Big Data and IoT: A Prime Opportunity for Banking Industry

Abdeljalil Boumlik(✉) and Mohamed Bahaj

Faculty of Sciences and Technologies of Settat, LITEN Laboratory,
Hassan 1st University, Settat, Morocco
boumlik.abdeljalil@gmail.com, mohamedbahaj@gmail.com

Abstract. Banking industry is one of the most complex and sensitive industries that experience enormous changes in daily basis. Likemany others businesses, Big data is a serious problematic, data management and real time monitoring fraud issues also are even bigger challenges in this sector, due to the huge quantity of data, coming swiftly and rapidly from different devices in structured and unstructured formats, waiting for instantaneously treatments and decisions. Most financial institutions and banks try to innovate and diversify payment processes to make it more challenging and secure to improve their digital skills. Understand customer's behaviors also become a successful key factor in the market at the same time, that's why Internet of Things (IoT) can be the best solution to solve the issue of collecting and sharing data via internet among different "things", as devices and objects (Sensors, ATMs, POS, Smartphones, Computers, payment gateways (ecommerce), notebooks, etc.). The architectural and technical sides remain a problem, since conventional database management system and existing banking systems are not capable anymore to handle, store and process this massive volume of data with sufficient real time. This paper, discuss Hadoop Distributed File System and MapReduce, as an architecture for storing and retrieving information from massive volumes of datasets that we can collect via Internet from different objects based on the advantage and potential of Internet of things.

Keywords: Big data · Internet of thing · IoT · Hadoop · HDFS · MapReduce · Fraud monitoring

1 Introduction

In recent years, data is increasing at dreadful frequency day after day, therefore manage, explore and visualize this big data is becoming a new challenge. This challenge includes analysis, capture, search, sharing, storage, transfer, querying and information confidentiality, not just about being huge in size, but the aim is how to profit and manipulate the big data we have. Traditional data processing applications are no more capable to deal with it. There is an unlimited interest and benefit to use big data technologies to increase productivity and economy of all business sectors to build efficient investment strategies in different private and public domains. These technologies have a high capability to expect conclusions, with low cost consumption in materiel perspective, increase

M. Ezziyyani et al. (eds.), *Advanced Information Technology, Services and Systems*,
Lecture Notes in Networks and Systems 25, https://doi.org/10.1007/978-3-319-69137-4_35

efficiency and improve decision-making in various areas like banking, finance, fraud control, real time monitoring and transactions processing. Banking and financial sectors consider as one of the most complex and sensitive worlds, who's suffering from scalability and inefficiency problems when processing or analyzing this data, also the performance issues become worse due to huge amount of data that are not be handled anymore by conventional database management system or processed by traditional systems architectures. Fraudulent transactions in real time are increasing each day, as an example BGFI Bank was losing 1.9 billion CFA [10] on 2017 due to credit card fraud. Recognizing large-scale patterns across several transactions and identifying strange behaviors in the banking systems from an individual or multiple users can change, and prevent bank loses as online fraudulent transaction. Therefore, banks specialists adopt important attention to new projects that combine between performance and analytics to prevent such type of major problems that cause financial loss to the banks in real time. From other side, the number of transaction made by cardholder's peer second was extremely increased, and become unmanageable by current traditional systems, thus it causes a delay during the treatment of requests, that become very significant and impact banks business strategies.

Recently, the biggest banks and financial institutions realize that their business is not just around processing transactions, but it's about engaging and distributing value through the transaction lifecycle by offering multiple services via diverse payment channels and several devices like Mobiles, Point of Sale, Contactless, Near Field Communications (NFC), wearables, Tokenization, Biometric chip, Smart Kiosk, Smart ATM in order to get have opportunity to increase their volume of transactions and generate more revenue. For this reason, the Internet of Things (IoT) emerge and offers this potential advantages to benefit from the market dynamics and trends.

