IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-5, Jun 2018 E-ISSN: 2347-2693

Credit Risk Management through Big Data Analytics

Deepika Sharma

School of Business, Shri Mata Vaishno Devi University, Jammu, India.

E-mail: deepika82sharma@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** Credit risk remains till date one of the biggest and most challenging issue in the lending financial institutions. Credit risk refers to the probability of default which may occur if the liabilities are not fulfilled under the terms of the contract, resulting into the loss of the financial institution or banks (the creditor). Difficulties in credit risk management arise because the credit default occurs mostly, unexpectedly. The databases of the banks around the world have accumulated large quantities of information about clients and their financial and credit history. These databases can be used for the credit risk assessment, but they are generally high dimensional and traditional data analytics may not be able to handle such large volume of high dimensional data. How to develop a high-performance platform to efficiently analyze the Big Data Analytics (BDA) that can lead to better and more informed credit decisions? This study seeks to answer this question by discussing a macroscopic view of emerging Big Data techniques for addressing the vital issues of credit risk across the various sectors of finance and aim to identify the suitable BDA tools for the purpose of managing Credit Risk.

***Keywords*—**Credit, Big, Risk, Hadoop, Finance

I. Introduction

There are multiple types of risk exposures in financial sector institutions including market risk, operational risk and credit risk, necessitates the existence of a sound risk management framework which has an embedded provision for addressing the credit risk, particularly in case of banks and other firms in the financial services industry (Eccles et al., 2001). Continuous increasing counterparty risk from individuals to sovereign governments along with rising variety of complex derivatives products indicates that credit risk management plays a crucial role in the overall risk management activities carried out by firms in the financial services industry (Fatemi and Fooladi,2006). Credit risk presents the probability of loss that the company (or a lender financial institution) incurs in the event of the counterparty default.

The default may occur if the liabilities are not met under the terms of the contract. These defaults results into losses to the company or to the creditors(Kliestik &Cug,2015). Banks and other financial institutions are particularly undefended against credit risk (Spuchakova & Cug, 2014). Credit risks are the major banking risk among other risks and causes major bank losses. The recent fraud of over eleven thousand crores rupees default by a famous business man at Punjab National Bank has been a classic case of credit risk and its wide ranging repercussions. Financial institutions across the world are facing the problems of outstanding loans that are unlikely to be paid back (bad debts) and financial institutions make use of certain credit models acting as a valuable tool to determine lending decisions and to measure the risks. In recent times credit scoring has become one of the key tool to as certain credit worthiness, ensure collections, reduce possible default risk and make better credit or lending decisions. Now a days more focus has been shifted to credit scoring, resulting into numerous useful techniques and credit scoring models which have been developed by the financial institutions and researchers to evaluate the creditworthiness (lee.et.al., 2002). The objective of credit scoring models are to classify and assign loan applicants to either a 'good credit' group that is likely to repay financial obligation or a 'not worthy credit' group whose application is likely to be denied because of high probability of default on the financial obligations.

This paper is organized into four sections. Section one describes Big Data features with reference to credit risk.Section2 discusses high-performance platform of BDA. Section3 evaluates the use of Big Data in Credit Risk Assessment and finally Section4 discusses the Credit Risk application of BDA tool and the suggested model.

II. Related Work

A. Big Data and Credit Risk

Credit management is the basis of the financial industry and Big Data Analytics provides the helpful insight to process the actionable information (Lin et al., 2015).Majority of banks these days are using big data analytics as a tool in credit risk management. The "Big" in

Big Data tells apart data sets of grand scale that old-style database systems are not able to process adequately. Big data refers to big datasets that can be captured and subjected to an analytics platform to detect patterns, trends and preferences that inform organizations about their customers in a variety of ways and it includes seven V's such as Volume, Velocity, Variety, Variability, Veracity, Visualization and Value (Uddin& Gupta, 2014). The objective of big data management is to ensure the efficacy in big data storage, analytic applications and security. However, many difficulties arise in management of big data (Siddiqa.et.al.,2016). Only advanced data mining and storage techniques can make the storage, management and analysis of enormous data possible. The bigger challenge for researchers and practitioners arise from the exponential growth rate of data, which surpasses the current ability of humans to design appropriate data storage and analytical systems to manage such huge amount of data effectively (Begoli & Horey, 2012).Analytics is now increasingly applied to databases in all fields which revolutionized the ability to identify, understand and predicts the developments. Researchers have paid attention to the credit risk analysis in financial institutions using data analytics and data mining techniques for predicting risk for analyzing the financial data (Srinivasan & Kamalakannan,2017).

