International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Big Data Analytics Techniques for Credit Card Fraud Detection: A Review

M. Sathyapriya[1], Dr. V. Thiagarasu[2]

[1]Assistant Professor of Computer Science, Gobi Arts & Science College (Autonomous) Gobichettipalayam, India

[2]Associate Professor of Computer Science, Gobi Arts & Science College (Autonomous) Gobichettipalayam, India

Abstract: ***Due to rapid advancement in internet technology, the use of credit cards has dramatically increased and it leads to increase in number of credit card frauds. The enormous collection of data due to human dependence on computers and automated system is not only helpful for researchers but equally valuable to investigators who intend to carry out forensic analysis of data associated with the variety of criminal cases. The conventional methodologies of performing forensic analysis have changed with the emergence of big data because forensic with big data requires more sophisticated tools along with the deployment of efficient frameworks. This paper presents a survey of techniques used in credit card fraud detection and this work provides a comprehensive review of forensic techniques to detect credit card frauds. Based on analysing the factors such as processing speed, latency, fault tolerance, performance and scalability, an evaluation is made about the techniques and proposed that Apache Spark is performing better for implementation of credit card fraud detection system when compared to other techniques.***

Keywords: Cyber crime, Big Data Analytics, Fraud detection, Apache Hadoop, MapReduce, Apache Spark and Apache Flink

1. Introduction

Cyber crime is any kind of crime that can be done in, with, or against networks and computer systems [10] [37].

Cyber crime is getting increased with the increasing threats due to online frauds and unethical hacking. With both information and cyber safety threat increasing, organizations must be ready to equip themselves with predicting and preventing cyber crime. Cyber crime experts are using big data tools to identify the potential threats and detect cyber crime incidents like credit card frauds. Big data analytics is enabling companies to analyze voluminous amount of data they gather during financial transactions, any locale-specific data and others as well. Fighting cyber crime is of utmost importance today due to increased risk of cyber theft. Big data tools are being used to fight cyber attacks. Big Data analytics can help detect fraud and identify theft and can facilitate digital forensic analysis. In this paper, a brief survey is made about various techniques used in analysing big data to detect the frauds related to credit cards by analysing large set of data. One aim of this study is to identify the user model that best identifies fraud cases.

2. Big Data – An Introduction

The world is becoming digitalized and interconnected so that the amount of data has been exploding every minute. To manage those data records, it requires extremely powerful business intelligence. The problem starts during data acquisition, when the large amount of data require us to make decisions about what data to keep, what to discard and how to store, so that data can be kept reliable and accurate. Big data refers to datasets whose size is beyond the ability of typical database software to capture, store, manage and analyze [23] [24] [30]. It can be described as a massive volume of both structured and unstructured data which can't be stored using traditional databases. An example of big data might be petabytes or exabytes or even zettabytes of data which consists of billions to trillions of records that are collected from millions of people all from different sources.

The sources of data may come from web, sales, customer contact centre, social media, mobile data etc. Big Data is a term associated with large datasets that come into existence with the features of volume, variety, velocity and veracity of data. Data variability, value and complexity are some other features that are used with big data [36] [24].
Volume: Vast amount of stored data that can be gathered and analyzed effectively.
Variety: Type of data that may be structured, unstructured, transactions, video/audio, text and log files.
Velocity: Rate of data change and speed of data available for analysis.
Veracity: Data integrity and extend of trust in the data to confidently use it to make decision.

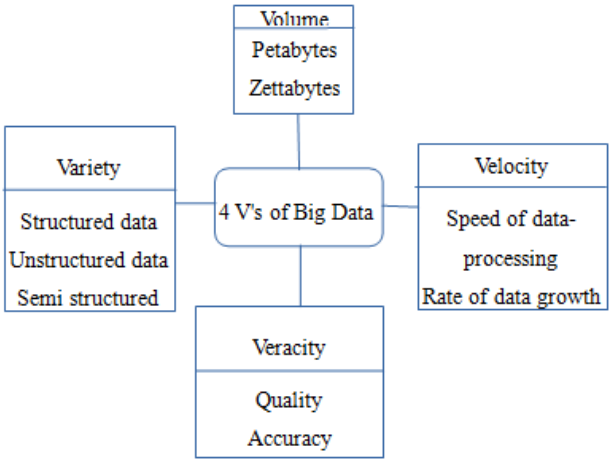

Figure 2.1: Big Data Properties [24]

