International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 3, May June 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Big Data in Finance Sector

Sinchana Gaonkar, Anitha Sandeep

(Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bengaluru)

Email: sinchanagaonkar99@gmail.com
anithasandeep@rvce.edu.in

****

Abstract:

Beginning from the last decade, big data technologies have contributed tremendously to various industries. The finance industry was mostly focused on structured data investigation until the advent of big data technologies. With the help of big data technologies, information stored in various sources of semi-structured and unstructured data can be harvested. Studies indicate that such information plays an important role in the decision-making process in financial institutions. The finance field is heavily focused on the calculation of big data events. Millions of financial transactions occur in the finance industry each day. Also, big data has a huge impact on financial services and products. Therefore, financial practitioners and analysts consider that it is important to analyse variousfinancial products and services using big data solutions.

***Keywords* — Big data, finance sector, stock analysis, risk management, data mining.**

****

I. INTRODUCTION

The Financial transaction system has developed significantly with time starting from the ancient barter system to the state-of-the-art e-commerce system today. The finance industry has thrived significantly with the rapid development of human civilization. Business intelligence and transactions in finance institutions had excessive human involvement earlier. But with digitization taking over, the transactions became more transparent and clearer and also generated enormous amount of data. Structured data is information managed by an organization for providing key decision-making insights. Unstructured data exists in multiple sources in increasing volumes and offers significant analytical opportunities. These digital footprints have become amenable for rigorous analysis using the new field called analytics in order to make clear and right business decisions. Over a period, the finance industry has produced a large amount of diverse data at a break-neck speed due to ever-growing young customers. This has led to the advent of a new generation of data analytics paradigm called Big Data Analytics [1].

Billions of dollars move across the globe every day. Analysts are responsible for monitoring this data with great accuracy, precision and speed to discover patterns in this data and create predictive strategies. This data is of great importance and also has enormous latent business opportunities. The value this data can offer to the business is dependent on how the data is gathered, processed, stored and interpreted. Since legacy systems cannot support unstructured data without significant complex IT involvement, analysts are increasingly adopting big data solutions. Big data technology enables human beings to visualize the underlying value of data and apply it to the financial field to enable modernization of the finance market. However, the application of big data technology in finance markets has its own risks and sufficient attention needs to be provided to it [2]. About 20 petabytes of data is generated around the world every day and this data is estimated to go up to 163 zettabytes by 2025. Financial institutions like investment banking firms, banks etc are the leaders in the use of big data.

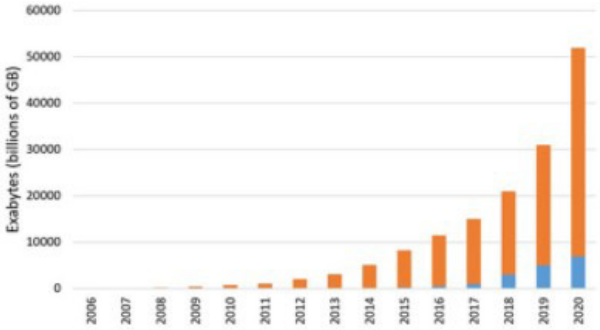

Fig. 1. Dynamics of the volume of information in the world

They have gathered enormous amounts of valuable information during their existence that need further analysis and processing. There are quite a lot of questions with respect to the applications of big data technology in financial services [3].

This study provides an overview of how big data technologies can be used in finance.

II. BIG DATA TECHNOLOGY

The term Big Data describes the huge amount of data. This data can be structured, semi-structured or unstructured. But the amount of data here is not what is important to the business but what organizations do to extract value from it is important. Big data can be analysed for insights that lead to better decisions and strategic business moves. Analysing this massive amount of data and making a better consumer experience possible with better data management has led to the genesis of Big Data Analytics (BDA). A huge volume of data is generated in less span due to financial transactions in finance, service, banking and insurance sectors. This is generated from different digital devices involving various types of data formats [1]. Big data solutions cut costs of on-premise hardware with limited shelf life. They also improve flexibility, scalability and security across all business applications. Big Data has 5 major characteristics called the 5Vs. They are Velocity, Volume, Variety, Value and Veracity. Volume represents the large amount of data generated every moment. The term Big Data in itself represents size which is enormous. Humongous data sets result from a large number of transactions. It is almost impossible to store and analyse this huge amount of data using conventional storage and computational technology. In the past, storing this massive amount of data was a problem but cheaper storage on platforms such as Hadoop and data lakes have eased the burden. Velocity dimension refers to the speed of generation of new data and the speed at which the data moves around. Reports suggest that The New York Stock Exchange captures about 1 terabyte of information every day. Variety dimension indicates that big data can be in various formats like structured, numeric data in traditional databases to unstructured text documents, images, videos, emails, financial transactions, stock ticker data etc.

