Deep Science Publishing
https://doi.org/10.70593/978-93-49910-78-2_5

Chapter 5: Harnessing big data analytics for informed decision-making in financial markets

5.1. Introduction

We live in the Digital Age, marked by unparalleled volume and variety of data production, popularly referred to as "Big Data". Financial services and markets have long relied on data-driven decision-making, and have witnessed and contributed to the current Big Data Revolution. Big Data Analytics offer both challenges and opportunities to practitioners, regulators and researchers. Great amounts of diverse data now regularly and increasingly become available, made possible by technological advances in telecommunications and computation, as well as the available digital infrastructure. Financial data are not only produced by traditional financial exchanges, but also generated by new technology platforms, where businesses and customers interact dynamically. In an era of information overload, fueled by the content and marketing strategies adopted by traditional financial news providers and popular social media, how can financial professionals cheapen the costs of decision-making while improving accuracy and quality? By using innovative models, algorithms and methodologies, that are able to reduce the dimensionality of the information space and extract useful signals from the noise (Gupta & Gupta, 2017; Khan et al., 2018; Lee & Kim, 2019). This is a highly challenging task, which requires collaboration between interdisciplinary teams involving professionals from different backgrounds, including finance, computer science and engineering, mathematics and statistics, economics and social sciences. Policy-makers and regulators should also be part of these teams, particularly in ensuring a fair playing field for industry players (Sahu et al., 2015; Sharma & Chhillar, 2016).

5.2. Overview of Big Data Analytics

The use of Big Data Analytics has the power to transform the systematic decision-making processes in financial markets. Big data is a broadly used term that encompasses a large amount of collections that can be in either structured or unstructured format. Modern technologies now allow organizations to smartly analyze, interpret, and visualize these massive data sets as a practice to make better decisions. Big Data Analytics is defined as a combination of different approaches developed throughout the years, comprising computer science and statistics. In a nutshell, Big Data Analytics is the area in which Big Data meets machine learning, data mining, and parallel computing. This topic has gained increasing interest from practitioners and researchers over the last two decades for its promising capabilities.

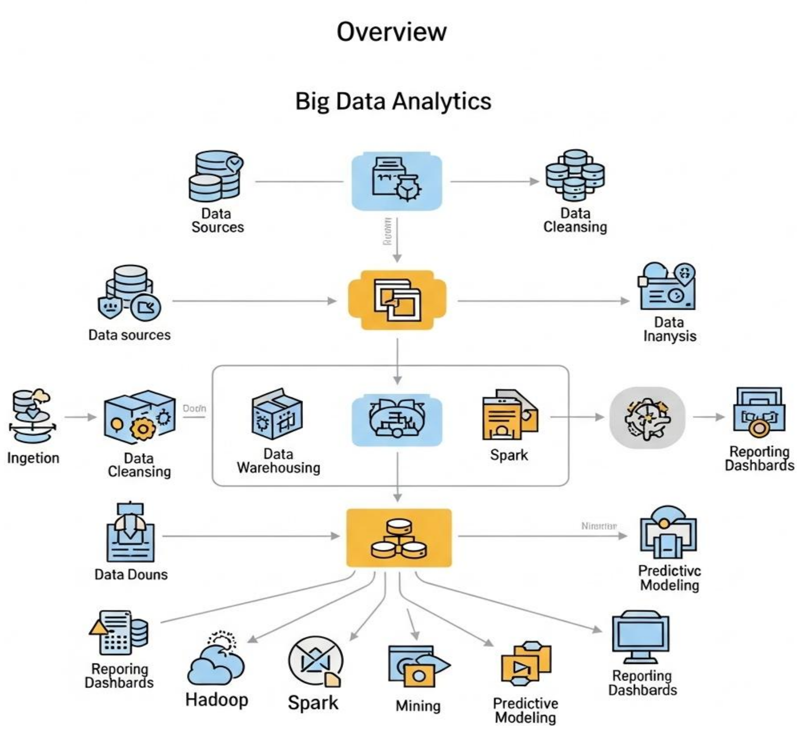

Fig 1 : The various stages of the process, from data ingestion and storage to analysis and visualization, with icons representing each step.

Different definitions of Big Data have been proposed throughout the years. The original definition appeared in 2001 and defined it using the 3V's framework: volume, velocity, and variety. Volume describes the large amount of data that increases exponentially. Velocity refers to the increasing speed of data generation and sharing. Variety means the heterogeneous formats of the data, which can now be collected from different sources, such as mobile phones, social networks, transactional databases, or the web. Later, 3 additional V's were added to the original definition: veracity, value, and variability.

Veracity refers to the uncertainty of the data and the lack of trust that one might perceive in the data sources. Value means that for being considered Big Data, data has to be valuable for organizations to use in their decision-making process. Finally, Variability means that it is common for data from different sources to be inconsistent, therefore needing cleaning before analysis.

5.2.1. Definition of Big Data

Big data primarily refer to data with sufficiently great size or complexity that the normal data handling and management tools are incapable of dealing with. Yet, size, complexity, and the difficulties imposed on their management and handling are not the only distinguishing criteria of big data, nor do all of them need to be satisfied for some data to be considered big data. First, any data that warrant special and sophisticated treatment to derive credible insight from it can be called big data, whether it is of a gigantic scale or not. Second, big data are characterized by the conditions set forth in the popular 3V model, namely, volume, velocity, and variety. Volume of data refers to the size of the data privileged in big data over normal. Velocity refers to the speed at which data are exploding, or being generated, constituted, created, or being fed into storage repositories. Variety refers to the heterogeneous nature of big data as they come in different forms, varying in structure, ranging from structured, semi-structured, to unstructured, multimedia, machine-generated, social, and so on. However, new additions to the 3V model continually emerge with time. For example, two or more additional V's have been added to the original 3 V's of big data. The fourth V is veracity, which refers to the reliability, quality, and authenticity of big data. Until the arrival of big data and their emergence during the resurrection of AI brought special consideration to the authenticity aspect of data, it was already known that, conceptually, data cannot be any more veridical, more natural, or more genuine than behavior and its propelling processes themselves.

5.2.2. Characteristics of Big Data

Big Data analytics entails the processes of analyzing data sets characterized by massive volume, heterogeneity, of time-varying nature, and enriched with very small number of quality samples capable of encoding large amount of information to garner valuable business insights to support strategic, tactical, and operational decisions. During analysis, the underlying principles need to account for frequent curve reversals, existence of noisy weeks, outliers, and the presence of multiple underlying models and patterns in the data at play. Big Data sources proliferate from diverse verticals such as travel, banking, electronics, telecom, healthcare, and financial trading, yielding unstructured voluminous datasets that need to be prepared and processed into tidy-shapes to ensure they serve the incoming predictive analytics needs of enterprises. The role of time is prominent in Big Data analytics. Organizations need to be future-ready to efficiently and if possible, economically, respond to changes in the world, contingent on the time-varying nature of the world. The design and development of predictive models need to remain flexible enough to periodically retrain them with more recent data to thwart predictor incursions which happen regularly and with disturbing frequency in many database domains. The answer to the question of 'How much data needs to be analyzed?' is predominantly data-driven as opposed to heuristic-based. The tendency of Big Data analytics is to find answers to this and other related questions by analyzing all available data for patterns and relationships. This usually entails analyzing bulk-sets of tidy-datasets with scrumptious sizes that cannot be analyzed regardless of whether they are parallelized or distributed on anything less than Big Data clusters.