In the context of credit card transaction analysis, volume corresponds to the thousands of credit card transactions that occur every second in every day. Variety refers to the type of data that are used in transaction process. Velocity refers to how quickly data can be processed for analytics. Veracity deals with analyzing the credit card transactions to make

decisions on it with the aim of finding fraudulent transactions if any. These factors are important for analyzing transactions to find frauds and taking immediate corrective action.

3. Big Data Analytics

Big data analytics is a technology that searches useful information such as a relation rule, a hidden value from huge data [28]. When data volumes reach big data proportions, parsing it for meaningful information requires very powerful data analytics. The domain of Big Data Analytics is concerned with the extraction of value from big data which are significant, previously unknown, implicit and potentially useful. These insights have a direct impact on making a useful decision from the interpreted data [17]. With the help of right analytical tools, big data can detect various frauds done with credit cards. The analytics tools perform the following activities:

- Collects data from many enterprise sources.
- Performs a deeper analytics on the data.
- Provides a fine view of security information.
- Achieves real-time analysis of streaming data.

4. Big Data Analytics in Cyber Crime

Big Data Analytics in cyber security involves the ability to gather massive amount of digital information to analyze, visualize and draw insight that can make it possible to predict and stop cyber attacks. Detecting fraud quickly requires real-time analysis of many structured and unstructured data sources. Fraud detection is one of the most visible uses for Big Data analytics. Most of the frauds are high-volume in nature. So a good opportunity is provided for analytics to identify patterns from high volume data and recommend preventive action. Many of the techniques employed to detect frauds requires recognizing identical/repeating pattern matches of people, places, systems, and events.

Compared to traditional approaches, big data analytics provides an efficient cyber security context by separating what is “normal” from what is “abnormal”, i.e., separating the patterns generated by authorized users from those generated by suspicious or malicious users. By providing means to discover changing patterns of malicious activities hidden deep in large volumes of organizations data, big data tools can indeed empower businesses to better understand if and how they have been attacked [15]. Big Data can be associated with the following fraud detection techniques:

Unsupervised Learning/Descriptive Analytics: Aims at finding the behaviour that deviates from normal behaviour or at detecting anomalies. These techniques learn from historical observations and do not require these observations as fraudulent or a non fraudulent activity.

Supervised Learning/Predictive Analytics: Aims to learn from historical information in order to retrieve patterns that allow differentiating between normal and fraudulent behaviour. This analytics can be applied to detect fraud as well as to estimate the amount of fraud.

Social Network Analytics: Aims at extending the ability by detecting the fraudulent behaviour in a network of linked entities. It also finds the relationship between entities by uncovering particular patterns indicating fraud.

5. Literature Review

Anushree Naik and Kalyani Phumamdikar [2016] suggested that using a Bayes minimum risk classifier, it gives rise to much better fraud detection results in the sense of higher savings. This architecture used HDFS for storing and fast accessing of the user logs [7].

K. R. Seeja and Masoumeh Zareapoor [2014] proposed a credit card fraud detection model that detects fraud from highly imbalanced and anonymous credit card transaction datasets. To find whether the incoming transactions of the customers belong to legal or illegal pattern, a matching algorithm is proposed and according to that transaction closer to the patterns is identified and decisions are made [21].

Bolton et al., [2001] presented a clustering based technique that analyses the user behaviour for detecting frauds. An alarm is triggered when a transaction violates the regular spending pattern [27].

Weston et al., [2008] presented a peer group based fraud detection method. This technique uses grouping behaviour of the transactions in identifying anomalies [12].

