments should not be made. The next class of dirty data is dirty metadata. Since the meaning of metrics is highly context-dependent, the wrong definition of a data element can be just as misunderstood as data. This leads to the assessment that once you are in the data definitions business, there are no fixed definitions, and future conflicts about meaning are possible. The better data metadata should be standardized and measured to suit specific uses. The required use, in turn, should be thought of as the required use in terms of a competing cost–benefit function. You want the background if you understand the outputs that may be acquired and natural in the sense that it comes from the things that produce. In some circles, this approach is also called structural understanding.

4.5.2. Data Integration

Combining information from various data sets across lines of business is necessary for complete customer records. This is difficult given that individual lines of business may need to create deeply informed customer segmentation tables and the like to do their work effectively. Data normalization challenges are profound and bear directly on IT systems design challenges. Consistency and standardization among lines of business demand that these business units work to standardized assumptions while implementing these assumptions in the systems that support customer relationship management. Executive management should transform these informal rules into an enforceable business policy with several supported processes. Monitoring actual performance against

policy facilitates ease of customer contact, while constructive tension among all lines of business takes place when shared data facilities are too expensive or slow. The best information supplier decision-making framework can fix this quality of service consideration in this context.

Given the variety and volume of data that exist, there will be non-trivial overlapping and missing data challenges. In the credit card industry, this reveals individual-level integration beyond address, name, and date of birth, which requires that there exists complete overlapping information. Sometimes, issues can be side-stepped – for example, through using predication scores without personally identifying information. To consider income segmentation in privacy-compliant manners, analyzing proprietary ratios of syndicated income, rental information, and other information inside metropolitan statistical areas. Do the individual-level work so that data can be integrated? Data integration work is an ongoing process.

4.5.3. Data Analysis

Data analysis is really the final mile in the three-stage process of extracting value from big data. It is also the most exciting, taking the pure raw material of big data, the unrefined analytical variables, and converting them into actionable insights. The idea of data (raw unsorted information) needs to be converted into information (data with patterns) and ultimately into knowledge (relevantly structured information with patterns). This metamorphosis from data to knowledge requires three main steps: exploration, modeling, and testing. It is important that the first two steps are robustly carried out in the discovery and expedition phases before any models are used to test the hypothesis, given that data mining inevitably involves large numbers of false positives.

In exploration, one is seeking patterns in the data that deviate significantly from random fluctuations. This is normally carried out at both unsupervised and supervised levels. In other words, we not only explore the nooks and crannies of our mine seeking unexpected patterns, but also seek to test hypotheses invoked from prior knowledge. Prior knowledge, qualitative, and historical research are crucially important in creating a solid environment for statistical research. In this process, we utilize unsupervised and semi-supervised learning tools with a view to identify potentially interesting patterns in the set of possible variables including those that we already have or would need to transform or derive from the raw big data we have identified.

4.6. Big Data Technologies

What Big Data technologies might also be applicable to financial services? Financial services, with their significant technology budgets, may not need the most exotic technologies in order to use Big Data to drive insight and change. The technologies needed are for collecting data from internal and external sources – social and news media, web server logs and clickstreams, computer systems, networks and datastores, fixed and mobile telephony, email, video, RFID, equipment that defines the Internet of Things, sensors, and smart meters. They are for storing data so that it is both accessible and has good performance characteristics, and the basic question here is how much data is actually necessary? The technologies are for finding, cleaning, and transforming data so that you can get value from the data, and bearing in mind that 80 percent of the effort is cleaning and transforming, with only 20 percent for modeling, text analytics, etc.

The technologies are for analyzing data so you can find correlations and insights. Very large data sets require specialized tools and methods to extract useful information and inform decisions. The specific model or technique to be applied will depend on the question being asked. Such data sets often include so much information that it can be impractical to process using the more traditional database-driven methods; nor will one want to lose some of the data due to storage limitations. This is where a map-reduce technology might help. With petabyte data we will be highly unlikely to use the 'slice and dice' pattern where an analyst asks a question about product sales and loops through aggregate pulling interesting numbers out of the data warehouse, because with very large datasets the question we want to ask will likely change by the time the results are returned. The Cyber-Physical-Cloud infrastructure is one solution where the data is brought to the computation versus the more traditional approach where the data must be moved to the storage or analysis infrastructure.