This paper is mainly implemented for banking domain, in which we present bank systems limitations and emergence needs for collecting data via different devices through internet and have also a fast access to such huge databases that requires an effective computing model. Based on that, we proposed also, a more sophisticated architecture that integrate Hadoop framework to get a complete system that deals with banks data and online fraud monitoring. Next, improve the architectures scalability and efficiency using big data environment, which is implemented on Hadoop (HDFS) distributed file system and MapReduce model as one of the most generally used parallel computing platforms for processing, stores and retrieve data from massive volume of datasets across multiple devices and sensors. The rest of this paper is organized as follows. Section 2, discusses related works and approaches that cover this common area. Section 3 presents a general view on Hadoop framework, MapReduce model and Internet of Things (IoT), then we highlight the advantages of this technologies Sect. 4, describes briefly the proposed solution from Transaction flow till Fraud detection systems. Implementation and evaluation is presented in Sect. 5. Finally, we conclude this paper, and discuss the future steps of our work.

2 Related Work

We can found several approaches that deal with fraud system in banking, finance, insurance, medical domains [1, 4, 5, 7, 8] depend on the needs of each institutions, but most of them are very simple and cover only limited scenarios or single parties of fraud, hence there is no enfranchisement and new challenges that get evoked. lasselare et al. presented a credit card fraud detection system that uses several intrinsic features to perform the detection process [1]. This technique is based on the buying behavior of the customer. The major parameters utilized are regency, frequency and monetary levels. These properties are used to predict frauds. However, the author was limited only on credit card fraud type and not a global system. A rule based fraud detection method that provided huge improvements in the detection process is described in [6]. This proposes to be a real-time system that has been implemented in a Turkish Bank. A cost sensitive credit card fraud detection method that uses Bayes Minimum Risk classifier is presented in [7]. The author claims to provide realistic views of the monetary gains and losses occurring due to fraud detection. A method that concentrates on providing effective fraud detection using imbalanced data is presented in [8], this solution is considers an extremely sparse and imbalanced data environment for performing the fraud detection process. An evaluation of accuracy provided by the Hadoop MapReduce environment on the Credit Card Fraud detection data is presented in [9], the Negative Selection algorithm is parallelized in the Hadoop environment for determining the accuracy. Authors Mahmoudi et al. presented a Modified Fischer Discriminant Analysis based anomaly detection method [2]. This technique uses Fischer Discriminant function to identify anomalies and Fraud. Halvaiee et al. presented an Artificial Immune System (AIS) based fraud detection model in paper [3]. This technique utilizes AIS to identify legitimate transactions from the fraudulent transactions it also limited to one type of fraud rules. Paper [4] present an ANN (Artificial Neural Networks) technique in the fraud domain based on the machine learning technique. Authors of paper [5] use a technique that utilize several classifiers, groups their results to identify fraudulent transactions.

From above results confirmed that recent solutions that deal with fraud systems, doesn't involve Big Data technologies as it should, none of them include Internet of Things (IoT) or discuss the way that they collect data. There are few articles that use Hadoop and MapReduce and only some of them involved the concept of IoT. We also conclude that these solutions have at least few defects or limitations. Moreover, we take the above as a source of motivation to provide efficient solutions based on IoT and Hadoop from architectural point of view and performance perspective.

3 Internet of Things and Hadoop

3.1 Internet of Things for Banking Domain

The Internet of Things (IoT) represents new opportunity for financial institutions and banks to find out more about what their customers require based on the information revolution, which offers an extraordinary level of data and data-driven customer services. Financial institutions realize that the customer experience becomes a key

differentiator to identify new ways to distinguish themselves in the marketplace, and to deep existing relationship and increase the customers based on today's competitive environment. By applying this technology, banks will provide exceptional services, and adaptable financial solutions and advices, that closely associate with day to day events in customer's lives that will impact positively the bank's revenues and gain many competitive advantages. It is a network of billion devices connected through internet, by doing so, this become an intelligent system of systems. These devices present in Fig. 1 can collect data that allows banks to provide a complete view of customer's finance status in real time. Consequently, banks can anticipate customer's needs through data collected and analyzed, then provides solutions that can helps customers take sound and smart financial decisions.