Aditya B. Patel et al., [2012] reports the experimental work on the Big data problems. It describes the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets [1]

6. Big Data Fraud Detection Techniques

A formal forensic investigation cannot be launched until meaningful and relevant data is not extracted from the entire data set [2]. The focus is on the different techniques that can facilitate the forensic investigator in analyzing the big data to find the underlying relationship among the data. Furthermore these techniques help the investigator in extracting meaningful and purposeful forensic evidence for detecting frauds from the large datasets:

A. Apache Hadoop

Apache Hadoop, the popular data storage and analysis platform, an open-source software framework for distributed storage of very large datasets on computer clusters. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks [26] [30] [33]. The key component of Hadoop is the Hadoop Distributed File System (HDFS), which manages the data spread across the various servers. It is because of HDFS that so many servers can be managed in parallel. HDFS is file based and does not need a data model to store and process data. HDFS can manage the storage and access of any type of data (e.g., Web logs, XML files) as long as the data can be put in a file and

copied into HDFS [13].

Hadoop offers two important services: It can store any kind of data from any source, inexpensively and at very large scale; and it can do very sophisticated analysis of that data easily and quickly. Hadoop stores terabytes, and even petabytes, of data inexpensively. Online businesses use Hadoop to monitor and fight criminal behaviour. Sites that sell goods and services over the internet are particularly vulnerable to fraud and theft. Hadoop is a powerful platform for dealing with fraudulent and criminal activity like this. It is flexible enough to store all of the data-message content, relationships among people and computers, patterns of activity etc. It is powerful enough to run sophisticated detection and prevention algorithms and to create complex models from historical data to monitor real-time activity [34].

Hadoop's advantages [11] include:
- A distributed functionality, making networks more robust, if one cluster fails, it continues to run.
- Efficient, doesn't require applications to send huge amounts of data across the network.
- Provides linear scaling in the ideal case, enabling easier design.
- HDFS can store a lot of data.

B. MapReduce
MapReduce is a programming paradigm for processing large datasets in distributed environments [19][20]. MapReduce is a simple programming model for processing huge data sets in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and assembles the results into a smaller, easier to analyze file. The basic notion of MapReduce is to divide a task into subtasks, handle the sub-tasks in parallel, and aggregate the results of the subtasks to form the final output.

Programs written in MapReduce are automatically parallelized so programmers do not need to be concerned about the implementation details of parallel processing. Instead, programmers write two functions: map and reduce. The map phase reads the input (in parallel) and distributes the data to the reducers [9]. Auxiliary phases such as sorting, partitioning and combining values can also take place between the map and reduce phases. MapReduce programs are generally used to process large files. The input and output for the map and reduce functions are expressed in the form of key-value pairs [14]. MapReduce is also used as a feasible technique to detect credit card frauds effectively [6] [16].

C. Apache Spark
Apache Spark is an open source cluster computing system that can be programmed quickly and runs fast. Spark relies on "resilient distributed datasets" (RDDs) and can be used to interactively query 1 to 2 terabytes of data in less than a second [9] [29]. It is a cluster computing framework that makes data analytics faster to run and also to write to distributed file systems. RDDs are collections of objects spread across a cluster and stored in RAM or on disk [22]. The Spark programs use a very similar map/reduce task that are used in Hadoop to run analytics on the large data sets. The difference is that Hadoop map/reduce programs are mainly used to run in batch mode, whereas Spark programs can be used to run on real time data as well as batch mode since it uses the concepts of caching data in memory.

Spark provides an easier to use alternative to Hadoop, MapReduce and offers performance up to 10 times faster than previous generation systems for certain applications [35]. Spark can simply be used as a framework to process data or to run machine learning algorithms and HDFS can be used to store the data. Combining these two frameworks together provides opportunities to solve credit card fraud like using big data analytics. In Spark Streaming if a worker fails, the system can recomputed to the lost state from the input data by following all the RDD transformations that preceded the point of failure [31].
Spark's advantages [11] include:
- Integrated advanced analytics.
- Use of data parallel processing.
- More efficient than MapReduce.
- Continuous micro-batch processing based on its own streaming API.
- Significantly faster than Hadoop, MapReduce for certain use cases.