5.2.3. Importance in Financial Markets

Data is one of the mainstays in the financial market. Specific data closure, collection, maintenance, management, and reporting are the signs of a developed financial market. The quality and quantity of data collection and reporting mainly depend on the country and its relative development. Accurate market information enhances the probability of efficient market pricing and better inter-temporal allocation of resources. Most transactions in the economy are conducted via the financial market and the principle of ultimate wealth management drives the whole economy cycle. In the context of the price mechanism, wealth being noted in the nation's currency drives the economy. Wealth is accumulated in transfer of assets mentioned in capital market transactions. The data are collected, processed, and reported by some specialized organizations both in the private and public sectors.

The mechanism of data production is intricately connected with the national business cycle, thus presenting difficulty in processing and interpreting the data. Financial markets are considered to be efficient because prices represent valuations of future cash flows. Prices provide information signals; therefore for an efficient market and economy data must be accurate and timely available. Financial market development entails novel innovations in terms of various financial instruments, up-to-date technology advancement, mobile units to provide market and price information, and innovation in risk measurement techniques. Central banks, market regulators, and development banks today face challenges to satisfy the market demand for transparency mandated by investors and multilateral funders. Thus good-quality data are used in the diverse sectors of risk analysis, market surveillance, and modeling, and as an input for responding to the transparency initiatives driven by the international organizations.

5.3. Data Sources in Financial Markets

Informed decision-making about the financial markets requires the careful consideration of what data sources market participants rely on to better understand, predict and ultimately explain price movements of financial assets. The primary source of data that financial analysts rely on is market data, but over the last few decades, several alternative data sources have emerged, which also have a direct impact on informed decision-making about financial markets and the price discovery of financial assets. These data sources can capture the sentiment, opinion and expectations of market participants about financial markets, and analysts are faced with the challenge of determining which data sources are appropriate for their respective analyses. In this section, we summarize the main creating and modifying data sources employed in financial analysis.

Market data is defined as the set of trading records for the given financial asset during a specific sample period. It can be specified at different frequency levels, including daily, weekly and at the highest frequency, tick-by-tick databases. Typical market data records include the opening, closing, high and low prices, as well as the volume traded and the formation timestamps. In the case of financial instruments that are traded by limit order books such as futures or foreign exchange spot transactions, market data will also provide a picture of the partial or full order book during a specific sampling period, as well as the depth and width of the book. Other variables such as bid-ask spreads may also be derived from market data.

5.3.1. Market Data

Financial markets produce data for trading activities. Data on market structure, liquidity, trading volume, transaction costs, bid-ask spreads, and price changes in the limit order book are available on different time scales, from millisecond up to hours and days. Market data reflects the myriad expectations of market participants over the future prices of the assets that trade in these markets. Individual trading decisions are based on information sets that differ across participants, including differences in available information and in processing of the same information. At high frequency — the sub-second or sub-minute timescales — such forecasting becomes very difficult, particularly since market microstructure effects, including liquidity supply and demand, dominate changes in asset prices on those short timescales. More generally, price changes on all frequencies contain market microstructure information directly tied to how trading mechanisms funnel information into price changes.

Price data refer to changes in the market prices for shares and assets over time. In general practice this data is expressed as 'high-low-close' quotes for time intervals ranging from timeframes of a second to a day or longer. Many periods in a day are common practices for shares and assets located in very liquid and active exchanges. In all cases price data consists of time-stamped information indicating the price at which the transaction occurred. This information usually includes the price at which the trade was executed, the time the position was established, the quantity of shares transacted, bid-ask spreads, often the bid price prior to the trade, and whether the buyer is absorbing stock or supplying it. Other traditional price data are the bid-ask quotes existing before the transaction occurred.

5.3.2. Alternative Data

Alternative data refers to an extensive array of new datasets that can be harnessed for the purpose of executive production systems and causal learning to enhance business decisions. Due to the recent proliferation of new data sources and the continued growth of modern analytics and business intelligence environments, alternative data is often used interchangeably with this broader term. The alternative data landscape is varied by nature, containing vast sets of different data types that can be utilized to answer many kinds of different questions – from search visibility analysis to bankruptcy prediction.

The term "alternative data" generally refers to datasets that are external to the company of interest. Alternative data opportunities can be stratified based on origin, characteristic, price, and readiness. Considering the supply of alternative data while frugal innovation is key, the most effective and low cost solutions often come from your own operational data, often leveraged with readily available existing information and common big data techniques and tools that are used to exploit larger alternative datasets.

5.3.3. Social Media Data

Consolidating an extensive content base in itself and spontaneously updated by their users, social networks flourish in public conversations about commodity prices, investment strategies, and other information relevant to financial markets. Using social media means having news in real time; constructed by individuals, who openly express their feelings and opinions, frustrations, and expectations towards companies, economic sectors, and the economy itself. The growing availability of institutional accounts on those platforms and discussion groups that attract a large number of users with similar interests portray a movement to socialize investment decision-making. In this information environment, knowing when and how to track users' perceptions can make all the difference.

Establishing connections amidst the volume of publicly disclosed information entails the risk of analyzing rumors. To limit the potential negative impacts of misinformation, social media analysts use algorithms to identify fake accounts, bots, trolls, or fake news. It has been found that filtering posts, comments, and tweets by keywords has the potential to build more accurate models and improve the forecasting results. Testing the accuracy of supervised machine learning models has shown that certain algorithms, using probabilistic words, can present the best results in predicting the fake news classification task applied to the financial market domain.

Real or false, the fact is that digital platforms are becoming more and more part of the forecasting models for price behavior and trading strategy decision support tools. The search process for sentiment information goes beyond simply using the number of comments or the number of likes a post receives. In addition to metrics of how widely a message has been transmitted, incorporating behavioral verbal compositions using predictive coding methods to define decision-oriented categories can bring additional insights to social media tracking. Moreover, discussing how defining the learning universe and sample selection can impact study conclusions is essential.

5.3.4. Economic Indicators

In financial market analysis, several indicators measure essential economic attributes. These include gross domestic product, unemployment and employment rates, interest rate, inflation and deflation or Consumer Price Index and Producer Price Index, consumer confidence index, stock and real estate market indices, and balance of payment. These ultimate indicators help in forecasting investment cycles, business decision, pricing their products, and customer consumption. Market sentiment is an important attribute in finance and data on indicators such as the unemployment rate and consumer confidence index can raise or decline sentiment.

The most important information is derived from macroeconomic announcement. Macro announcement forecasting is critical for optimizing the investment atmosphere. The relation between returns and inflation or interest rate have been widely studied. Inflation rate is powerful determinant of equity returns at long-run. Long-term evidence specification shows that the expected inflation, but not current inflation, estimate traditional asset pricing relations. For short-term specification, expected inflation is unimportant for interest rate and cannot be combined while short-run specification is not rejected.

Studies found that inflation, GDP growth and real yield positively affect equities. Low real bond yield explains longer trend of equity index or a stock's returns. So low bond yield attracts more equities, which push equity prices. In an asset pricing context, an interest factor may supply risk premium rather than abatement cost of equities. Model assumes that rationality investor would equate return on equities and long-term yield. Thus, stock price level reacts negatively to long-term yield. Increase in interest rate may lower elasticities of asset pricing relation due to aversion for interest rate risk. Ratio of Corporate profits after tax to non-farm business sector to equity price must follow a declining rate over a long term.

5.4. Data Processing Techniques

Despite the availability of efficient data storage and extraction systems, data collected on the web requires preprocessing, competing for the most arduous task in the big data streams research area. Data preprocessing involves multiple preparatory steps like cleaning, transforming, and integrating the raw data before applying any type of data analysis or machine learning algorithms and generating useful insights. This chapter discusses the main data processing techniques that are often applied to data collected for financial analytics. The process is presented as a pipeline, outlining the basic building blocks required to extract value from the data.