Fig. 2. The Five Characteristics of Big Data

Veracity refers to the extent of reliability of the data. The correctness and quality of big data is less controllable due to the presence of structured, semi-structured and unstructured data. Hence, veracity concerns the inconsistency in the data. The Value dimension indicates the business value that is extractible from the data. The noise component present in the data can be significantly more compared to the useful data present very often. Thus, these five dimensions succinctly capture the entire characteristics of big data [1].

A. Introduction to Apache Hadoop

A large volume of data can be stored in distributed file and can be processed or analysed in a distributed manner with the help of Map Reduce architecture in the Apache Hadoop framework. The Apache Hadoop framework consists of the following modules: (i) Hadoop Distributed File System (HDFS), (ii) Hadoop Yarn, (iii) Hadoop Common and (iv) Hadoop MapReduce. The HDFS is a distributed file system and it stores data distributed over various machines present in the cluster. Hadoop Yarn manages computational resources that are present in the cluster and also schedules user applications. Hadoop common consists of utilities and libraries that are required by other modules present in the Apache Hadoop framework. MapReduce architecture enables distributed processing and consequently enables processing of large volume of data at a decent speed.

B. Introduction to Apache Hadoop

MapReduce programming is limited to batch processing and real-time data cannot be processed using it. Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. Spark can work over the Hadoop framework or as a standalone platform to perform real-time data analytics in a distributed computing environment. Complex analysis in business intelligence, streaming and batch processing, machine learning etc. can be performed using Apache Spark. It makes use of horizontal clustering for speedy and efficient computation. Spark runs an application in Hadoop about 10 times faster when running on a disk and 100 times faster in memory. Spark enables implementing machine learning tasks like classification, optimization, regression, dimensionality reduction, clustering etc. efficiently [4].

Apache Spark employs MR model for an extended computational framework subsuming interactive database queries and online processing through streaming. Different workloads like iterative codes, batch program, interactive database queries, streaming data etc. can be handled by Spark. Spark stores intermediate values in memory during execution

and it is faster than distributed programming due to the reduced number of read/write operations to the hard disk [1].

Spark core engine supports APIs in Java, R, Python or Scala and it also supports 'Map', 'Reduce' operations. Spark also supports SQL queries (Spark SQL), Machine learning (MLlib), Graph algorithms (GraphX) etc. Alluxio provides shared memory facility by sharing the same data among multiple applications. In Spark, cluster management is achieved in three ways viz. Hadoop Yarn, Standalone and Mesos. Spark can access the data from local file system or any distributed file system like Hadoop Distributed File System (HDFS).

III. APPLICATION OF BIG DATA IN THE FINANCE SECTOR

Financial sectors use big data tools to predict and analyse stock movement and securities in market by building a predictive algorithm. Another area in which big data is of significant use is in analysis of a large amount of stock market exchange data and then use it to make certain important decisions [5]. Intelligent decision systems with data mining and machine learning can be used in stock market forecasting. Investment portfolio managers and stock market traders can process and analyse a large volume of unstructured data to identify the best companies to invest in. Banks and other financial institutions in turn use unstructured public data like product review, news, supplier price and data, and process them with big data tools and hence produce a mathematical model that helps stock traders to decide whether to buy or sell stocks. Another significant field in which big data is used is in extracting sentiments from news, which analyses and concludes if the article is neutral, positive or negative [6]. A survey indicates that around 30% of companies in the finance sector have begun implementing big data and are targeting on techniques to use big data to reduce risks and manage finances more efficiently.