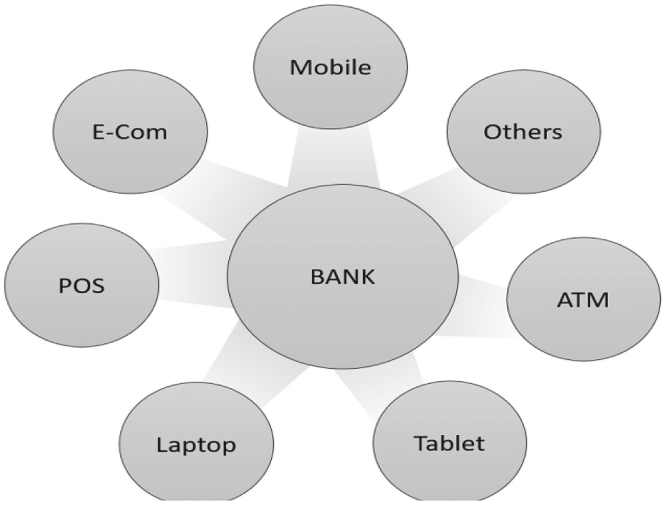

Fig. 1. Different source of data for financial institutions.

3.2 Hadoop Framework to Process Big Data

Hadoop is an open source distributed processing framework from Apache. Developed in Java language, Hadoop used for storing a very large amount of data sets in different huge number of computer clusters. Hadoop's key advantages are related to his ability and flexibility during processing of large scale data, manage and control hardware failures or fault tolerance in the software level, cost effectiveness, scalability and robustness in real time processing. The Hadoop framework core consists of three major components that are Hadoop Distributed File system (HDFS), NameNodes and DataNodes. The Hadoop distributed file splits data and store it in large blocks to different nodes in the cluster system. Therefore, performance, access, operations and visualization of big data will be executed in parallel throw Hadoop HDFS layer, which means data will be processed faster and more efficiently than it would be in any other most conventional super-computer architecture. Hadoop makes replication of data automatically, due to

Master/Slave architecture in which the Master called NameNode and Slave called DataNode. NameNode is the responsible part of managing file system and mapping of files to their considered blocks which knows which data node stores which blocks, manage block replication, store all metadata in the RAM…etc. Otherwise, DataNode is a slave node consist to stores and reads blocks of files on top of native host (file system), also is responsible to forward a stored block to another server on another frame and replicate to a third server. Below we present the HDFS Architecture (Fig. 2).

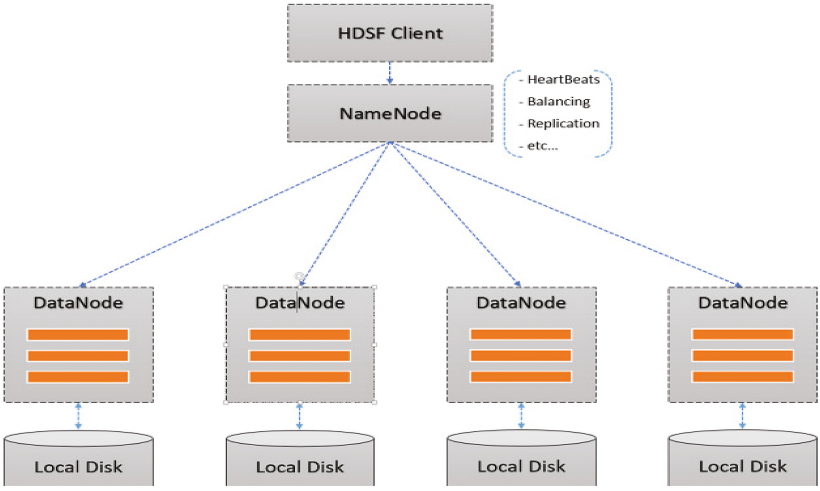

Fig. 2. Hadoop HDFS architecture

3.3 MapReduce Model

In addition to the HDFS, MapReduce is core part of Hadoop. It is a programing model, which allows parallel processing of large volume of data. The MapReduce concept is simple and easy to understand, below is a graphical representation for all logical data flow with key functions (Fig. 3).

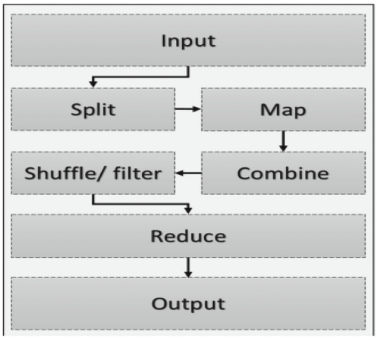

Fig. 3. MapReduce logical flow

The MapReduce jobs contains two major tasks, Map and Reduce are prepared to dividing whole workload into number of tasks and distributing them over different machines in Hadoop cluster. The Map task refers to a job that perform filtering and sorting. It takes input data, create a key/value pairs, and prepare them in a queue, as results and it will be sorted and sent to the Reduce task. In Fig. 4, we illustrate the graphical representation of the Map job flow.