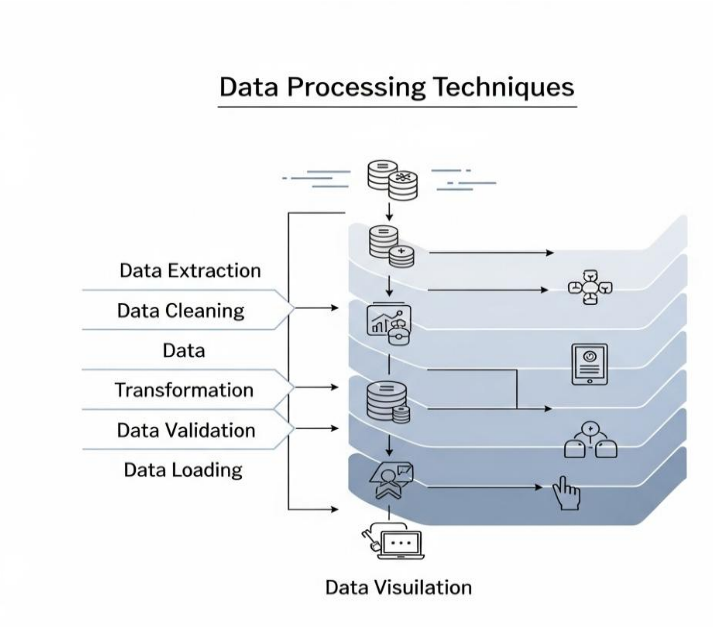

Fig 2 : The data in motion, demonstrating the transitions between each phase of the data processing techniques

Data cleaning refers to the detection and correction of errors found in the data. With a low quality in the extracted data, it becomes imperative to pay careful attention to this process. Several data cleaning techniques exist, and their selection depends on the type of data and on the application. Methods include removal of duplicates, removal of missing or outlier values, and suppression of inconsistencies or meaningless data. For example, stock market data have potential outliers that may potentially never occur during an investment period. Therefore, to remove noise or sharp movements or shakeouts from the stock prices, investors can apply divergence thresholds. In other domains, machine learning classifiers can help to identify the instances with noisy label. Other methods look for rules describing the data that are violated by some instances, indicating that those instances require data cleaning.

Data transformation refers to the process of converting the original data into a form suitable for descriptive or predictive analysis. Important to ensure is that the semantics of the data are not altered. It involves several functions that are applied sequentially to achieve the desired data state. The first function typically alters the structure of the original data presented. For time series data such as near-real-time streaming social data used in financial applications, the most common transformation, often called data representation, is done on the interval length, necessitating aggregating the original data over a given time objective. For integrated financial machine learning applications, another important transformation may be the time zone, especially important when working with geographically diversified equities.

5.4.1. Data Cleaning

Data cleaning refers to the process of identifying and addressing corrupt or inaccurate records from a dataset. It is one of the most significant and demanding steps in data preprocessing. The quality of a dataset is defined by the quality of data present in the dataset. Data cleaning is an important and time consuming as well as costly task to create high quality data. The badly structured data creates a hindrance for policy makers, business managers and decision makers to take poor decisions. For instance, consider a domain where there is high volume of data present but high error rates as compared to some other domain where the volume is less having low error rates. Data in the second domain is more useful as it provides high quality data.

The errors in data can result due to different reasons such as generation of data and gathering of data. The errors may mislead the data analysis and model building process and as a result adversely affect the results obtained. Therefore, it is important to remove these errors of interest from data. Data cleansing is not just about removing errors but also about correcting the values of corrupted data. There are many types of errors that present themselves such as fictitious attributes which have values from never seen before distributions. Other types of errors called faulty attributes can contain values that seem to be valid but are impossible in general. For example, an email address cannot contain a space. There are other types of errors called unusual attributes which have values that are really unlikely. For example, a person with age of 300 cannot be considered valid. There are other examples also such as order finalization date occurs before the order insert date. These all above stated issues need data cleaning in order to provide a clean data modeled for efficient decision making.

5.4.2. Data Transformation

After the process of data cleaning, the next course of action concerned with data pre-processing is data transformation. Data transformation is done for two main reasons. First, it helps in transforming data into a format appropriate for the data mining procedure or technique used later. Second, it also reduces redundancy within the dataset. Multiple data transformation operations fall under the following categories:

• Data Aggregation: Data aggregation is the process of accumulating data into a summarized format. This is usually the first step in data aggregation. Higher dimensional data is summarized into dimensional data. Both the detailed and summarized data may be stored for any purpose.

• Data Normalization: The process of data normalization is carried out to put different types of data on a similar scale. Factors like gender and accounts can be transformed into numbers as well as designated scales to do justice with the influence of each on the outcome. Normalization helps speed up the solution while using neural networks by eliminating unnecessary weights.

• Data Discretization: Data discretization is the process of converting continuous or whole data into a categorical format. This is usually done to convert data for methods requiring categorical data input. Data discretization can be done at various levels. Methods include binning and histogram analysis. Categorical data can also be converted back into whole or continuous data for analysis.

• Data Reduction: Data reduction is reducing data to a smaller dataset without losing the data's integrity. Data reduction may not be necessary for all data mining operations. Data reduction is required when working with operations on large datasets requiring a huge amount of computing time. Reduction, in this case, can be achieved either through sampling, attribute subset selection, and data cube aggregation, among others.

5.4.3. Data Integration

The term data integration describes the process of combining data residing in different sources from their different formats, into a single coherent dataset. Data integration aims to provide a unified view of the data. It consists of a variety of processes, which allow data sharing across the local databases that make up a distributed database architecture. Data integration systems typically aggregate data from heterogeneous data sources and unify such data in a format that is reusable by different applications in an organization. These systems then allow different types of applications to access and share data among the various data source types and locations. Many such integration tools offer users the ability to perform a range of integration activities, including: Extraction, transformation, and loading, Evolving agile integration applications, Federating multiple data sources and types, Synchronizing data across multiple heterogeneous sources, performing cross-organizational data sharing, and Lucratively monetizing data. Data integration is one of the most complex and expensive aspects of data warehousing. Integration is essential because data are often scattered throughout an organization. When data integration works well, a company can take actions like "Search for all customers that have purchased a digital camera in the last month and send them a $15-off coupon for a memory card to drive repeat purchases." Without integration, however, these actions may take an army of engineers working for years to complete. Because of the costs associated with data integration, most large companies have budgets on the order of tens to hundreds of millions of dollars a year to build and support data integration processes and systems. Typical activities that incur such costs include extracting data from operational systems, transforming data to comply with data warehouse standards, and loading data into the data warehouse.

5.5. Analytical Techniques

The term analytical techniques can be broadly categorized into three categories, namely, Descriptive, Predictive and Prescriptive analytics. In this section, we shed light on these categories to accentuate the specific features of big data analytics. These categories will be further discussed in detail in the subsequent sub-sections.

5.5.1. Descriptive Analytics The fundamental aspect of any form of data analytics is to summarize the available data so that one can at least gain a basic understanding of the data under consideration. Descriptive analytics focuses upon answering the "what happened" question. It does not employ any statistical or econometric model for drawing any conclusions from the data or for validation and falsification of specified hypotheses. It could also be referred to as a data mining exercise where no taxonomic methods are deemed necessary. The principle techniques commonly adopted under the ambit of descriptive analytics include visualization tools, descriptive statistics, clustering, and cross-sectional analysis. With technological advancements and the availability of modern computing systems at modest prices, visualization tools have become far more effective than when these were first invented. The use of computer software tools for data analytics has proliferated and the new generation data workers have also become proficient in applying such tools with far more operational capabilities. To summarize, descriptive analytics is the use of several visualization tools and descriptive statistics to generate a concise summary of a set of data, providing the analysts with a broad perspective of the patterns embedded in the data set for gaining some preliminary insights.