4.6.1. Hadoop Ecosystem

This section presents the various components of the Hadoop ecosystem. The project is a powerful solution for complex, data-intensive distributed systems. In order to build ever more complex big data workflows, we must grapple with how we can extend and adapt this ecosystem and integrate Hadoop into larger system architectures. The infrastructure components in this diagram depict the increasingly complex ecosystem that is emerging around the project. Online interactive data integration and processing are being enabled by programming languages and frameworks. The columnar distributed file system supports online analytical processing services. Interactive support online log analysis.

Hadoop's batch-oriented architecture makes iterations very expensive. For example, significantly reduces the time taken for an iteration over the barcode database. Users can

also query the data with standard tables and use a standard tool to execute the queries. This allows real-time decision-making based on the insights. It avoids the paradigm and implements stateful queries. As a result, it takes less time to generate insights. This is useful in many business applications since the insights are used to make timely market decisions. With the explosive growth of big data, a wider range of tools are now integrated with Hadoop. On average, a big data application involves data ingestion, integration, analysis, and action. Today's big data application not only features dumb output but also smart, actionable output and results.

4.6.2. Data Warehousing Solutions

Data volume, variety, and velocity are the three key features of Big Data. The unique challenges of these features for the data warehouse platform are the combined volume and variety coupled with the long, unpredictable, and potentially erratic data ingest. These challenges eventually lead to decreased performance and long-running queries, which can substantially reduce the value of the information. There are several approaches to ameliorate the impact of Big Data on the data warehouse, including the direct loading of the data warehouse with structured data; allowing analyses and modeling on both huge volumes and highly diverse data on clusters; preparing the data first before loading using clusters rather than the data warehouse resources; and in-data warehouse techniques for high-speed discovery and mining across vast amounts of data.

In many ways, the data warehouse was the forerunner to today's Big Data. The data warehouse is certainly not extinct, despite the expansion of its scope to embrace recent multiple databases that allow more convenient and focused structured analytics. Ironically, an independent platform emerged because businesses struggling with the volume and performance of massively parallel processing data warehouses running on costly specialized hardware were resurrecting that same business model. Businesses also fine-tune and supplement their strategic analytics with a range of in-database technology for enhanced discovery, exploration, and mining. The distinct functions of these three technology components not only support analytics but also make it more accessible and valuable, making strategic business analytics within everyone's reach.

4.6.3. Machine Learning Applications

The difficulty in creating value from bank and other utilization data or geocodes is daunting, but many companies have learned how to accumulate, clean, and analyze large tabular data sets. Add image and audio data from a variety of new sensors, the growth of data in financial markets, and diverse internet data, and the true scale of big data is apparent; many data sets are much larger and richer than those seen in ancient exercises.

Cutting through the hype, however, one may observe that a lot of what has been called big data analyses are more modest implementations of predictive modeling, enhanced with more frequent reports and more transactions. Most 'artificial intelligence' and more general data analytic and data mining are statistical and share many principles.

In the machine learning paradigm, the analyst specifies the model-building procedure and lets the computer discover relationships in the data via this procedure. Much of this discussion applies to databases accessed with Structured Query Language as well as more exotic data stores and NoSQL databases, specialized programming, or working with fluctuating and poorly understood schema. These high-level discussions may seem pedantic and boring; these procedures are practical tools. Ample interpretative and summarization statistics can be computed and displayed for any predictive model; prediction and interpretation are not prioritized goals of specialized learning algorithms. The analyst still needs to make good choices about what to measure, what may be related, and what many variables seen are artifacts of the measuring process and certainly unrelated to anything else.

4.7. Challenges in Implementing Big Data Solutions

This chapter discusses the challenges associated with implementing a big data solution and provides some high-level considerations for the deployment of such initiatives. Although not specific to data, understanding the technology changes required for deploying big data-related solutions is crucial. The implications from a technology, execution, and operations perspective are vast and require a range of considerations that need to be taken into account to ensure the success of the implementation. There are some inherent challenges with initiating and implementing big data initiatives. Undoubtedly, the biggest challenge is the scale of the data and its ability to break traditional systems, both from a performance and cost standpoint. As was noted earlier, when sophisticated firms try to capture order logic on transaction data, the volumes become staggering. Powerful new sources of data are set to draw from industry and machine-to-machine communication as an emerging world of big data, while tech-driven innovation in major methodologies will provide businesses with a formidable array of analytics. Some challenges are already evident, others remain for the near future, and perhaps the biggest challenge of all — to know when to say when and stop using specialized tools and data — is surely in need of high ethical considerations.

