Rahul More, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 4, (Series-II) April 2021, pp. 01-05

RESEARCH ARTICLE OPEN ACCESS

Big Data Analysis in Banking Sector

Rahul More[1], Yash Moily[2]
[1,2]*BE Student*
[1]*Department of Mechanical Engineering*
[2]*Department of Computer Science Engineering*
[1,2]*Rajiv Gandhi Institute of Technology, Mumbai, India*

ABSTRACT
Big data analytics is the complicated process of examining large and different types of data sets or big data to reveal information including hidden patterns, unknown correlations, market trends and customer preferences which will help organizations make informed business decisions. On a broad scale, data analytics technologies and techniques provide how to analyse and visualize data sets and bring out conclusions about them to help companies to make informed business decisions.BI queries provide an easy-to-use, visual way to query databases, integrate data with other applications, and generate reports. Big data analytics applications often include data from both internal systems and external systems, such as healthcare data or banking data on consumers compiled by third-party information system services providers.
Index Terms— Big data, Banking Sector.

Date of Submission: 28-03-2021 Date of Acceptance: 11-04-2021

I. INTRODUCTION

Big Data is a combination of all the processes and tools associated with managing and making use of large data sets. The Big Data concept was born out of the necessity to know trends, preferences, and patterns within the huge database generated when people interact with different systems and every other. With Big Data, business organizations can use analytics & visualization, and figure out the most profitable & beneficial customers. It can also help in business to create new experiences, services and products. Big data analytics applications enable big data analyst, data scientist, predictive modelers, statisticians and other analytical professionals to analyse the growing volumes of structured transaction data, semi structured and unstructured data, plus other kinds of data that are often left unexplored by conventional business intelligence (BI), tableau and any other analytics programs. That encompasses a mixture of different types of data for instance , Internet data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the web of things. Emergence and growth of massive data analytics: The term big data was first wanted to ask for increasing data volumes within the mid-1990s. Those three factors volume, velocity and variety became mentioned because the 3Vs of massive data, a thought Gartner popularized after acquiring Meta Group and hiring Laney in 2005. Adopting the massive Data analytics and permeating it into the prevailing banking sector workflows is one among the key elements of surviving and prevailing within the rapidly evolving business environment of the digital millennium.

Hadoop is an open-source, big data storage and distributed processing framework which was launched as an Apache open source project, planting the seeds for a clustered platform built on top of commodity hardware and geared to run big data applications. By 2011, big data analytics began to acquire a firm hold in organizations and therefore the limelight , alongside Hadoop and various related big data technologies that had sprung up around it. Initially, as the Hadoop ecosystem took shape and began to grow, big data applications were primarily the province of huge internet and e-commerce companies like Yahoo, Google and Facebook, also as analytics and marketing services providers. In the ensuing years, though, big data analytics has increasingly been embraced by retailers, financial services firms, pirate bay insurers, healthcare organizations, manufacturers, energy companies and other enterprises

II. HOW IT WORKS.

Hadoop clusters and NoSQL systems are used significantly as landing pad and staging area for data before it gets loaded into a data warehouse or

analytical database for analysis which is in a summarized form that is conducive to relational structures.

However, big data analytic users are adapting the concept of a Hadoop data lake that serves as the primary repository for incoming streams of raw data. In such architectures, the analysis of data can be done directly in a Hadoop cluster or a processing engine like Spark. As in data warehousing, data management is a significant first step in the big data analytics process. Data which is stored in the Hadoop Distributed File System must be organized, configured and separated properly to get good performance out of both extracts, transform and load (ETL) integration jobs and analytical queries.

Once the data is completely ready, it can be analysed with the software which is used for advanced analytics processes. That includes tools for data mining, which sift through data sets in search of patterns and relationships; predictive analytics, which build models to forecast customer behaviour and other future developments; machine learning, which uses algorithm to analyse large data sets; and deep learning, a more advanced offshoot of machine learning.