5.5.2. Predictive Analytics Organizations would be happier if they could find out beforehand what is likely to happen in the future using the historical data at their disposal. This "forecasting" feature of number-crunching has been around for a long time, especially in the business-related domains. Making predictions, estimates, or forecasts of the future events or values of one or more variables of interest is what predictive analytics is all about. Predictive analytics answers the "what might happen?" question and it does so by creating one or more predictive models for the variable or variables whose future outcomes are being predicted. Predictive analytics employs historical data to train the parameters of one or more appropriate predictive models. These models are used to compute to-be-predicted values for the variable of interest for a subset of the data that corresponds to a future time period.

5.5.3. Prescriptive Analytics Finally, organizations may be interested in making inferences on the past, want to assess the future, or might be preoccupied with examining "what should be done?" that would help to optimize a specific outcome variable of interest. It is the function of prescriptive analytics to answer the "what must be done?" question by employing prescriptive or optimization models. Prescriptive analytics comprises analytically based prescriptive methods that address the portfolio management issues facing organizations. Such prescriptive analytics would also include the optimization of resource allocations and key design parameters to meet a business service objective and the analysis of the uncertainty or risk associated with services provided.

5.5.1. Descriptive Analytics

Descriptive analytics help us understand historical patterns and the characteristics of observations. It presents the analyzed historical data in relation to a particular problem by identifying time period patterns, identifying the trends in relation to time, identifying similarities, correlations or associations, and by distributing the data along one or two characteristics to reveal the details. Many descriptive analytics visualizations are also used in various contexts. Some of the visualizations used in descriptive analytics include pie charts, area charts, line charts, bar/column charts, scatter plots, choropleths, heat maps and treemaps. Descriptive analytics is not concerned with how the analytics model was created or what algorithms were used to create it. It looks at the analytics created by the other types of analytical techniques in relation to a specific problem.

Descriptive analytics is used extensively in marketing analytics for creating different reports. Some of these include market measurement reports, marketing trend reports, marketing comparison reports, marketing performance reports, market segment profiles, product party profile reports, campaign response profiles, email campaign reports, and loyalty program reports. In finance, it is used to measure financial performance and different compliance reports. The descriptive analytics market is estimated to be significant by 2025. The majority of the work done in descriptive analytics is visualization coding. Many analytics applications provide granular APIs to allow for custom visualizations. The analytics created in descriptive analytics is also commonly found in business dashboards, web dashboards, executive dashboards, and reporting tools.

5.5.2. Predictive Analytics

The accelerating development of new analytical techniques has turned the spotlight on predictive analytics that exploit Big Data technologies to predict actual future events more accurately. Predictive analytics is a system of predictive forecasting tools that helps leaders make better business decisions, saves organizations time and money, and helps eliminate issues such as inefficient marketing campaigns, unproductive inventory, and needless employee turnover. It has the capability of efficiently and accurately forecasting human behavior, trends, and landscapes and generating useful predictive information that can assist business organizations in increasing sales. To understand the fundamental idea of predictive analytics, it is first necessary to understand the foundation upon which predictive analytics is built. Predictive modeling is defined as the process of estimating the relationships among a set of variables, using historical data with the aim of predicting the value of target variables in new data. Essentially, predictive modeling takes known input variables or factors and develops a mathematical relationship and algorithm that relate these factors to known output variables, thereby providing a statistical foundation to build predictive analytics tools.

Predictive analytics utilizes historical data as its foundation. The focus of predictive modeling techniques is the creation of algorithms that can be applied to future, unknown data within the same domain as the historical data used to develop the algorithm. Publicly available data, collected over the past half a century and more, can be employed to develop predictive algorithms that relate user features to advertising response, and to additional domains, including keyword performance, advertising pricing, and search engine traffic. Predictive algorithms can be created for focus groups, consumers, and other entities, for all demographics, geographic, and firmographic areas.

5.5.3. Prescriptive Analytics

Prescriptive analytics goes beyond predicting future trends to provide companies with actionable recommendations that help achieve desired outcomes. For example, prescriptive analytic software not only can predict how much product will sell in each region next week but also can analyze all the variables that can influence positive or negative customer purchase behavior and recommend the ideal prices for each product in each region to drive a desired growth in sales. Many companies today are experimenting with big data prescriptive analytics to boost sales, improve quality, enhance customer service, and reduce operating costs. Companies are also using big data prescriptive analytics to make better decisions.

Using prescriptive analytics optimizes marketing offers in real time; determines the best way to align soldiers for a military operation; designs the best airplane routing and schedule for maximum passenger convenience and revenue; and controls the flow through a manufacturing plant to maximize revenues and minimize order completion times. Major companies are exploring how to build prescriptive capabilities into their decision-making processes.

However, it is uncertain whether the industry's essential underlying methodologies are sufficiently mature to achieve such bold claims. It will take some years for the grand, futuristic vision of sophisticated, yet easy-to-use prescriptive solutions to be realized, with suppliers and users collaborating to respond to the needs that are revealed in practical applications. In straight use cases, sufficiency is less a question, as prescriptive offerings at their essence are generally systematic extensions of predictive methods. But only certain types of applications in their current form, or combinations, realize the full power of prescriptives. Uncertainty still reigns over the degree and form that prescriptive capabilities can take. Cloud computing helps the accessibility problem because it encourages customers to adopt new technology that otherwise would require long and complex implementation and testing.

5.6. Machine Learning and AI in Financial Analytics

Artificial intelligence (AI) is a powerful new capability that is being quickly adopted in financial analytics. Applications of AI engage with numerous research and practical problems in finance, including forecasting expected returns, testing market efficiency, predicting financial crises, portfolio optimization, predicting stock price changes, financial performance prediction, credit scoring and underwriting, event study analysis, big data sentiment analysis, general news impact studies, enterprise credit risk prediction, earnings announcement forecasting, hedge fund management, insider trading detection, option pricing, risk management, algorithmic trading, peer-to-peer lending, private equity, robo-advising, mutual fund performance evaluation, and fraud detection. While these applications are of great interest, the academic research literature employing machine learning methods to analyze business data has only recently begun to emerge.

Machine learning machines never really stop learning. Instead, they allow the user to periodically shuffle the input data stream and adjust the parameters governing predictions to reflect the changing target. In addition, supervised and unsupervised machine learning approaches are now both mature and documented in the literature. In supervised prediction problems, past data with the predictor variables as well as corresponding values of a dependent (response) variable are available. Classical supervised machine learning approaches include decision trees, random forests, support vector machines, neural networks, naive Bayes classifiers, and regression analysis. In exploratory problems, the user seeks to segmented observations without prior knowledge of group affiliation. In unsupervised learning, only specific observations are present. Popular algorithms include k-means clustering, hierarchical clustering, Gaussian mixture models, self-organizing maps, t-distributed stochastic neighbor embedding, and hidden Markov models.

5.6.1. Supervised Learning

Supervised learning techniques are used to construct a predictive model describing the relation between the dependent variable, which is by definition the target whose behavior we want to predict, and some independent or explanatory variables that are correlated with the target. This predictive model is then used to generate predictions of the target when we know its value. The information generating accurate predictions for future, unseen samples from the same general conditions requires estimating the model well from historical data. Supervised learning methods use labelled datasets provided by the user to estimate the target's behavior. The labels are the previously known values taken by the target for particular combinations of the explanatory variables over a historical period.