3.3.1 Mapping Phase

See Fig. 4

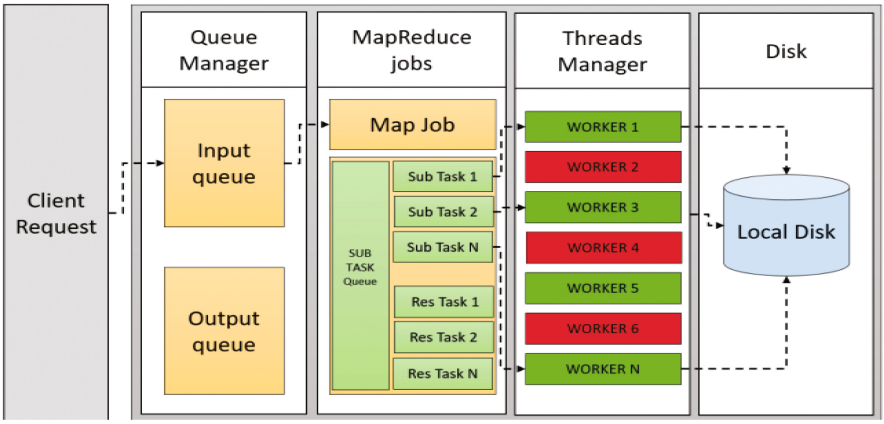

Fig. 4. Map job architecture

3.3.2 Reduce Phase

After all the Map tasks, have completed successfully, the master controller combine and aggregate the results from each Map task and process them to be as a sequence of key-value pairs. At that time, the Reduce job correspond to this request take as input the returned Key and linked values to it as illustrate in the Fig. 5.

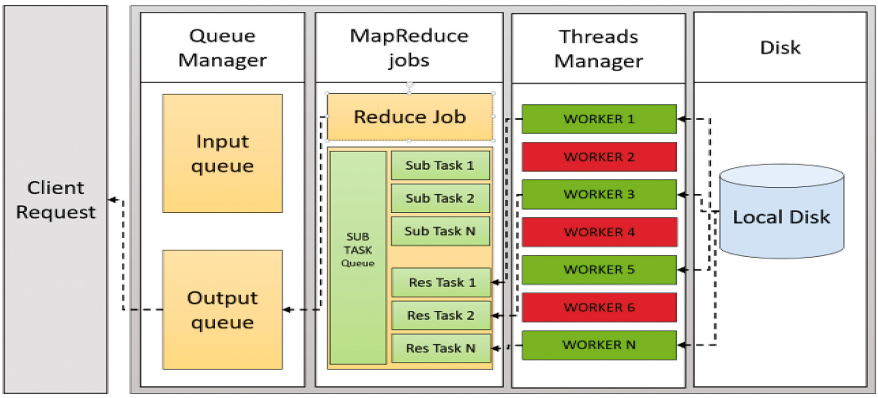

Fig. 5. Reduce job architecture

Also, reduce job, combine the values of the input key by reducing the list of retrieved values corresponding to the initial occurrence. Accordingly, the reduce job generate the output as a set of key-value pairs.

4 Research Design and Methodology

4.1 Overview

Recently, banks and financial institutions reconsidering how they model their enterprises. The statistical modeling and analytics insights become a key role in the industry to improve optimization, forecasting and operation decision process. Therefore, Banks and financial institution continue to focus on revenue progress and higher borders through operational efficiency, and especially better risk management, and improved customer intimacy. All these above and others factors force the innovation and invent solution and software architectures to deal with these challenges with necessity of taking into consideration duration and costs front of advantages to lead in the markets. The banking systems also have the problem of Volume, Velocity and Variety and it's considers as exciting factor which mean one of the most reason behind Big Data innovation to capitalize this data for strategic advantages depends on the niche (Fig. 6).