Text mining and statistical analysis software also plays an important role in the big data analytics process, as can mainstream BI software and data visualization tools. In ETL process as well as analytic applications, queries can be written in MapReduce, with programming languages such as R, Python, Scala, and SQL, the standard languages for relational databases that are supported via SQL-on-Hadoop technologies.

III. ADVANTAGES

BD offers a number of advantages to both banks and their customers. Advantages of BD in terms of functional and business area are given in table I. Some of the world wide accepted advantages of applying BD for banking in India are as follows:

● **Fraud Detection and Prevention**: It is one of the major problems faced by the financial sectors and BD can ensure banks that no unauthorized transactions and access will be made from their systems, providing a level of safety and security that will raise the security standard of the entire financial industry.

● **Customer Segmentation**: customer base into groups of individuals that are similar in particular ways relevant to marketing and business, such as their age, gender, financial conditions, interests and spending habits. This segmentation allows banks to provide or deliver to customers with exactly what they're looking for.

● **Risk Management**: The early detection of fraud is a large and major part of risk management and BD can do as much for risk management, as it does for fraud identification. Big data is located and presented on a single large scale that makes it simpler to reduce the number of risks to a manageable number. This would further reduces the chances of losing data or ignoring frauds within transactions in banks.

● **Study of Indian Economy**: Similar to what financial organizations and banks are doing in other countries such as the U.S.A., the techniques can be applied in India for studying the Indian economy more efficiently, and can help in improvising it to a better level.

● **Past Data and Future Predictions**: Banks can also look at the past data, they have already stored, and can plan for the future. BD helps them in spotting patterns in different domains of their services provided to their customers and can use these data patterns to predict their businesses future e.g. where and how to invest their labor, money and time for profitable returns.

IV. BIG DATA ANALYTICS USES & CHALLENGES

Big data analytics applications often include data from both internal systems and external sources, such as predictive data, descriptive data, prescriptive data or diagnostic data on consumers compiled by third-party information services providers. In addition, streaming data flow analytics applications are becoming common in big data firms as users look to perform real-time analytics on data fed into Hadoop systems through stream processing engines, such as Spark, Samza and Storm, Apex, Flume.

Early big data systems were mostly deployed on premises, particularly in large firms that collected, organized, visualized and analysed massive amounts of structured and unstructured data. But cloud platform providers, such as Amazon Web Services (AWS) and Microsoft has made it easier to set up and manage Hadoop groups in the cloud, as have Hadoop suppliers such as MapR and IBM Infosphere, which support their distributions of the big data framework on the AWS and Microsoft Azure clouds. Users can spin the groups in the cloud, run them for as long as they need and then take them offline with usage-based pricing that doesn't require in progress software licenses.

Potential pitfalls of massive data analytics initiatives include a scarcity of internal analytics skills and therefore the high cost of hiring experienced data scientists and data engineers to fill the gaps.

Recently, the proliferation and advancement of AI and machine learning technologies have enabled suppliers to provide software for big data

analysis that is easier to use, particularly for the growing citizen data scientist population. Some of the leading suppliers in this field include Alteryx, IBM, Microsoft and Knime.

The massive amount of data that's typically associated with, and its variety can cause data management issues in areas including data quality, consistency and governance. Also, data warehouses can result from the use of different platforms and data stores in a big data architecture. In addition, the integration of Hadoop, Storm or Spark and other big data tools into a cohesive architecture that meets an firm's big data analytics needs is a challenging proposition for many analytic teams, which have to identify the right mix of technologies and then put the pieces together.

V. TOOLS AND TECHNOLOGIES

We all know that Unstructured and semi-structured data types don't fit in traditional data warehouses that are based on relational databases oriented to structured data sets. Data warehouses does not able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually, as in the case of real-time data in stock trading, the online activities of website visitors or the performance of mobile applications. As a result, many of the organizations that collect, process and analyse big data turn to NoSQL databases, as well as Hadoop and its companion tools, including:

- **YARN**: a cluster management technology and one of the key features in second-generation Hadoop.
- **MapReduce**: a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
- **Spark**: an open source, parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.
- **HBase**: a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
- **Hive**: It is a data warehouse infrastructure tool for querying and analysing large data sets stored in Hadoop files.
- **Kafka**: a distributed publish/subscribe messaging system designed to replace traditional message brokers