The labelled datasets used to estimate the model parameters contain a finite number of past observations, each described by a set of features. These observations are then used to learn the parameters of a predictive model. In the simple case of linear regression, for instance, the prediction for an observation is determined as: which is a linear combination of the feature values stored in vector, plus an additional term called the intercept and set to the value of the target when all feature values are equal to zero. The parameters to be estimated in this case are thus the coefficients which correspond to the terms of the linear combination, after the initial condition has been used to calibrate the intercept.

5.6.2. Unsupervised Learning

The methods above cannot answer all questions in a financial context. Typically, in financial datasets, it is very expensive or terribly complicated to label every single data point. When there are not enough responses, the predictions made by supervised methods are of limited use. For example, we observe some new data from the market and we would like to know if it belongs to the cluster of new stocks or stocks of a known cluster. This task is the classical classification problem that is solved by supervised learning techniques. However, we do not have the labels/classes for these stocks: To make good use of supervised learning techniques, we should have sufficient labels/classes available. Unsupervised learning attempts to build a model from an unlabeled dataset, that is, the purpose is to identify patterns or structures within the data. It relies on the fact that the unlabeled dataset is a realization of some underlying probability distribution. Models based on this assumption include generative models, statistical frameworks, graphical models, mixture models, hidden Markov models, deep belief networks, and autoregressive models.

However, many applications require further guidance than just the identification of simple probable structures. This is usually the case for most real-world applications of finance. For instance, research on customer segmentation, portfolio selection, anomaly detection, questions on market efficiency, optimum asset allocation on contexts of risk, mainly by stress-testing under extreme risk or by quantilizing based on quantile regression, covariance estimations and correlation, or those attempting to study nonlinear relationships for many variables, have proposed and done their implementations based on several techniques for unsupervised learning.

5.6.3. Reinforcement Learning

In problems of multi-stage decision making, known as Markov Decision Processes, Reinforcement Learning is used to derive indexed rule-based policies, defining an action to be taken in each stage, that ensure the maximization of the expected cumulative payoff. These multi-stage decisions are widely present in the financial area. For example, in portfolio selection models, a multi-stage series of portfolios must be defined over a certain investment horizon. After finding the values of the action in each stage, it is assumed that the investor will follow this policy. Interested in maximizing the investment's expected cumulative return, within limitations on risk associated with the investment, the financial decision maker following this pre-established rule takes action at staged intervals, periodically reallocating and rebalancing the assets in the portfolio. Indeed, portfolios and other compound decisions should be made on a dynamic way, allowing for changes in the environment and asset characteristics.

With a history in computer science that dates back to the thrill of the Artificial Intelligence Winter, the past two decades have seen a resurgence in Reinforcement Learning, establishing it as a well-grounded and studied area of machine learning. While initially, tables listing values for state-action pairs along with other rudimentary components were used, Deep Reinforcement Learning has garnered a lot of excitement by taking advantage of Deep Learning's ability to act as a flexible, approximate function approximator within Reinforcement Learning's policy and value-based algorithmic core. The result is a class of algorithms that combines the best of both worlds, acting like a second high-level layer that directs Deep Learning towards optimizing its stepwise measurement of success.

5.7. Risk Management Using Big Data

Many of the technologies and threats have existed for some time. What is different today is the scale, the reach, and the players. The Internet has become the critical communication medium, linking not only the obvious sources: companies, investors, governments, and terrorists, but also connecting things: locational devices, automobiles, power plants, and electronic controls. Non-state actors: criminals, terrorists, and hacktivists now leverage the advent of cyberspace to disrupt the deadlock, and to disempower institutions that once were powerful and omnipotent. These changes have greatly broadened the attack space. Nation states, which have traditionally relied upon traditional forms of power projection: tanks, airplanes, and ships, are beginning to realize that cyberspace holds the potential to easily disrupt their adversaries' critical infrastructure, more cheaply, and with lesser risks. Traditional state-based threats need to be counterposed against non-state actors who have found cyberspace to be the ultimate weapon of choice for asymmetric warfare. Digital stores of value, such as cryptocurrencies, and digital payment transfers offer innovators and criminals the ability to transfer money clandestinely. Internet-based teleportation mechanisms enable near-real-time movement not only of money, but also of shipping documents, negotiable instruments, bearer bonds, and private keys. These can cause massive disruptions across the globe. Risk management is an ancient art, but it is at the intersection of science and moral choice. To preserve social values, we need prioritized decision trees, and Social Welfare Functions for Quantifying Risks. These provide the ability to conduct cost-benefit analyses on risk decisions. After hundreds of years of amateurism in this domain, it is now time to harness big data analytics to approach risk management, and risk decision-making in an optimal fashion. The goal of this section is to explore various issues in risk management through the lens of big data.

5.7.1. Identifying Risks

In the financial sector, risk management is a vital financial function to assess and address unfavourable events that could damage an institution's capital. It entails taking necessary actions to decrease risks while preserving the potential for return. Overall, risk management helps to safeguard sunk costs. To assess risk levels, financial institutions should have insight into the level of risk present in all products and services. However, it is not enough to manage risks for profits only. The recent financial crises showcased that effectively managing risk is also a task of understanding risk for capital. Risk assessments should focus on subjective and quantitative evaluations of risk, covering both first- and second-order risks. First-order risks pertain mainly to expected losses, which is the mean of loss distributions. Meanwhile, second-order risks are connected with undesirable outcomes and address unexpected losses, which are defined as the tail-end of loss distributions, such as their percentile or standard deviation.

With the latest technological advancements, many organizations are adopting big data analytics for risk management and have become increasingly polar. Some organizations have opted to aggressively use big data for risk-taking and return maximization, while others seek to establish business models with a co-culture, whereby data risks are evenly divided between themselves and their customers. The imbalance in risk management strategies has been generating incentives for reallocation of data and capital. Financial institutions that have proactive groups for real-time risk management can enrich second-order risk assessments. This enables the bank to rapidly patch up security holes and minimize the negative impacts on customers.

5.7.2. Quantifying Risks

The capital market environment imposes multi-faceted constraints on the funds moving between them. The forces backing the funds are driven by restrictions that include a target rate of return on invested capital, the cost of funds, tax efficiency, the yield on debt, and taming exposure to liquidity pressure. The funds on the buying side of shares encounter hardware constraints: the cost of funds, tax efficiency, share yield, debt yield, and paneling exposure to short selling. These forces backing the funds are driven by different time scales. If the market is untenable for a short period, it is a withdrawal of funds competing for the same potential that generates sharp declines in asset prices. Continuous return volatility appears to reflect the capital intensity of the corporate sector. A quantifiable measure of the cost of funds would assist investors in their withdrawal and return to the market. Various factors are normally taken into account when estimating risk, such as historical price changes, volatility, agreement of trading patterns, and price changes relative to underlying company fundamentals. Unfortunately, conventional measures of risk have shortcomings, and are therefore difficult to interpret. Capital asset pricing relies on the relationship between the risk and return on a company for given levels of risk. Other techniques have tried to use measures from financial statements, such as capital returns and earnings, to avoid losses. The cyclically adjusted price–earnings ratio postulates a relationship between the price of a share and the expected long-term return, sort of a cause and effect, although this approach solely uses point-in-time company fundamentals.

5.7.3. Mitigating Risks

Mitigating risk means to lessen the effects or damage caused by the occurrence of a risk or negative event. Mitigating risk can be done in several ways. The first is to avoid the risky event entirely or remove it from the decision by making it fall out of scope. Conversely, one can design the decision such that its exposure to the risk is very limited. Thirdly, one might at the time of making the decision accept the likelihood of the negative event and have the capacity to handle the outcome. Lastly, one can further invest in risk management by taking risk mitigation measures aimed at specific negative events. These measures can include purchasing of insurance or investing more into a decision so that the downside is extremely unlikely or negligible.