D. APACHE FLINK
Apache Flink is an open source stream processing framework developed by the Apache Software Foundation. The core of Apache Flink is a distributed streaming dataflow engine written in Java and Scala [4] [5]. Flink executes arbitrary dataflow programs in a data-parallel and pipelined manner [3]. Flink's pipelined runtime system enables the execution of bulk/batch and stream processing programs [18] [8]. Flink provides a high-throughput, low-latency streaming engine [31] as well as support for event-time processing and state management. Flink applications are fault-tolerant in the event of machine failure and support exactly-once semantics [25].

Flink provides a true data streaming platform that uses high-performance data flow architecture. With its strong compatibility option, Flink enables developers to use their existing processes on MapReduce, Storm, etc. directly in Flink's execution engine. It is known for processing big data quickly with low data latency and high fault tolerance on distributed systems on large scale. Its defining feature is ability to process streaming data in real time.

Apache Flink includes a lightweight fault tolerance mechanism based on distributed checkpoints [25]. A checkpoint is an automatic, asynchronous snapshot of the state of an application and the position in a source stream. In the case of a failure, a Flink program with check pointing enabled will, upon recovery, resume processing from the last completed checkpoint, ensuring that Flink maintains exactly-once state semantics within an application.

Flink's advantages [11] include:
- A true stream processing framework.
- The use of algorithms in both streaming and batch modes.
- An aggressive optimization engine.

- Speedier processing.

7. Evaluation of Big Data Fraud Detection Techniques

This section extends the analysis of the techniques proposed previously for the big data forensic investigation. There are different factors that can influence the selection and performance of the forensic techniques. These factors should be analyzed closely before carrying out the implementation. The following table has been populated based on the available resources and the sensitivity of the data to be used for investigation.

Table 7.1: Analysis of Techniques and its Factors

Factors / Techniques	Processing Speed	Latency	Fault Tolerance	Performance	Scalability
Hadoop	Medium	High	High	Slow	Medium
MapReduce	Slow	High	High	Slow	Medium
Spark	Fast	Low	High	Fast	High
Flink	Fast	Low	High	Fast	High

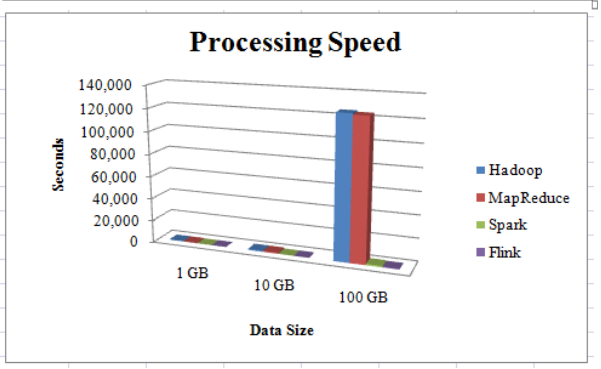

Figure 7.1: Processing Speed

In Figure 7.1, it can be seen that the processing speed of Apache Spark and Flink remains the same even though the data size increases. But Hadoop and MapReduce increases its processing time if the data size is too big to handle.

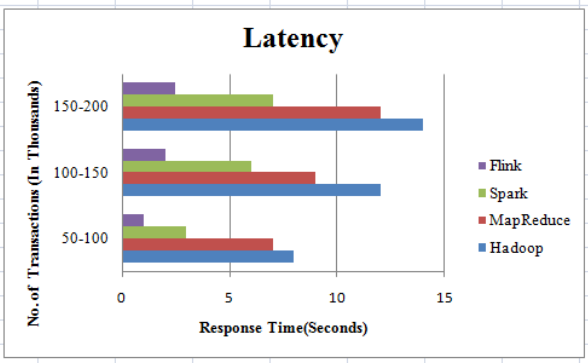

Figure 7.2: Latency

The response time should be at minimum while executing the data/transactions. Among the four techniques, Apache Flink has low latency while processing Big Data.

While analysing the Fault tolerance factor, Spark system replicates the input data in memory which is a most useful solution for handling faults in between the execution of transactions. Data lost due to failure can be recomputed from replicated input data.