A. Financial Risk Modelling Using Big Data Technology

As markets are becoming increasingly interconnected, associated financial risks are also increasing. Applying big data to risk management has become very important as the amount of data grows tremendously every day. In the finance sector, big data technologies are most frequently used to build predictive models for fraud prevention and models to monitor and analyse behaviour of customers. Important financial decisions like loans and investments now heavily rely upon unbiased machine learning. Decisions are made based on predictive analysis. Economy, business capital, customer segmentation etc. are used to identify potential risks and bad investments.

In this paper [7] an approach is discussed to evaluate risks in financial contracts. This approach considers that financial analysis can be reduced to 3 main steps viz. evaluation of cash flow of all contracts, application of specific analytical transformations to these cash flows, which results in granular analytics and then aggregate granular analytics to a desired level. The data input to the system are risk factors and contract data. Contract data determines the dates, types and amount of cash flow that is generated by a contract. Risk factors indicate the state of economic and financial environment under which cash flows should be calculated. Some significant risk factors are foreign exchange rates, interest rates, prices of shares etc. The paper demonstrates how a large number of contracts can be evaluated for risks using Apache Spark. Scalability is studied by doubling the size of CPU cores and input data and linear scalability is observed starting from 32 cores. So, this paper concludes that analysis of 96 million financial contracts can be done on 512v CPUs with nearly linear scalability. The paper basically shows how existing analytical approaches can be applied to huge amount of data using big data technologies.

In order to successfully use big data solutions for risk analysis, companies first need to work on collecting the necessary internal and external data and utilize them. This will provide the company with a good understanding of the data sources and also the volume of data they are generating. Data from externals sources like tweet data can be used effectively for financial risk modelling. Graphical Gaussian models can be made use of to estimate the relationships between bank tweet sentiment variations. The paper [8] shows how to combine tweet based systemic risk networks with those obtained from financial market data, using the posteriori Bayesian mean of the complete variance-covariance matrix. Two different sources of information, market prices and tweets are integrated here to estimate systemic risk networks. Therefore, an integrated analysis process is necessary to efficiently analyse big data to predict risks.

Despite the fact that big data technologies have great impact in risk management, there are still some obstacles to overcome. Some of these obstacles being lack of expertise among the employees to handle big data, data protection issues, money constraints etc. Increasing number of companies are taking small steps towards incorporating big data solutions in their business and this will enable them to identify areas of risk in their companies.

B. Stock Market Prediction Using Big Data Technologies

Big data analytics on stocks help investors to determine what is the best time to buy or sell stocks based on the predicted stock prices. A lot of research has been in progress on how to efficiently use big data in stock market prediction and some of them are discussed here.

One of the proposed approaches for this is to make use of a feed forward neural network. As mentioned earlier, Apache Spark supports machine learning libraries and this enables application of machine learning algorithms to big data. In this approach, an adaptive moment estimation optimizer is used and Relu is used as the activation function. The collected data is pre-processed and divided into train and test set just like in any machine learning approach. This data is then fed to various models like support vector machines, multi-layer perceptron model, ARIMA model etc. Results indicated that feed-forward neural networks give the best accuracy for predicting stock prices [9].

In this paper [10], a system is proposed where data is collected, pre-processed and then transformed from high frequency data to a ratio matrix. Then an outlier mining algorithm is used to find out the anomalies in the data and predictions are made based on the positions of the anomalies. Evaluations on real exchange data are done and the method is found to be a lot more efficient than the traditional data mining algorithms. Another proposed approach for prediction is to use hierarchical clustering, step-wise regressions and an ANN model for detecting historical patterns in stocks and predict daily stock prices by optimal significant variables using feature selection. Big data frameworks based on R and Hadoop have been made use of for processing and the accuracy is found out using the RMSE values of stocks [11].

This paper [4] proposes a system which uses robust Cloudera-Hadoop based data pipeline to perform complex analysis for any scale and type of data. Here, certain US stocks are analysed in order to predict daily gains on real time data obtained from Yahoo Finance. In this approach, multiple opensource modalities of Apache Hadoop ecosystems are used in order to build a cloud architecture. The linear regression-based learning model is trained with the training set and then it predicts the correlation between stock prices based on coefficients in the regression model. Mean Average Error and R squared value are also calculated to support the study.