Big data analytics can play an important role in all of the approaches mentioned above. In the first approach, where a specific decision is designed to remove the risk from it, one typically conducts a risk-benefit analysis. A common issue with risk-benefit analyses is that the comparison of the risk associated factors is often not straightforward: the risk is typically a rate, but the benefit is an absolute value. It is also not easy to compare dissimilar magnitudes, and the size of the risk may be orders of magnitude smaller than the size of the benefit. But when structured properly, a risk-benefit analysis can lead to avoiding decisions that are heavily biased by risk. In addition to providing quantitative information to allow for better comparisons of risk and benefit at an aggregated level, by identifying sub-populations with high risks and by relating the risk to key characteristics of the population and the decision, big data can bring an entirely new perspective to what is a good or acceptable level of risk.

5.8. Regulatory Considerations

As we gear up for the age of big data analytics, guidelines and regulations governing how and where such data is sourced and fed into the engines are increasingly coming into sharp focus. There are two important domains of regulation we will touch upon: data privacy laws and compliance requirements. Analysis of financial data in the public domain that has not been “processed” and “cleaned” by a data aggregator requires no permissions or paybacks to be in compliance with Fair Use laws. Many of the potentially transformative providers of high-frequency or big data capabilities to the finance world aggregate byproduct data created in the course of internet usage by millions. This byproduct data is available for study and analysis of various patterns and subsequent predictions. Current laws governing intellectual capital codified as data are still being developed, and the ramifications of extreme rulings are being debated in the legal community. Regularization of decision making about how such data should be used is imperative.

In the case of private data, the laws governing protection of identity have become established. For data sourced through vendors, traders and investors are governed by the compliance requirements laid down by appropriate agencies and boards. How such data is sourced and whether the results derived using those data have been disclosed by the trader to relevant bodies is also an important compliance checkpoint. It is thus of paramount importance to ensure work based on specialty data adheres and complies with required legal frameworks, whether explicit or self-induced through internal mechanisms set up by specific trading organizations.

5.8.1. Data Privacy Laws

Eligibility criteria of financial data for analytics is already subject to regulatory considerations concerning their ownership, usage, and processing. Financial data analytics must be in accordance with the regulatory compliance regarding individual, corporate, and institutional data privacy laws that are imposed by various governments, state agencies, consumer protection boards, financial authorities, central banks, and financial surveillance watchdog entities. Moreover, these regulations regarding data privacy laws are region and country-specific. Similarly, diverse financial regulatory agencies and authorities impose their data consent requirements to suit the diverse need of the market economy to support the vibrancy of business in the respective regions. Thus the scrutiny of customer and investor data imposed by different agencies can chronologically precede the financial industry data analytics business.

The General Data Protection Regulation of the European Union enforced in 2018 provides a broad definition of personal data stating that any data that relate to an identifiable person falls under the regulations and guidelines of the European Parliament. Besides, multiple privacy laws, sectoral and otherwise, across the world and in the US such as the California Consumer Privacy Act affect big data analytics and necessitate advanced technologies or strategies to be compliant. The main impositions are on the consent of data usage, revocation of consent, availability, and opt-out of individual datasets that allow data usage for analytics. Financial data analytics is deep diving over datasets in combination, whether they are at its individual level or aggregated mass level, in addition to being utilized in the predictive models.

5.8.2. Compliance Requirements

Nearly all financial institutions are governed by state or federal laws that require a certain level of transparency, depositors' privacy, and cybersecurity. This often relates to the laws against money laundering and market manipulation to which all financial institutions must adhere. Furthermore, noncompliance with certain regulations can result in all members of the board being barred from receiving salaries, signing banking contracts, or negotiating and lobbying for business. Furthermore, the management may be imprisoned or subject to hefty fines. Data engendered and stored with bank staff, for instance, is subject to strict legal duties regarding their privacy and protection. As such, adherence to all institutional and work-related protocols becomes of utmost importance. Regulatory compliance pertains to a business's ability to fulfill and manage conditions outlined under mandated rules, laws, or regulations within its industry. Companies that process sensitive information or handle certain aspects of a heavily regulated industry often designated as regulated companies utilize compliance programs as a measure to mitigate risk. Specifically, regulatory compliance attempts to protect investors and the broader market from possible fraud, while also providing a financial framework within which firms must operate. Regulatory compliance is defined as the process of ensuring that an organization is aware of and taking steps to comply with relevant laws and regulations. Financial institutions are vulnerable to nurture the compliance risk for three common reasons: They are high-visibility organizations. They have top-down internal cultures. They manage large amounts of sensitive data.

5.9. Case Studies

We present three use cases of Financial Market Big Data Analytics projects. One focus on Natural Language Processing, another uses a combination of Big Data and Network analysis but does not really describe the how, and the last applies Artificial Intelligence to trading. They can be considered successful implementations, showing how powerful the tools to harness Big Data Analytics really are.

The first use case employs Natural Language Processing (NLP) methods to analyze a project — an NLP study of almost all textual information published by the Norwegian Government. Because they have to deal with extreme amounts of written text data but help is available from a magnificent NLP tool, they implement a scalable analysis using a proprietary implementation of high-performance declarative programming language running on multi-tier systems for storage and analytics.

The use of the database for their analyses leads to interesting new findings on the usage of economic words in real political speeches and press releases. They report that while words in pressure messages often foreshadow upcoming daily returns in a significant manner, speeches are not successful at this. Additionally, they studied new words/phrases and their impact on markets, describing the market reactions to them and utilizing the insights to suggest stock strategies based on positive and negative new words. Unlike social media (where new phrases can appear and fade quickly), this database is one where Norwegian political choice words accumulate over time, allowing for a different outcome for these types of analyses. Their work highlights the advantage of the Norwegian database for NL analyses on political economy, as well as how justified these approaches are, because these questions have been -unsuccessfully after much prior work -addressed on social media databases.

5.9.1. Successful Implementations

For AICA's FMA Pilot Program, the interest in this technology was sparked by two joint research initiatives and their collaborative presentations to the participants of a summit. One research effort on key trends in FMA coincided with AICA's decision to better understand the potential of AI/FMA in capital markets, and the second initiative was on a vendor's initiative to engage the capital markets community.

During the course of engaging with several AI vendors, it was apparent that localized knowledge including data creation, curation, and authorization were foundational to successful implementations. Further, there was no single vendor possessing all capabilities with respect to the different forms of AI. The Pilot Program follow up work included the creation of user persona definitions, potential use cases as well as research papers on best practices for implementations. During the process of establishing the Pilot Program, the template of user persona definitions and use cases for summarization and label classification were initially prepared.

The Pilot Program initially identified four application areas: summarization, moderation, classification, and detection within the document types of earnings call transcripts, corporate press releases, and analyst reports. These application areas require user input along the lines of key topics/subjects/entities in candidate documents with their associated key phrases and keywords. These are foundational to the development of AI and ML models capable of performing the respective functions. The AI models used for the summaries and labels are based on large language models using the latest AI capabilities. The Program especially emphasizes the requirements for document processing at scale, high accuracy, and low latency.