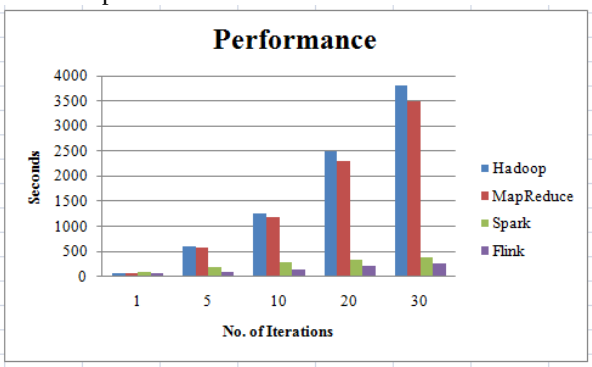

Figure 7.3: Performance

From the Figure 7.3, it is noted that Spark has small variability in the execution time when compared to other techniques.

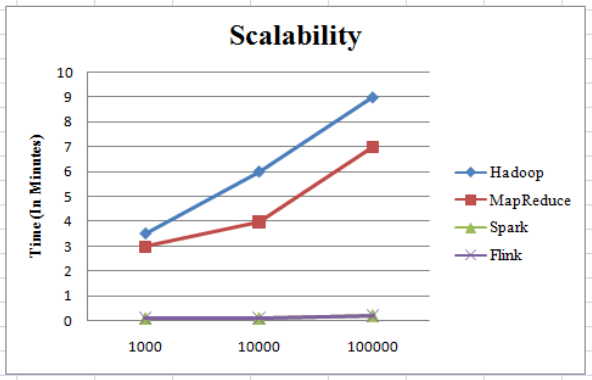

Figure 7.4: Scalability

With scalability factor, Spark and Flink processes the data smoothly if there is need to increase nodes for processing Big Data. Although the number of node increases, Spark performs better with Big Data.

8. Conclusion

In this paper, the techniques such as processing speed, latency, fault tolerance, performance and scalability that were previously being used in a different context have been analyzed that can facilitate the forensic investigator in performing big data forensic. In this work, all the four techniques have been evaluated on the basis of factors such as processing speed, latency, fault tolerance, performance and scalability and spark is suggested as a better performing technique among others. Combining the results, Apache Spark is considered as an efficient technique that helps to implement credit card fraud detection system.

References

[1] Aditya B. Patel, Manashvi Birla, Ushma Nair , "Addressing Big Data Problem Using Hadoop and Map Reduce", International Conference on Engineering, 2012.
[2] Alessandro Guarino, "Digital Forensic as a Big Data Challange", ISSE Securing Electronic Business Processes, 2013.