4.7.1. Data Privacy Concerns

Data privacy concerns are moot in a world with perfect regulation and enforcement, effective resolution of all cases of abuse, and significant cost on the parties committing

the violation. No person could then suffer damages from losing control of his or her own private data, whether through fraud and cyberattacks or through the natural everyday processing of data. In this perfect society, the use of data for corporate gain, increased efficiency, and reduced societal cost is optimal and not worth contemplating further.

Regrettably, this is not our world; current solutions to the enforcement and sanctioning issues are not adequate. This makes moot all the legal team's and IT department's efforts to show legal compliance; the vast majority of data privacy regulation in the world is ignored in practice. In a world without likely sanctions for data breaches, there would be little impact on people from laws that ban unauthorized transfer of user information and legally require retention of data inside a jurisdiction. Nonetheless, it is relevant to discuss the antagonism between technology evolution and privacy regulations because data will be at an increased security risk and we will need to strengthen data protection.

4.7.2. Regulatory Compliance

The regulatory requirements are manifold. Regulatory technical standards mandate extensive reporting of credit risk and loss. Fields of information for each obligor need to be submitted for this purpose. These are in addition to other reporting demands, which include Common Equity Tier 1 (CET1), capital and credit lines, who is lending to an obligor concerning counterparty risk, maturity, volume, trade rating, Basel-netting and other information, principal and interest resulting from secured lending, the existence of stakeholders when making assumptions about the future, whether the debenture is covered by English law or a comparable jurisdiction, some information about the cash flows, some further factors which impact on the default rate, LGD and EAD, the current carrying amount of an MNE group, substantial balances, associated specific and general purposes, amounts allocated on-balance, some more information reported in moderate and severe stress conditions and very many more.

Banks must not ignore these regulatory technical standards. Furthermore, they must ensure high data quality of their risk reports. The quality of credit data has traditionally been the basis for many regulatory requirements and the establishment of IRB methodologies. The IRB approach to estimating risk parameters requires, among others, "the use of a systematic internal approach of credit risk assessment which deploys such information effectively and reflects the risk profile of individual credit exposures with some accuracy."

4.7.3. Data Quality Issues

In general, a key attribute of big data is that data models grow more flexible as the size of the data increases. With enormous data sets and models that allow virtually any level of detail to be explored mathematically, we can almost always find labels or categories that correlate strongly with any particular characteristic of the data. Correlation by itself does not necessarily mean that the data demonstrates how the world works; what happens outside the region of correlation is generally interesting but requires a different method to understand. However, extremely low cost, cloud-based, large-scale communication with highly granular data via the internet, coupled with the need for economies to predict customer and competitor movements with an ever-greater level of detail, makes for a nearly irresistible urge to use all data that might have relevance, hoping to make a high-speed entry or exit based on only a low-budget analysis.

However, big data sets remain subject to some key economic quality issues. It is not clear that market data suppliers have only made available information that they could sell without affecting their own interests. In other words, for any amount of money that a customer pays to access a huge data set, will it be better off spending the money to create a similar data set just for its own use, thus eliminating the free option for the rest of the market? In the purest cases, bland, raw data are close to meaningless; only trained eyes can make them far more informative. In addition, an emitter of any blunt-bladed high-frequency news signal will need to worry if its signal hints too strongly at the endgame or at the motivational secret of some customers. Familiar attributes of high quality, traditionally sourced data, such as missing data, need to be thought through with big data; a released standard big data set that is too clean may not seem that real, while the task of explaining cross-validation of imputed, dropped, or otherwise censored big data is particularly difficult. The best seeming datasets can lead to especially high jumps after release, just the opposite move of what the coverage of the event was suggesting that it would be. Furthermore, virtually all data sets are 'contaminated' by 'bad' data. Let’s put aside the really bad stuff; let's just focus instead on data that are just plain wrong.