The paper [4] describes 5 steps to implement the proposed system- First the data set is obtained from Yahoo finance which contains 13 columns and 2905 rows. This collected data is then injected into HDFS using Apache Flume. Flume consists of three main components viz. Source, Channel and Sink. The source receives some external events and stores it in a channel. Channel obtains data from source and buffers them till they are consumed by sink. Sink consumes the data (events) from the channel and delivers it to the destination. Data set is stored in HDFS on Cloudera. Then the data is pre-processed using a python API named PySpark. Sequence file is converted to RDD (Resilient Distributed Dataset) in this step. Finally, Mlib library in Spark is used to perform regression analysis on the data. Linear regression function is fit to the training data set. For the training data set, returns of the USO are predicted and mean squared error is computed. The Mean Average error turns out to be 1.95% and this suggests that a linear regression model is not suitable for predicting stock return margins form the data with high dimensionality.

Machine learning is changing trade and investments. Instead of merely analysing stock prices, big data analytics also considers political and social trends that affect the stock market. Machine learning monitors trends in real-time and this enables analysts to collect and evaluate appropriate data and make smart choices.

C. Other Applications of Big Data Technologies

Finance companies generally have a huge customer base and losing these customers is painful for companies. Also bringing in new clients is very important for the business. Banks are increasingly shifting from being product centric to customer centric. Various metrics can be used by finance companies to measure probability of losing their clients. Big data technologies play an important role in performing such analysis. A large amount of data is generated from the communication between the customers and the financial institutions. Data mining techniques can be applied to this customer data and the results can help in managing complex customer relationships, identify key customer and also predict the possibility of the customer leaving the firm. These results help the staff to develop plans to retain customers [12]. Big data analysis help companies to know customer details like personal behaviour, demographic details, transaction details etc. This helps banks to target customers based on their interest and behaviour.

Financial fraud detection and risk management are some important aspects that banks need to focus on. Big Data tools can help analyse and find the risk levels, forecast them, and help decision makers to avoid risks thereof. Economic crimes include malicious overdraft, money laundering, forging credit cards etc. Such economic crimes are on the rise lately. Data in multiple databases like crime history database of an area, bank transaction database etc. can be integrated, abnormal patterns

in this data can be found using big data tools, and this helps in solving economic crimes [12].

Security risks posed by credit cards can be effectively mitigated using analytics that interpret buying patterns. When credit card information is stolen, banks can freeze the card instantly and also notify the customer about the same. Predictive analysis can be utilized to detect money laundering and other related fraudulent activities. The enormous volume of data obtained from different sources help in monitoring different platforms closely. This leads to an increased probability of detecting plans to engage in fraudulent activities even before it happens. The use mode of customer credit card can be analysed and the use of credit card by customers can be monitored dynamically. In case of unusual use of credit card by a customer is noticed, the bank can take necessary measures to prevent loss. Data mining tools can be used to identify and analyse fraud patterns and therefore provide warnings to financial institutions and remind them to strengthen their management and supervision [13].

Algorithmic trading is being widely used in financial institutions. It is an automated process where computer programs execute trades at high speeds which is not achievable by humans. This approach uses mathematical models to execute trades at the best prices and also trades are placed in a timely manner. Models are trained using massive amount of historical data in order to enable less risky investments. Since the algorithm can be developed for both unstructured and structured data, incorporation of social media data, stock data etc. can help the algorithm make better trading decisions.

IV. CASE STUDIES OF COMPANIES USING BIG DATA TECHNOLOGIES

Big data technologies are increasingly being used in finance companies for various applications. Some of them are discussed in this section.

The United Overseas Bank (UOB) has leveraged Big Data Technologies for financial risk management. The company developed a big data-based risk management system. This system enables completion of risk calculation tasks in just few minutes which used to take around 20 hours before the introduction of the system.