B. High Performance Platforms of BDA

With the Development of high-performing platform to efficiently analyze the Big Data Analytics (BDA) leads to more informed credit decisions in terms of parameters like, Memory/storage, Processing landscape for data analytics, Network resources for data analytics, energy considerations and capacity of different platforms for assimilate scaling in different forms (Singh & Reddy,2015).

1) *Memory/storage:* The size of data processing is apparently the most significant factor. If the data fit into the system memory, then clump are typically not required. The whole data can be processed on a single setup. GPU, Multicore CPUs etc. are few of the platforms that can be used to speed up the data processing. If the data doesn't fit into the system memory then other clump choices such as Hadoop, Spark etc. can come across. Hadoop has developed frameworks and tools and though for repetitive tasks it functions slowly, however Hadoop and Spark clusters are efficient in handling big data. The user must decide if (s)he needs to use off-the-shelf tools which are available for Hadoop or if (s)he wants to optimize the cluster performance in which case Spark is more appropriate.
2) *Horizontal Scaling:* Horizontal scaling affect distributing the workload across many servers with precise commodity machines also known as "scale out", where several independent machines are taken collectively in order to increase the processing ability. Some of the categorical horizontal scale out platforms comprises peer-to-peer networks and Apache Hadoop.
3) *Vertical Scaling:* Vertical Scaling affect installing additional processors, additional memory and faster hardware, typically, within a single server. It is also known as "scale up" and it usually includes a single instance of an operating system. Most popular vertical scale up examples are High Performance Computing Clusters(HPC), Multicore processors, Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGA).
4) *Network resources for data analytics:* Batch based processing technologies like 'Apache Hadoop' permit to course large volumes of data. 'Apache Hadoop' is used to do the processing of data intensive applications (Li. et. al., 2013). It uses a Map/Reduce programming model to course a large volume of data (Thusooet.al., 2009).Map/Reduce runs through the divide-and-conquer way of breaking down a matter into many little parts.
5) *Energy considerations in big-data analytics:* Here counting is done augmenting energy utilization in global cloud systems by first analyzing the energy utilization profiles for transport, storage, and provides in different situations from storage, software and processing as a service sides of cloud computing.

C. Discussion on the Big Data techniques for addressing the important issues of credit risk

BDA has been in the forefront of credit risk analysis since last one decade or so to address various issues like huge storage, privacy and many more. A brief description of those issues follows here.

1) *Privacy and Security:* As the expanding rate of data is so fleeting and old methods of encryption algorithms are only applicable to finite amounts of data, new encryption algorithms are needed that are applied to big data confidentiality(Siddiqa.et.al., 2016).For this, improved techniques and technologies are evolved to secure, trace and scrutinize big data processes in terms of infrastructure, application and data. With the studies of (Terzi.et.al.,2015) security and privacy issues for big data under 5 titles as Hadoop security, cloud security, monitoring and auditing, key management and anonymization.
2) *Data Collection& Sharing of Information:* The front end of data-driven supervision is the collection of information about participants in the financial system. Data collection is an exercise in instrumentation of the system, and a key design consideration is the appropriate measurement resolution(Flood.et.al.,2016). The intent behind financial stability of data requires four key aspects which includes coverage, frequency, granularity and

details as data collection of the system is the major component.

3) *Storage and Processing Issues:* The data produced is in such a large amount that the storage is not adequate as well as it requires time to process which is not possible.
4) *Data Modeling and Analysis:* Data analysis, like the other aspects of big data processing, faces escalade issues that create both problems and opportunities(Flood.et.al.,2016). As different sources of big data are unfamiliar to financial econometrics which also limit the acceptable specifications.

D. Technical challenges and suggested model

Data scientists are facing many challenges when dealing with Big Data. One challenge is how to collect, integrate and store, with less hardware and software requirements, tremendous data sets generated from distributed sources (Chen.et.al., 2014b;Najafabadiet. al., 2015a). Efficient methods of data mining are needed to get accurate results, to monitor the changes in various fields and to predict future observations (Oussous.et.al., 2017).

Technical challenges are:

1) *Fault Tolerance*: Data management in efficient form stays a problem for both cloud and big data.
2) *Scalability:* The scalability issue of Big data has lead towards cloud computing, which now adds up to multiple distinct workloads by differ performance goals into abundant batches. It demands sharing of expedient which is costly and also brings various obstacles like how to run and execute different jobs to meet the aim of each workload cost effectively.
3) *Quality of Data*: A cost is involved for the assembling and storage of data. If we use large data for decision making or for predictive analysis in business better results are achieved.
4) *Heterogeneous Data:* Different type of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more are unstructured data. Structured data is arranged into highly mechanized and organized way. Thus structured data is in full integration with database but unstructured data is raw and unorganized.
5) Tools established by various establishments to process and analyze Big Data.
   a) *Hadoop:* Hadoop is an open source project hosted by Apache Software Foundation. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment (T,2009).
   Hadoop mainly consists of:
   - File System (The Hadoop File System)
   - Programming Paradigm (Map Reduce)