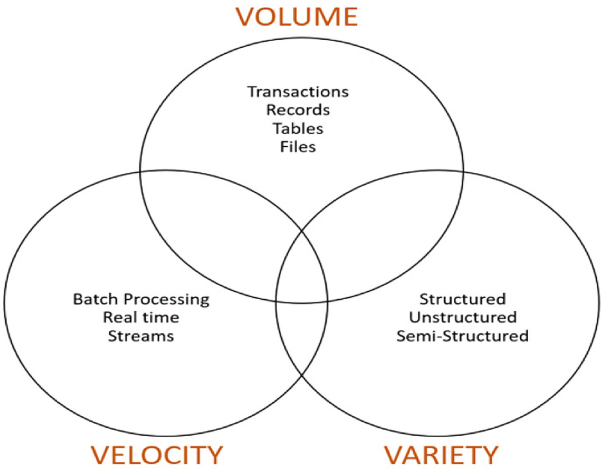

Fig. 6. Big data's important 3 axes.

4.2 Risk Management and Fraud Systems

In this paper, we focus only on fraud monitoring systems as a first step, we describe how things works with this type of systems, the current architecture used to build a full Fraud system, and the issues occurred on this type of architectures. Using Big Data technologies in this kind of services can make enormous changes on the current version of processing in performance perspectives, which is considered as a key factor and most important aspects in real time fraud monitoring logic, because every millisecond become very important to prevent financial impact on the banking institutions.

4.2.1 Existing Fraud Systems and Limitations

Today's financial institutions need a real-time automated system to detect fraud through multiple channels and masses of transactions a day. Current architecture used in Banks does not reach yet the level of sensitivity required in this type of systems, either, most of uncomfortable performance systems are mostly fraud systems due to the large size of data that need to be processed and compared with defined banking rules before triggering a financial decision. All the above, should be done in few seconds, even milliseconds to improve effectiveness and impact of a risk.

4.2.1.1 Limitations

Existing architecture will never help financial institution to become more efficient to detect fraud, due to many reasons that we cite below with current architecture too:

- None of the system use the IoT technology to collect data.
- Needed use of ETLs for Extract, transform and load functions.
- Lose original raw data for any new information after processing.
- Limitation in term of processing common pool of storage data.

- Considerable amount of time during refreshment of data through Real time dashboards.
- Many applications that act on the data stored on relational databases and obtain required information.
- The collection layer (big data layer) give raw data from different sources, which lead to a delay on the treatment before reaching back the requested system due to the diverse data structured and unstructured format.

4.2.1.2 Existing System's Designs

In the below Fig. 7, we try to describe the main workflow of the banking systems, without including all parties. As you can see, we have many data sources and some of them are linked with online applications, which means that we have an activity for 24/7 nonstop that should be manageable. All data are stocked in different databases with structured