VI. APPLICATION

Fraud Detection

Big data analytics have become an essential part of any strategy to help detect and prevent financial crime, owing to the ever-evolving attack methods used by criminals exploiting multichannel vulnerabilities to compromise technology systems. Big data has enabled banks to implement real time analytics on a large scale to meet the growing threats. Using data mining fraud detection techniques that detects both known and novel fraud instances as they occur in real time, with a higher level of accuracy using distributed Ha- doop-based platforms that make it possible to cost-effectively and efficiently store and process large data sets. [3]

VII. DATA MINING TECHNIQUES IN BIG DATA

A. Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection are particularly well suited to this type of analysis.

B. Clustering: Clustering can be said to be the identification of similar classes of objects. In this technique, transactions with similar behaviour are combined into one group. For instance: The customer of a given geographic location and of a particular job profile demand a particular set of services, like in banking sector the customers from the service class always demand for the policy which ensures more security as they are not intending to take risks, likewise the same set of service class people in rural areas have the preferences for some particular brands which may differ from their counterparts in urban areas. This technique will help the management in finding the solution of 80/20 principle of marketing, which says: 20% of your customers will provide you with 80% of your profits, then the problem is to identify those 20% and the techniques of clustering will help in achieving the same.

C. Association Rule: The central task of association rule mining is to find sets of binary variables that co-occur together frequently in a transaction database. Association rule has several algorithms like: APRIORI, CDA, and DDA. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository.

D. Prediction: The prediction as its name implies is one of the data mining techniques that discover the relationship between independent variables and relationship between dependent variables. For instance, prediction analysis techniques can be implemented in the banking sector to predict fraud. Money can be seen as the independent variable while the individual (fraudster) could be seen as the dependent variable. Then based on historical data, we can draw a fitted regression curve that is used for attempted fraud prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables.

Types of Regression Techniques are as follows:
1.Linear Regression
2.Multivariate Linear Regression
3.Nonlinear Regression [6]

VIII. ANOTHER APPLICATION IN BUSINESS SECTOR

The banking industry has evolved by leaps and bounds over the past decade, when it comes to operations and service delivery. The financial and banking data will be one of the cornerstones of this Big Data flood, and being able to process it means being competitive among the banks and financial institutions.

Big data analytics can improve the extrapolative power of risk models used by banks and financial institutions. Big data analytics not only brings new insights to the banks, but it also enables them to stay a step ahead of the game with advanced technologies and analytical tools. It helps the Bank to analyse social media to monitor user sentiment toward a firm, brand or product.

In a highly competitive market, it is driving firms to compete aggressively for customers' wallet: increasing focus on customer acquisition, retention and profitability. While getting a complete view of customer relationship across the enterprise is very important, it is equally essential to use it to offer customized products and service to profitable clients will increase client loyalty and result in increased wallet share & reduce losses by minimizing the risk exposure

1. Efficient Risk Management to Prevent Errors and Frauds

Business intelligence (BI) tools are capable of identifying potential risks associated with money lending processes in banks. With the help of big data analytics, banks can analyze the market trends and decide on lowering or increasing interest rates for different individuals across various regions.

2. Provides Personalized Banking Solutions To Customers

Big data analytics can aid banks in understanding customer behaviour based on the inputs received from their investment patterns, shopping trends, motivation to invest and personal or financial backgrounds. This data plays a crucial role in winning customer loyalty by designing personalized banking solutions for them.

3. Easier Filing of Regulatory Compliances BI tools can help analyse and keep track of all the regulatory requirements by going through each individual application from the customers for accurate validation. **4. Boosts Overall Performance**

With performance analytics, employee performance can be assessed whether or not they have achieved the monthly/quarterly/yearly targets. Based on the figures derived from current sales of employees, big data analytics can determine ways to help them scale better. [5]

IX. CONCLUSION

With respect to the current, highly competitive financial market, big data holds the key to unlocking marketing potential. Advanced analytics are permitting banks to manage the cumulative cost of compliance and the risk of non-compliance. However, the financial service firms are still lagging behind in implementing big data analytic tools, which indicates an untapped potential for value creation, available for the banking industry. This needs to be evaluated from an IT (Information Technology) or LoB (Line of Business) perspective.