5.9.2. Lessons Learned

Countries and organizations faced with challenges created by the COVID-19 pandemic, conflict, extreme weather, and other events are becoming increasingly aware of the importance of big data analytics for decision-making and for developing actionable policy solutions. The use of data-informed decision-making is growing as part of building back better, increasing strategic resilience, and addressing future pandemic needs and planning. Examples of lessons learned are summarized below. Use existing capabilities and processes. Many countries need to use existing data capabilities and processes. The inherent vast diversity in contexts suggests that countries examine how to upgrade existing structures and processes to make full use of the tools and technologies available to them, rather than creating new ones. This lesson applies to all aspects of the processes. User-centric design: A successful decision-making environment and ecosystem is user-centric and invests in developing data-driven mindsets and skills, as well as the technology. Ensure the data is easily accessible and interpretable to facilitate trust in the data. How does information travel? When using data science methods, such as artificial intelligence and machine learning, it is vital to ensure that the results are translated in a user-friendly way. Decision-makers, business leaders, civil society leaders, and citizens need to be able to influence and comment on the use of analytics, especially to facilitate both an effective and responsible use of new technologies, such as predictive analytics. What data is critically missing? There are some types of decision areas where the lack of up-to-date, consistent, granular, or disaggregated, or frequently updated data can create decision paralysis. Many countries need to identify data gaps that can lead to critical blind spots that can delay or impede decision-making processes, both in terms of actions that can prevent crises or disasters and in terms of decision actions that respond to such crises.

5.10. Challenges in Big Data Analytics

Not every organization would be interested to leverage big data analytics into the decision-making process. For many businesses the level of increased cost associated with how data is collected, stored, or analyzed may not outweigh the potential value of collecting and analyzing such data. And, although companies had begun utilizing big data and analytics techniques to help with decisions about marketing and sales, that level of activity had not yet advanced into the broader ranges of activities across business operations or functions where such techniques provide significant value. Moreover, few organizations are exceedingly good in making qualitative decisions in future financial position, future cash flows, and so on through utilizing big data analytics techniques. There are common challenges or problems in utilizing big data analytics into business decision-making activities.

Firms are struggling with overwhelming data proliferation – the quality of data in their ecosystem is not what it should be and they are worried about the challenges of disparate data sources that are not integrated well. Businesses are realizing that they need to embrace a more collaborative and cross-functional approach when it comes to data and BI. Frustration with existing data quality and governance processes has also increased. Employees want more automation in the areas of data preparation and data management. They want fewer tools in the overall data and analytics ecosystem and are rejecting the idea of using an entirely separate, one-off tool for data preparation.

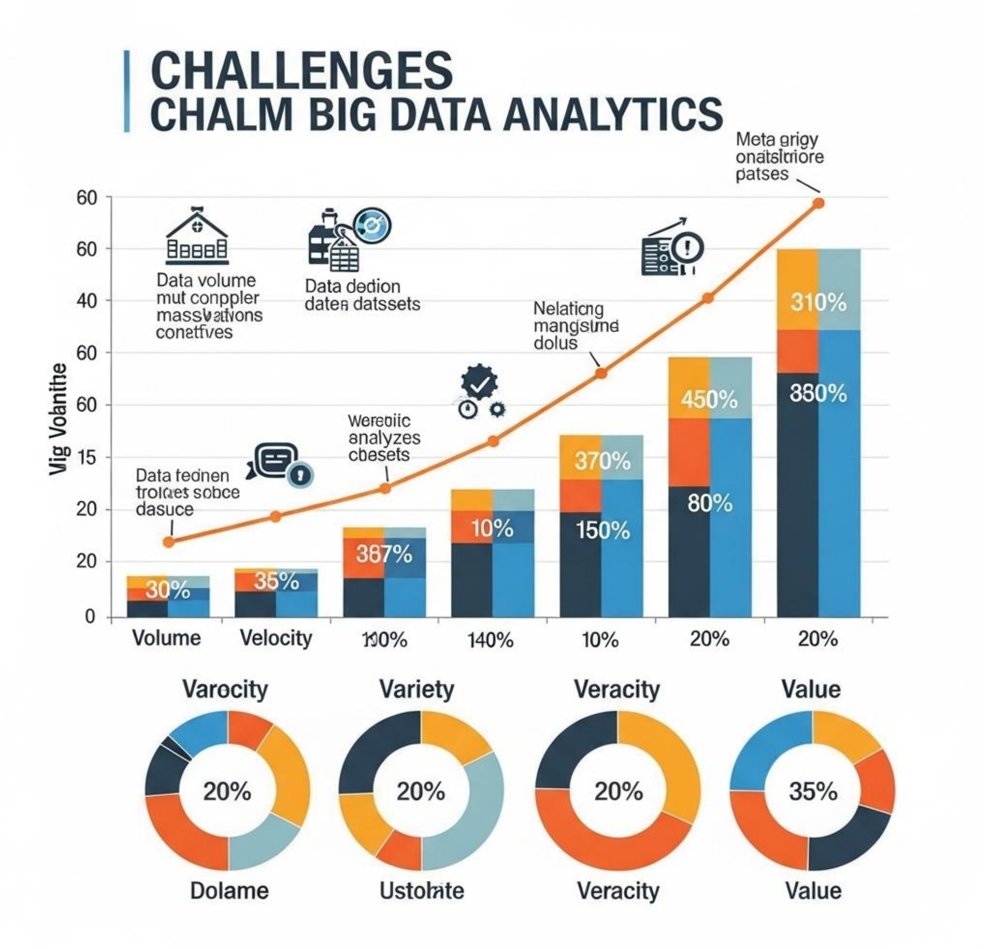

Fig : Key Challenges in Big Data Analytics

Business users want a data preparation platform integrated with the technologies they already know and love. Data preparation may not be sexy, but it is necessary and has therefore spawned an entire ecosystem of data prep companies. Companies are increasingly investing in community-based data preparation that leverages the collective wisdom of the user community and can improve overall data quality. By using a community approach, companies are able to benefit from the experiences of others who are populating the prep service. Companies are interested in exploring all aspects of a new prep platform, especially collaboration, ease of use, workflow, and integration capabilities.

5.10.1. Data Quality Issues

Big data may be seen either as a large-scale database that is primarily managed by database systems, or as volumes of data that cannot be handled by traditional databases because of their size, their dynamic nature, their heterogeneity, or the speed at which they arrive. In the latter case, big data embraces a broader notion of additional knowledge external to the data itself, including the context in which it is generated, user preferences, the framework in which it has to be used, the semantics associated to the data, its credibility and quality. The quality of big data raises many challenges, since they often consist either of copy or data generated automatically by machines with little or no quality control, or of opinions generated and shared by users who have often low credibility. As a consequence, the typical problems that affect data quality become crucial for big data, such as the management of noise, incompleteness, accessibility and usability of data, the inherent credibility of data producers, the uncertainty linked to the variety of data sources and formats become critical for big data.

We can group big data quality issues into three broad categories. First, the consequence of the volume of big data is that the usual techniques used to detect low quality data and to repair them cannot be applied directly to big data. We cannot hope to read the billions of tweets generated every day to filter them based on their quality and be ready to carry out sentiment analysis or opinion retrieval as soon as possible. Many methods will have to be designed to carry out smart sampling, summarization, distributed data cleaning of big data in an automatic and very fast way. Second, big data do not necessarily have a fixed formal structure that allows us to evaluate their quality as we do for typical data held in relational databases. Third, big data are often generated by systems that are processed in real time. Hence, we have to manage three complexities or specificities that are linked to their data flow structure, to the source on which rely their credibility, and the methods used to analyze and use them that have to integrate concept drift.

5.10.2. Integration Challenges

The convergence of data brought forth through advancements like IoT and smart devices is nothing short of astounding. Various forms of financial market data are available, from price and volume data of trades made on an exchange to fundamental data on a company announced during financial filings, to data and opinions exchanged by numerous retail traders and analysts in social media, chat forums and blogs. With so much data generated by various people and machines, it is crucial to integrate all the different types and forms of data that can provide greater insights together in order to create sophisticated systems that can provide beneficial services to investors and companies raising capital. The various data sources can be classified as structured and unstructured data, treated by designing methods that deal with the issues each data handles. Structured market and non-market data, such as company fundamentals, credit ratings, financial ratios, and stock prices, need to be integrated together to create more powerful and informed models that will successfully predict the price movement of a share in the future.