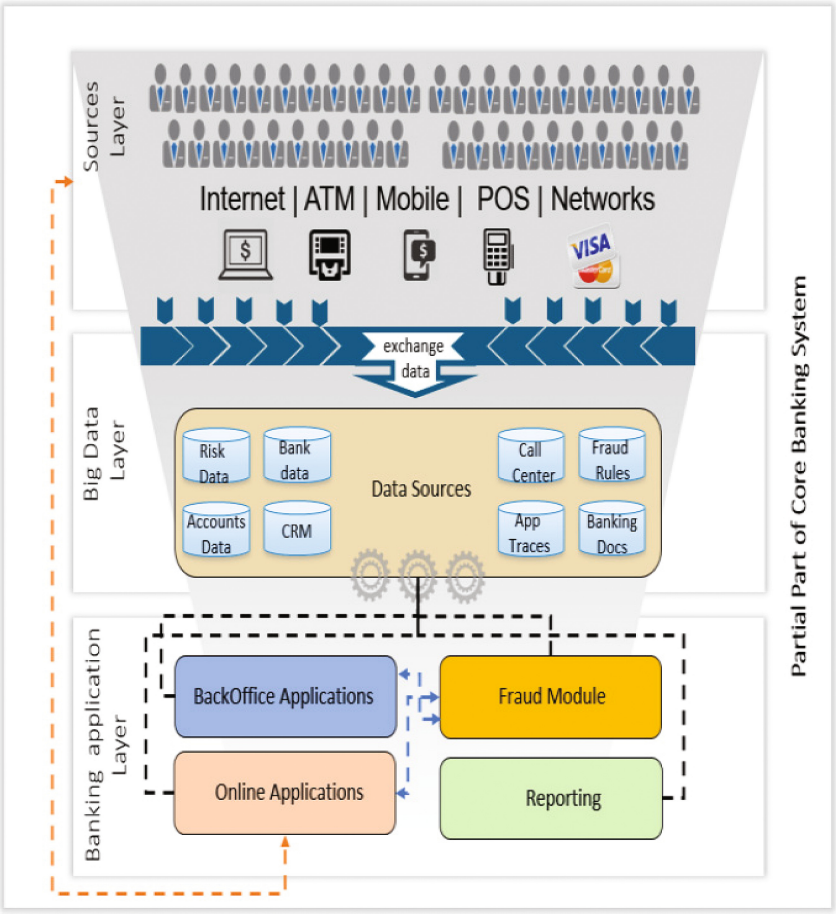

Fig. 7. Example of existing fraud systems' architecture

and unstructured format due to difference in channels and communication's level. The access to such data require many development skills, controls and ETLs that should be done on the application layer to represent data correctly, also, the way used to access to different databases still traditional (First in, first out). All the above impact consumed timing during the execution of queries, retrieve information, and trigger corresponding actions.

4.2.2 Proposed Solution and Design

In this article, we propose a system's design using powerful big data technologies like Hadoop and MapReduce. The proposed design works on providing a more efficient and more exact fraud detection system. It consists to integrate Hadoop framework inside current system's architecture and use MapReduce algorithms to get a direct impact in

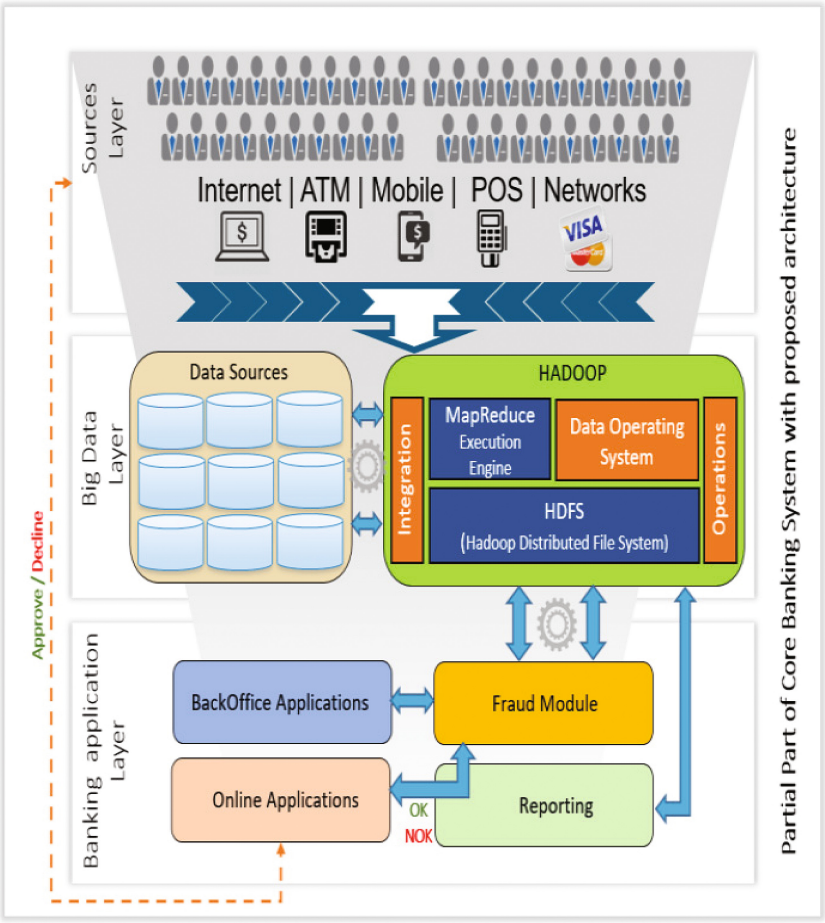

Fig. 8. Presentation of our proposed solution

term of performance frequency, and information distribution that is ignored in many previous solutions and in the current systems too.

The main objective is to identify fraudulent claims before it happens or during the processing of any online transaction that meet any fraud rule defined by the banks risk department in the database. In our architecture, we focused on MapReduce, one of the crucial enabling approaches for meeting increased Fraud systems demands by using high parallel processing, data storage, analytics and online processing on a large number of commodity nodes.