   b) *Map Reduce:*It is a programming model for computations on enormous amounts of data and an execution framework for large scale data processing on clusters of commodity servers. It was originally developed by Google and built on well-known principles in parallel and distributed processing (Lin &Dyer,2010).Map Reduce program consists of two functions –Map function and Reduce function.
   c) *HDFS*: "HDFS is a distributed file system Intended to run on top of the local file systems of the cluster nodes and store extremely large files suitable for streaming data access and highly fault tolerant"(Shvachko.et.al.,2010).
   d) *Hive:*"Hive is one of the Hadoop-related sub-projects. Hive offers a warehouse structure in HDFS"(Thusoo.et.al.,2009). Hive gives an SQL-like user interface to query data kept in various databases and file systems that assimilate with Hadoop(Zoltán.et.al., 2011).

III. Methodology

The various component of the suggested model has been derived from the extensive literature survey as illustrated in the following Table no.1

Table:-1

Related Studies	Outcomes
Blazquez & Domenech,2017	Sources of Big Data
Chen.et.al.,2012, Wu & Ding.,2014	Aggregate Clean Enhance Data
Kollár.et.al.,2015, Weng.et.al.,2016	Analytic Model
Abdou & Pointon.,2011. Abdou.et.al.,2011	Credit Report
Prekopcsak.et.al.,2011	Decision tools
Power.et.al.,2015	Credit Decision

A. Data Sources: Secondary Data Sources like website of government agencies and regulators (RBI, SEBI etc.) will be used along with data bases like Prowess and Bloomberg. Data may be sourced from credit rating agencies and credit card companies.

B. Tools and Softwares: Hadoop and Spark are open source data analytics cluster computing framework may be primarily used.

IV. Result and Discussion

The inter-connected model of Credit Risk:

Gathering information is a critical issue in building a credit scoring model. In general, through loan application forms, customer bank account(s), related sector(s), customer credit history, other financial institutions and banks,

market sector analysis and through government institutions, banks may gain competitive advantages by building a robust credit scoring model. Thefig.1. Describes the numerous components and stages of suggested credit risk appraisal model using BDA are:-

A. *Data sources:* Data sources recognized are Internet Web Pages, Discussion Forums, Chats and message common in and among Social Networks, Remote Sensing Networks and all kinds of day to day dealings done through internet based applications,
B. *Aggregate Clean Enhance Data*: Extract, Clean, normalize, transform and Load.
C. *Apply Analytical Model*: Charge of processing and analyzing data or the solution of the equation used to describe changes in a system by mathematical function.
D. *Credit Reports:* Individual's credit history
E. *Credit Scores:* Analysis of a person's credit files by numerical expression
F. *Credit summary Characteristics Report:* Various risk involved in credit management and their controlling parameters.
G. *Indices:* Default swap index
H. *Decision Tools:* Customer management system.

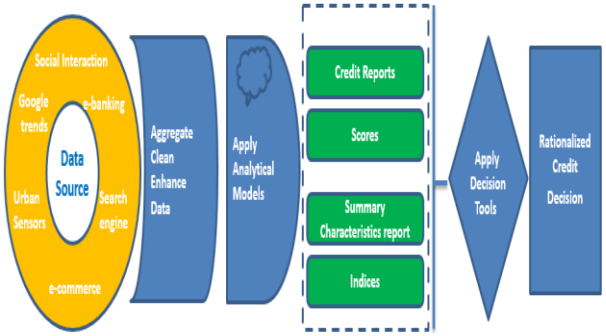

Fig 1: Big Data Credit Risk Decision Model

V. Conclusion and Future Direction

This paper reviews various data processing platforms that are now available and confers some of the popular software frameworks such as Hadoop and Spark. Additionally, a brief dialog has been presented over diverse platforms based on some of the significant features and related vivacious issues of Big Data in finance. The presented credit risk analysis model has been designed using various important tools and features of BDA. However, this is a broad view but a more detailed research over the inclusion of each component in this model needs to be carried out. It would be an excellent research idea to focus over designing a hybrid model for credit assessment so that it adds to the quality of credit risk decision and may arrest the incidences of credit defaults and banking scams.