Danske Bank, which is the largest bank in Denmark, had a very low fraud detection rate of 40% and had up to 1200 false positives per day. The company was trying to find efficient methods for fraud detection. The company then joined hands with Teradata (database and analytics service provider company) and employed advanced Big Data Solutions for improving fraud detection methods. Since then, the bank saw a 60% decrease in false positives and true positive rate increased by 50%.

JPMorgan Chase and Co. has a huge customer base and massive amount of data is generated everyday including transactional information, credit card information etc. They have adopted Big Data technologies like Hadoop to handle this data. Using Big Data Analytics, they are now able to get deeper insights into customer trends. Customers are analysed individually and these analysis reports are generated within seconds.

A company named Flowcast, based in San Francisco, provides an artificial intelligence platform which helps finance companies to make data-driven credit decisions. Smartcredit, which is the company's flagship product, utilizes a diverse set of data types in order to provide more information than other risk models [14].

V. CONCLUSION

Data is being generated across the world rapidly in the finance sector due to digitization and widespread use of technology. It is deemed very important to make use of technologies to handle this data and extract meaningful patters from them. Therefore, big data technology can be very useful in such cases to handle enormous amount of complex data. Financial institutions have been using big data tools in a variety of ways to deliver better business outcomes for their organizations. Big data in finance has led to significant innovations that have enabled convenient, personalized, and secure solutions for the industry. As a result, big data analytics has managed to transform the entire financial services sector.

REFERENCES

[1] Ravi, V., Kamaruddin, S., "Big data analytics enabled smart financial services: opportunities and challenges", In Reddy, P.K., Sureka, A., Chakravarthy, S., Bhalla, S. (eds.) Big Data Analytics, pp. 15–39. Springer, Cham (2017).

[2] R. Yang, "Research on the Risk and Supervision Method of Big Data Application in Financial Field," 2020 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS), 2020, pp. 695-698, doi: 10.1109/ICITBS49701.2020.00153.

[3] A. V. Bataev, "Analysis of the Application of Big Data Technologies in the Financial Sphere," 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (ITQMIS), 2018, pp. 568-572, doi: 10.1109/ITMQIS.2018.8525121.

[4] Z. Peng, "Stocks Analysis and Prediction Using Big Data Analytics," 2019 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS), 2019, pp. 309-312, doi: 10.1109/ICITBS.2019.00081.

[5] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu ,Shuo Feng, A survey of machine learning for big data processing , EURASIP Journal on Advances in Signal Processing, 2016.

[6] A. Jaiswal and P. Bagale, "A Survey on Big Data in Financial Sector," 2017 International Conference on Networking and Network Applications (NaNA), 2017, pp. 337-340, doi: 10.1109/NaNA.2017.46.

[7] S. Kurt, J. Heitz, N. Bundi and W. Breymann, "Large-Scale Data-Driven Financial Risk Modeling Using Big Data Technology," 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 2018, pp. 206-207, doi: 10.1109/BDCAT.2018.00033.

[8] Cerchiello, P., Giudici, P. Big data analysis for financial risk management. J Big Data 3, 18 (2016).

[9] P. Singh and A. Thakral, Stock market: Statistical analysis of its indexes and its constituents, in 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bangalore, 2017, pp. 962966.

[10] L. Zhao and L. Wang, Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm, in 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 9398.

[11] S. Jeon, B. Hong, J. Kim, and H. Lee, Stock Price Prediction based on Stock Big Data and Pattern Graph Analysis:, in Proceedings of the International Conference on Internet of Things and Big Data, Rome, Italy, 2016, pp. 223231.

[12] H. Zhang, Y. Li, C. Shen, H. Sun and Y. Yang, "The Application of Data Mining In Finance Industry Based on Big Data Background," 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, 2015, pp. 1536-1539, doi: 10.1109/HPCC-CSS-ICESS.2015.198.

[13] Tao Liu. Application of data mining technology in the financial field [J].Tianjin Science Technology, Feb. 2015, Vlo1. 42 NO. 2, 51-52.

[14] A. Schroer, The numbers add up: The outsized role of big data in finance, January 3, 2019. Accessed on: May 28, 2021. [Online]. Available:https://builtin.com/big-data/big-data-finance-bankingapplications