The suggested solution and architecture is presented in Fig. 8, in which we describe each element and the powerful role of Hadoop and MapReduce during the treatment to achieve the best effectiveness and efficient for detecting fraud and reduce the execution and cost.

In the above architecture, we will focus only on the role of Hadoop and his impact on the processing. In this architecture, the source of data is presented in the Top of the Fig. 7. This, include many data collected using Internet of Thing (IoT) technology to collect data via different channels, sources devices and actors in both mode Online (real time) and offline (processed by machines or human) before it came to the staging table to be processed with the core banking system. In the below section, we will explain all parties, elements and in the architecture, with their key roles.

- HDFS: As you know, HDFS provides Hadoop's efficient scale-out storage layer. It is used to serve and allowing wide variety of data access methods to operate on data stored in Hadoop. The required data by Fraud module will be kept for both Online and Batch processing in real time in Hadoop's Distributed File System. The main of our solution in this architecture is to use Hadoop sub programs like MapReduce and Machine learning to understand fraud patterns and trigger the rules matching.
- MapReduce: is used to be able to connect powerful large clusters of computers. In our proposed design, MapReduce is applied to large batch and online orientation processing of transactions, and, organize and reduce the result of the map job from each node into unified response to a query. In addition, this is the reason why MapReduce program is designed for applications such as monitoring and stream processing.

From initial testing perspective, we conclude that the integration of Hadoop will solve the performance issue in term of response-time during online transactions. Which means, that the full core banking system will be impacted positively in time consumption propose and more efficient during the generation of alert process, thanks to the powerful MapReduce algorithm that detect and much fraudulent rules with fraudulent activities and respond back to separate Fraud system in order to decline the processing of the concerned transaction. The parallel processing provided by MapReduce on large number of commodity nodes is a very good adventure in this type of systems in which time matters and quick action make difference.

5 Conclusion

In conclusion, our paper introduces some preliminary knowledge of Fraud systems its fraudulent rules and behaviors. The data become very important to detect fraudulent behavior that's why Internet of things because a major factor that collect huge data from different devices and share them via internet to prevent any fraudulent actions that's why data in banking institution become big data after each day that is why we tried to create a new design to involve big data technologies in Fraud systems as a new generation of fraud monitoring and fraud risk management in both type's Real time and Offline. Our analysis of Big Data technologies like Hadoop and MapReduce proves its huge potential, to reduce the detection and/or processing time of treatment to prevent Fraud before it happens. Another important advantage that we offer in our proposed system was the ability to handle all types of fraud and not limited to single type or scenarios, which mean a global solution based on IoT. All this, leads to the conclusion that the best solution for detecting fraud in the banking domain system is, at present, the optimized response during treatment of transaction in real time, and the optimized and research in terms of technologies and in terms of models of analysis, which make a huge difference between all international fraud systems providers.

References

1. Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B.: APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decis. Support Syst. **75**, 38–48 (2015)
2. Mahmoudi, N., Duman, E.: Detecting credit card fraud by modified fisher discriminant analysis. Expert Syst. Appl. **42**(5), 2510–2516 (2015)
3. Halvaiee, N.S., Akbari, M.K.: A novel model for credit card fraud detection using artificial immune systems. Appl. Soft Comput. **24**, 40–49 (2014)
4. West, J., Bhattacharya, M.: Payment card fraud detection using neural network committee and clustering. Comput. Secur. **57**, 47–66 (2016)
5. Zareapoor, M., Shamsolmoali, P.: Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Comput. Sci. **48**, 679–685 (2015)
6. Duman, E., Buyukkaya, A., Elikucuk, I.: A novel and successful credit card fraud detection system implemented in a Turkish bank. In: 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW). IEEE (2013)
7. Bahnsen, A.C., et al.: Cost sensitive credit card fraud detection using Bayes minimum risk. In: 2013 12th International Conference on Machine Learning and Applications (ICMLA), Vol. 1. IEEE (2013)
8. Wei, W., et al.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web **16**(4), 449–475 (2013)
9. Hormozi, E., et al.: Accuracy evaluation of a credit card fraud detection system on Hadoop MapReduce. In: 2013 5th Conference on Information and Knowledge Technology (IKT). IEEE (2013)
10. http://www.jeuneafrique.com/404469/economie/gabon-bgfi-bank-secouee-fraude-massive-aux-cartes-visa-prepayees