References

[1] Baliga, J., Ayre, R.W., Hinton, K. and Tucker, R.S., 2011. Green cloud computing: Balancing energy in processing, storage, and transport. Proceedings of the IEEE, 99(1), pp.149-167.
[2] Begoli, E., & Horey, J. (2012). Design principles for effective knowledge discovery from big data. Software architecture (WICSA) and european conference on software architecture (ECSA), 2012 joint working IEEE/IFIP conference on (pp. 215-218). IEEE.
[3] Begoli, E., & Horey, J. (2012). Design principles for effective knowledge discovery from big data. Software architecture (WICSA) and european conference on software architecture (ECSA), 2012 joint working IEEE/IFIP conference on (pp. 215-218). IEEE..
[4] Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y. and Li, L., 2016. Big data analytics and big data science: a survey. Journal of Management Analytics, 3(1), pp.1-42.
[5] Dragosavac,(2015)"Big Data Analytics for LendersandCreditors" https://www.kdnuggets.com/2015/10/big-data-analytics-lenders-creditors.
[6] Eccles, R., Herz, R., Keegan, M. and Phillips, D. (2001), "The risk of risk", Balance Sheet, Vol. 9No. 3, pp. 28-33.energy in processing, storage, and transport, Proc. IEEE 99 (1) (2011) 149–167.
[7] Fatemi, A. and Fooladi, I. (2006), "Credit risk management: a survey of practices", Managerial Finance, Vol. 32 No. 3, pp. 227-233.
[8] Flood, M.D., Jagadish, H.V. and Raschid, L., 2016. Big data challenges and opportunities in financial stability monitoring. Banque de France, Financial Stability Review, 20.
[9] Katal, A., Wazid, M. and Goudar, R.H., 2013, August. Big data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.
[10] Klieštik, T. and Cúg, J., 2015. Comparison of selected models of credit risk. Procedia Economics and Finance, 23, pp.356-361.
[11] Lee, T.S., Chiu, C.C., Lu, C.J. and Chen, I.F., 2002. Credit scoring using the hybrid neural discriminant technique. Expert Systems with applications, 23(3), pp.245-254.
[12] Li, Y., Chen, W., Wang, Y., & Zhang, Z. L. (2013). Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 657–666). ACM.
[13] Lin Z, Whinston AB, Fan S (2015) Harnessing Internet finance with innovative cyber credit management. Financial Innovation, 1(1), p.5.
[14] Lin, J. and Dyer, C., 2010. Data-intensive text processing with MapReduce. Synthesis Lectures on Human Language Technologies, 3(1), pp.1-177.
[15] Mashanovich,(2017)"Using Big Data and Predictive Analytics for Credit Scoring" https://dzone.com/articles/using-big-data-and-predictive-analytics-for-credit.
[16] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald R.,Muharemagic, E., 2015a. Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), p.1.
[17] Oreski, S., Oreski, D. and Oreski, G., 2012. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert systems with applications, 39(16), pp.12605-12617.

[18] Oussous, A., Benjelloun, F.Z., Lahcen, A.A. and Belfkih, S., 2017. Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences.

[19] Prekopcsak, Z., Makrai, G., Henk, T. and Gaspar-Papanek, C., 2011, June. Radoop: Analyzing big data with rapidminer and hadoop. In Proceedings of the 2nd RapidMiner community meeting and conference (RCOMM 2011) (pp. 1-12).

[20] Shvachko, K., Kuang, H., Radia, S. and Chansler, R., 2010, May. The hadoop distributed file system. In Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on (pp. 1-10). IEEE.

[21] Siddiqa, A., Hashem, I.A.T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A. and Nasaruddin, F., 2016. A survey of big data management: Taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71, pp.151-166.

[22] Singh, D. and Reddy, C.K., 2015. A survey on platforms for big data analytics. Journal of Big Data, 2(1), p.8.

[23] Spuchľaková, E., & Cúg, J. (2014). Lost Given Default and the Credit Risk. Proceedings of ICMEBIS 2014 International Conference on Management, Education, Business, and Information Science, Shanghai, China, EDUGait Press, Canada (pp. 12-15).

[24] Srinivasan, S. and Kamalakannan, T., 2017. Multi Criteria Decision Making in Financial Risk Management with a Multi-objective Genetic Algorithm. Computational Economics, pp.1-15.

[25] Terzi, D.S., Terzi, R. and Sagiroglu, S., 2015, December. A survey on security and privacy issues in big data. In Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for (pp. 202-207). IEEE.

[26] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., et al. (2009). Hive:a warehousing solution over a map-reduce framework. Proceedings of the VLDBEndowment, 2(2), 1626–1629.

[27] Uddin, M.F. and Gupta, N., 2014, April. Seven V's of Big Data understanding Big Data to extract value. In American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the (pp. 1-5). IEEE .Vol. 25 Issue: 4, pp.422-434.

[28] White, T., 2009. Hadoop: the definitive guide: the definitive guide:"O'Reilly Media, Inc.".