[3] Alexander, Rico Bergmann, Stephan, "The Stratosphere Platform for Big Data Analytics", the VLDB Journal, 2014.
[4] "Apache Flink" [Online], Available: http://flink.apache.org/.
[5] "Apache Flink" [Online], Available: http://github.com/apache/flink
[6] R.Anbuvizhi, V.Balakumar, "Credit / Debit Card Transaction Survey Using Map Reduce in HDFS and Implementing Syferlock to Prevent Fraudulent",International Journal of Computer Science and Network Security, 2016.
[7] Anushree Naik, Kalyani Phulmamdikar, Shreya Pradhan, Sayali Thorat, Prof. Sachin V. Dhande, "Real time Credit card transaction analysis", International Engineering Research Journal (IERJ), Vol 1, Issue 11, 2016.
[8] "Blog: Apache Flink" [Online], Available: www.odbms.org/blog/2015/06/on-apache-flink-interview-with-volker-mark1/
[9] Brian Ye, Anders Ye, "Exploring the Efficiency of Big Data Processing with Hadoop MapReduce", School of Computer Science and Communication (CSC), Royal Institute of Technology KTH, Stockholm, Sweden.
[10] Cameron S.D. Brown, "Investigating and Prosecuting Cyber Crime: Forensic Dependencies and Barriers to Justice", International journal of Cyber Criminology, 2015.
[11] "Comparing Hadoop, MapReduce, Spark, Flink, and Storm" [online], Available: http://www.metistream.com/comparing-hadoop-mapreduce-spark-flink-storm/
[12] D. J. Weston, D. J. Hand, N. M. Adams, C. Whitrow, P. Juszczak, "Plastic card fraud detection using peer group analysis", Advances in Data Analysis and Classification, 2008.
[13] Dr. Hugh J. Watson, "Big Data: Concepts, Technologies and Applications" [Online], Available: http://www.watson-tutorial-big-data-business-analytics-collaborative.pdf/
[14] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M, " Efficient Analysis of Big Data Using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, Vol.2, 2014.
[15] Dr. Tariq Mahmood and Uzma Afzal, "Security Analytics: Big Data Analytics for Cyber security A Review of Trends, Techniques and Tools", 2nd National Conference on Information Assurance (NCIA) 2013.
[16] Elham Hormozi, Mohammad Kazem Akbari, Hadi Hormozi, "Accuracy evaluation of a credit card fraud detection system on Hadoop MapReduce", The 5th conference on Information and Knowledge Technology, IEEE Publications, 2013.
[17] "Gartner Report: Big Data will Revolutionize Cyber Security in the Next Two Years" [online], Available:http://cloudtimes.org/2014/02/12/gartner-report-big-data-will-revolutionize-the-cybersecurity-in-next-two-year/
[18] "Hadoop, MapReduce, Spark, Flink" [Online], Available: www.infoworld.com/article/2919602/hadoop/flink-hadoops-new-contender-for-mapreduce-spark.html.
[19] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters", Communications of the ACM, vol 51, pp. 107-113, 2008.
[20] Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux, David S. Allison, and Miriam A.M. Capretz, "Challenges for MapReduce in Big Data", Proc. of the 10th 2014 world congress on services, 2014.
[21] K. R. Seeja and Masoumeh Zareapoor, "Fraud Miner: A Novel Credit card fraud detection model based on Frequent Item set Mining", The Scientific World Journal, 2014.
[22] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica, "Spark: cluster computing with working sets", HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 2010.
[23] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics", Journal of Big Data, 2015.
[24] Palak Gupta, Nidhi Tyagi, "An Approach towards Big Data –A Review", International Conference on Computing, Communication and Automation (IEEE), 2015.
[25] Paris Carbone, Gyula, Stephen, Seif,Kostas, "Light weight Asynchronous Snapshots for Distributed data flows", Cornell University Library, 2015.
[26] Palak Gupta, Nidhi Tyagi, "An Approach towards Big Data –A Review", International Conference on Computing, Communication and Automation (IEEE), 2015.
[27] R.J.Bolton, D.J. Hand, "Unsupervised profiling methods for fraud detection", Proceedings of the VII Conference on Credit Scoring and Credit Control, 2001.
[28] R. Magoulas and B. Lorica, "Introduction to Big Data", Release 2.0, Issue 11, Feb 2009.
[29] "Real-Time Big Data Analytics: Emerging Architecture", Mike Barlow, O'Reilly media, 2013.
[30] Shahzaib Tahir, Waseem Iqbal, "Big Data−An Evolving Concern for Forensic Investigators", IEEE Transactions, 2015.
[31] "Spark Streaming Programming Guide" [Online], Available: http://spark.apache.org/docs/1.1.1/streamingprogramming-guide.html.
[32] "Streaming Data with Apache Flink" [Online], Available: http://yahooeng.tumblr.com/past/135321837876/benchmarking and streaming-computation-engines/
[33] T. Giri Babu, Dr. G. Anjan babu, "A Survey on Data Science Technologies & Big Data Analytics", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 6, Issue 2, Feb 2016.
[34] "Ten Common Hadoopable Problems-Real World Hadoop Use Cases" [Online], Available: http://blog.cloudera.com/wp-content/uploads/2011/03/ten-common-hadoopable-problem-find.pdf

[35] "The Big-Data Ecosystem Table" [Online], Available: http://spark.incubator.apache.org/
[36] X. Chen, S. Member, and X. Lin, "Big Data Deep Learning: Challenges and Perspectives", IEEE Access, Vol 2, 2014.
[37] Z. Spalevic, Zaklina, "Cyber Security as a global challenge today", Singidunum Journal of Applied Sciences, 2014.