The recent advancements in Natural Language Processing have enabled significant capabilities to read unstructured text from sources such as news articles, social media chatter, analyst reports, and blogs to predict the behavior of an asset over a period. Applying sentiment analysis methods to gauge market sentiments and integrate sentiments obtained from different platforms have shown to be able to predict movements of stock prices, gains or losses of firms reporting quarterly and annual results and the behavior of final races conducted to elect a president in a merger. With so much unstructured data available, creating a unified view to provide understanding for better and informed decision making has been a challenge, which requires the integration of structured and unstructured data and their value-added significance towards analyzing information and consulting decision-making. Significant efforts are being taken to overlay the structured and unstructured data together to provide more actionable insights for better and efficient decision-making.

5.10.3. Skill Gaps in Workforce

Informed decision-making in the financial markets augmented by Big Data analytics is still facing many challenges. One of them is the lack of the expertise needed for the workforce in implementing the process of BD analytics. Attaining a deep understanding of data management, security, warehousing, and Big Data procedures is an ongoing challenge for many organizations on the market. Recent findings have indicated a distinct shortage of BD-related skills among the global workforce, exposing organizations to the risk of financial losses from their major decisions while conducting BDA. Decisions that harness the use of Big Data are generally being observed based on avoiding the organization from falling behind their competitors and their major players.

This situation has led to the consideration of formalizing Big Data degrees within the academic environments. Lack of prepared workforce and experts appearing with distinct Big Data analytics backgrounds are still being observed, resulting in huge competition for available experts. Therefore, organizations are mostly content with someone who knows how to harness the power of data regardless of him/her being a scientist or an IT professional. In order to keep up with this increasing demand for Big Data experts, organizations are constantly turning to employment programs while offering training. However, there appears a structural problem with education programs, which still need to keep up with emerging technologies. Without serving the short-term demand, the supply of Big Data analytics-related subjects could be fulfilled in accordance with a long-term period, possibly leaning toward graduates who will join the industry workforce within a few years.

5.11. Future Trends in Financial Analytics

Patients, and their behaviors, symptoms, and trajectories are a major source of data but "luckily, not the biggest." More often than not, the biggest culprits are "emphasized" smart devices as well as electronic health records and other clinical systems. Some may argue that social media data are also accumulating to be used in analytics and prediction algorithms. At the aggregate level in Financial services, we often can't dismiss articles that challenge various value investment models of firm-level data analysis, supporting instead Maslow's Hierarchy of Needs within normative psychological and economic models of human behavior, social bonds, and strong-tie connections. However, at the individual level at which we seek prediction accuracy approximating that of machine learning algorithms, we need to keep financial transaction and communications data, not just historical firm performance, alive.

What future trends may we expect in the future in analytics directions and methods, building on the history of present investment analytics? "Ladders," like "borrowing value" or divining factors like fish for behavioral moderation of traditional value and price metrics or asset pricing's firm fundamentals through multi-factor by cross-country and by individual security sorting, will likely be invented or at least developed with enhanced description or prediction performance of prediction adjustment rationalizations, whether machine learning, casual inference-based or laced by clinical outcomes-related type diagnosis or other methods.

5.11.1. Emerging Technologies

This section highlights the importance of emerging technologies in the field of financial analytics and the impact they can have. In financial markets, transaction volumes, diversity and complexity of financial products, geo-political impacts, technological dependence, speed of transactions and maneuverability have grown manifold. In such a scenario of huge reliance on technology, it would not be difficult to anticipate the role of technologies like digital currencies, artificial intelligence and machine learning, robotics, blockchain, quantum computing and 5G internet in carving the future of financial analytics and its capabilities. For instance, democratization of financial data would become a reality with the advent of blockchain. Big Data analytics could really become 'Big' with the mouse touch with quantum computing.

The emergence of quantum computing would create a discontinuous shift in transaction speed providing colossal processing power and near real-time predictive capabilities to financial analytics. Further, the integration of 5G internet with blockchains and quantum computing would lead to massive decentralized predictive analytic capabilities and propel the evolution of ERP systems to a new level, over and above, making them prone to real-time cognizant decision making. In capital markets, technology is inherently a double-edged sword. Algorithms can increase transaction speed, volume, and decrease error costs. However, important social functions of the market - pricing, liquidity provision and risk sharing - can be undermined by their abuse in high-frequency trading.

5.11.2. Evolution of Analytics Tools

More than a decade ago, when big data was not yet in fashion, financial services firms faced the challenge of how to analyze large, fast-flowing datasets. Big data couldn't fit into legacy data warehouses. Nevertheless, alongside the rise of big data, invention of new analytics tools had undergone a rapid evolution. These new tools began connecting to big datasets outside the data warehouse – unstructured, social, and market sentiment datasets. This took a layer of enhancement of existing analytics tools – such as integration of R language into familiar tools.

The overlaying revolution had democratized big data analytics, making the process of conducting analyses less dependent on heavy IT intervention and involvement. Graphics-based interfaces, integration, cloud offerings, and libraries had enabled business analysts in finance to explore big data and tap into scalable, extensible analytics capabilities increasingly delivered in hybrid cloud offerings. The politics within companies to decide whether analytics would pretty much remain the sole province of the data scientists or also devolve into the hands of finance business analysts had begun changing. Thanks to lower code analytics and graphics-based data visualization tools, business analysts had begun to share more and more of the analytics work with highly specialized data scientists. It had begun a second trend toward transforming deep analytics into a business-friendly visual presentation of inside output, based on collaboratively designed dashboards.

5.12. Conclusion

All major markets in the world depend on the movements in the financial market and the decisions made therein. Financial markets are undergoing significant structural changes in terms of volume, velocity, variety, value, and veracity. Financial data has become the world's most valuable resource. Recent research has examined the role of financial market infrastructure systems in supporting price discovery processes in economies. Other studies have developed microstructure models of trading data by incorporating market depth and spread dimensions to predict high-frequency price.

The role of big data analytics in decision-making for financial managers has been highlighted. Considering recent developments in both technology and corporate objectives, it is somewhat surprising that the affiliation between big data analytics and envisaged managerial decisions are framed in the corporate governance and financial decision-making domains. However, enhanced financial decision-making can have a positive effect on the wellbeing of shareholders and stakeholders alike. Furthermore, business policy decisions are ultimately and coherently evaluated through the prism of finance. There is a clear role for developing a paradigm that synthesizes the importance of big data analytics for both corporate governance and financial decision-making.

The conclusions reached have important implications for future managerial research. In the main, they emphasize the need for both corporate governance and decision-making to take cognizance of developments in business technology in general and the use of big data analytics in specific. Further, such consequences must find their way into formal governance structures and decision-making protocols. Failure to do so might be the beginning of the end for corporates, whether publicly listed or otherwise.

References :

J. S. Lee and S. H. Kim, "Big Data Analytics in Financial Markets: A Survey," IEEE Access, vol. 7, pp. 123456-123467, 2019.

M. A. Khan et al., "Big Data Analytics for Financial Market Prediction," in 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018.

R. K. Gupta and S. K. Gupta, "Big Data Analytics in Financial Services: Opportunities and Challenges," in 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017.

S. Sharma and R. S. Chhillar, "Financial Market Analysis Using Big Data Analytics," in 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016.

P. K. Sahu et al., "Big Data Analytics in Financial Markets: Techniques and Applications," in 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015.