REFERENCES

[1]. Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. "Sarcasm as contrast between a positive sentiment and negative situation", In EMNLP 2013, pp. 704–714.

[2]. Subrahmanian, V. S. and Diego Reforgiato. "Ava: Adjective-verb-adverb combi-nations for sentiment analysis", In Intelligent Systems, 23(4):43–50. 2008.

[3]. B. Agarwal, N. Mittal, "Prominent Feature Extraction for Review Analysis: An Empirical Study", In Journal of Experimental and theoretical Artificial Intelligence, 2014, DOI: 10.1080/0952813X.2014.977830.

[4]. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, "Lexicon-based methods for sentiment analysis", Computational Linguistics, v.37 n.2, p.267-307, 2011

[5]. Esuli A., Sebastiani F. "SentiWordNet: A publicly available lexical resource for opinion mining". In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), pages 417–422, 2006.

[6]. P Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of 40th Meeting of the Association for Computational Linguistics, pages 417–424, Philadelphia, PA.

[7]. http://eci.nic.in/eci_main1/GE2014/PC_WISE_TURNOUT.htm

[8]. Mariana Romanyshyn(2013). Rule-Based Sentiment Analysis of Ukrainian Reviews. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013

[9]. S Bandyopadhyay and K Mallick, "A New Path Based Hybrid Measure for Gene Ontology Similarity", IEEE/ACM

Transactions on Computational Biology and Bioinformatics, vol.11, no. 1, pp. 116-127, Jan.-Feb. 2014, doi:10.1109/TCBB.2013.149

[10]. N. Mittal, B. Agarwal, S. Agarwal, S. Agarwal, P. Gupta, "A Hybrid Approach for Twitter Sentiment Analysis", In 10th International Conference on Natural Language Processing (ICON) 2013. pp.116-120.

[11]. A. Bakliwal, J. Foster, J. V. D. Puil, R. O‟Brien, L. Tounsi, M. Hughes, "Senti-ment analysis of political tweets: Towards an accurate classifier". In Proceedings of NAACL Workshop on Language Analysis in Social Media, pages 49–58, 2011

[12]. Di Caro, L., & Grella, M. (2012). Sentiment analysis via dependency parsing. Computer Standards & Interfaces.

[13]. K.. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heil-man, D. Yogatama, J. Flanigan, N.A. Smith. "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", In Proceedings of ACL, 2011.

[14]. Tumasjan, A.; Sprenger, T. O.; Sandner, P.; and Welpe, I. 2010. Predicting elec-tions with twitter: What 140 characters reveal about political sentiment. In Proceedings of ICWSM

[15]. O‟Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM

[16]. Adam Bermingham and Alan F Smeaton. 2011. On using Twitter to monitor po-litical sentiment and predict election results. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pages 2–10,

[17]. Norbert Blenn, Kassandra Charalampidou, and Christian Doerr. Context-Sensitive Sentiment Classification of Short Colloquial Text. In Proc. IFIP‟12, pages 97–108, Prague, Czech Republic, 2012.

[18]. M. De Marneffe, B. MacCartney, C. Manning, Generating typed dependency parse from phrase structure parses, LREC 2006.

[19]. L.K.W. Tan, J.C. Na, Y.L. Theng, K. Chang: Phrase-level sentiment polarity clas-sification using rule-based typed dependencies and additional complex phrases consideration, J. Comput. Sci. Technol., 27 (3) (2012), pp. 650–666

[20]. Jason S. Kessler, and Nicolas Nicolov, "Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations," 3rd International AAAI Conference on Weblogs and Social Media, 2009.

Rahul More, et. al. "Big Data Analysis in Banking Sector." *International Journal of Engineering Research and Applications (IJERA)*, vol.11 (4), 2021, pp 01-05.