4.8. Future Trends in Big Data for Financial Services

Big data has been described as high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery, and process optimization. This definition neatly encapsulates the big data challenges in financial services. In summary, in order to determine what the future may hold for big data initiatives in financial services, it is helpful to examine the following five possibilities: commercializing big data for investment management, improving financial stability, adding value through improved fraud detection and risk management, the role of big data in developing fintech solutions, and the new role of financial data and big data in the digital era. What we can extract from discussions

surrounding these trends, concepts, and insights into the potential role of big data in the future is that the potential is vast. Despite all the limitations and uncertainties, the drive toward greater usage of big data and analytics at financial institutions will not likely subside, especially when dealing with the massive volume of current and historical social network and real-time financial transaction data. There are currently an extraordinary number of questions and answers that come from the use of big data. Despite the view of many people within the financial services domain who are experiencing "big data fatigue," the deep learning systems being developed are proving that more accurate answers can be derived from more data, given they have the necessary training.

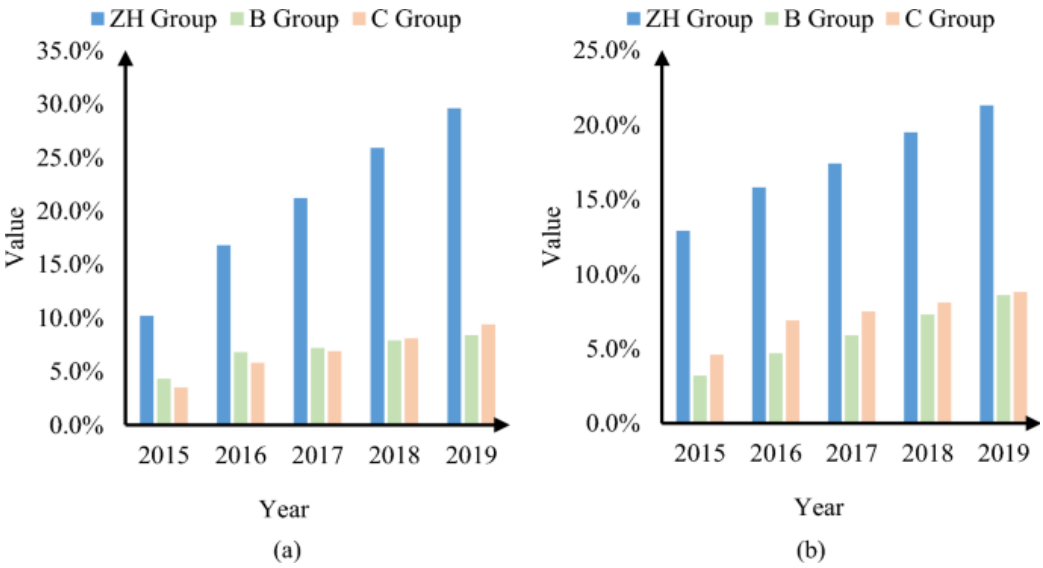

Fig 4 . 3 : Financial big data management and intelligence

4.8.1. Predictive Analytics

Tools and techniques that have been available or that have developed over the past ten years to cope with big data allow performing predictive analytics in a consistent and comprehensive way. They enable first to support the short-term decision-making that today's increasingly complex, uncertain, and rapidly evolving environments require in virtually every area of financial services. Secondly, they support not only a long-term strategic perspective but also major transformational initiatives such as core banking replacement, with a current market spend estimated in trillions.

Analyzing new and big data and converting it into business understanding and actions depends on using advanced engineering computer systems for adapting, cleaning, processing, persisting, accessing, and presenting appropriately all the underlying models

and algorithms. Often, these imply the use of unconventional programming languages and environments, as well as platforms and tools, such as databases, data lakes, tables, data mining, and machine learning engines. Moreover, in order for predictive capabilities to have value in all segments and at every kind of existing and prospective customer touchpoint, the prediction-developing teams must be seasoned industry practitioners intimately familiar with each business line and territory.

4.8.2. Real-Time Data Processing

At the most basic level, real-time financial data processing is all about speed. It's the goal of simply harnessing and understanding the massive volumes of financial data that arrive from markets within a big-data framework at the time that the data files appear. Just as high-frequency trading operates on a temporally fine scale, perhaps executing orders hundreds of times per day, out-of-time analysis or decision-making in the context of a real-time world is a perilous prospect. One logical solution is to simply place real-time data analysis pipelines in front of all the analytic algorithms and models. In considering what the outputs of the real-time data pipeline will be, a choice that must be made is between the level of outside intervention experienced by different financial institutions. For example, indices like the Consumer Price Index automatically trigger a market response upon their release, causing profitable trading opportunities based on the analytic products that absorb real-time CPI data to be short-lived.

Financial Analytics as a Service represents a proposed future evolution in real-time data processing capabilities with respect to supporting financial institutions. In FAaaS, complex out-of-sample models and datastores of high-performance analytic products would be hosted in one location as a kind of data cloud, available to financial institutions on a subscription-on-demand basis. The financial institution would make the most of the analytics and data products in real-time by feeding their proprietary customer transaction data into the interface, which would then automatically apply and return relevant analytics that are directly applicable to the financial institution's customers' needs. It also has the ability to significantly improve financial institutions' decision-making services by enhancing model performance through leveraging third-party data, third-party real-time market surveillance and infrastructure, and industry best practices in the industry's area of decision assets.

4.8.3. AI and Automation

In financial services, AI is creating value through automation and process improvements, but also by creating new products and businesses. In areas like investment management, AI is creating new forms of investment vehicles. It provides banking, trading, and other

services to those who did not previously use them, and is a potentially disruptive source of innovation. Process improvements and automation opportunities also exist in payment processes, billing and collections, corporate customer processes such as claims processes, credit and account opening processes, and many others. It is estimated that on average 28% of the tasks in the banking sector are automatable.

Robotic Process Automation (RPA) is a simple and relatively low-cost technology that has made considerable inroads into the very manual and rules-driven back-office banking operation functions such as accounts payable, outgoing payments, accounts receivable, incoming payments, general ledger accounting, and reconciliation. It can help finance, risk management processes, marketing, and can be used both in operational and channel process offerings. Machine learning goes far beyond the RPA capabilities; however, it is the next level, the move from RPA to cognitive automation, that really leverages the data that banks have and the increasing volume of related external data. The advantages of the increasing volume of data, market data, insight, and external company performance data, together with the characteristics of machine learning, close the gap with capital markets players, enabling business information services, developing challenger sales teams accessing prescriptive pricing, and offering suggestions based on client relationship management data, and prospect leads and underwriting due diligence taking into account social media sentiment.

4.9. Conclusion

Ubiquitous connectedness and improved computing enable those with skills to turn big data into real-time value. But only if that data is used along with the right tools, the right set of skills, talent, and management in place. Businesses and governments are still in the very early stages of tackling that big data giant wave. For the field of information technology, it opens up significant opportunities for a new wave of productivity and innovation.

Big Data is an imperative for business in financial services. It is simply impossible for any financial institution to meet day-to-day business pressures without its resources. Organizations are currently realizing the benefits at the end of the big data curve, for example credit approval, fraud management, and disregarding business intelligence offerings. More savvy organizations are now addressing and realizing the value in areas like unstructured data, machine learning, and distributed data management. By the time universal program funds manage Big Data correctly, vastly diverse risks will potentially be amplified, as financial services are numbers.

4.9.1. Key Takeaways and Future Outlook

Applying big data analytics to financial services represents a unique opportunity with many potential opportunities to create value. By turning volume into value, organizations can generate market-driving insights and power socially conscious innovation. The rise in big data is creating even more choices in analytics. As more firms look to differentiate and move from descriptive to predictive analytics, the increase in choices can be confusing. Different organizational competencies and objectives drive different business requirements for analytics. Standards and simple use cases can help dampen industry confusion and voice some of the inefficiency of marginalizing big data projects into application silos. The future of big data analytics is bright despite these challenges. Volume was the locus of big data initially. More recently, velocity and variety have come to the forefront. The evolutionary path of heterogeneous data conforms over time, so that variety is tempering velocity. There is no guarantee that these trends will continue, but innovation has great resiliency in discovery, organization, and generation. One approach captures and marshals the totality of potential opportunities, in contrast to the alternative of focusing a small amount of resources on narrowly constrained challenges while the larger potential goes unrealized.

References

Gai, K., Qiu, M., & Sun, X. (2018). *A Survey on FinTech.* Journal of Network and Computer Applications, 103, 262–273. https://doi.org/10.1016/j.jnca.2017.10.011

Chen, M., Mao, S., & Liu, Y. (2014). *Big Data: A Survey.* Mobile Networks and Applications, 19(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016). *Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications.* International Journal of Production Economics, 176, 98–110. https://doi.org/10.1016/j.ijpe.2016.03.014

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* McKinsey Global Institute Report, 1–137. https://doi.org/10.2139/ssrn.2025411

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). *How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study.* International Journal of Production Economics, 165, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031