Machine Learning for Big Data Analytics with Blockchain Based Security

Prof. Harshal D. Misalkar, Prof. Omprakash A. Jaisinghani, Prof. Anup W. Burange

Abstract- **Big data is more than storage of and access to data. Big data Analytics plays an vital role in making sense of the data and exploiting its value. But it's a momentous confront to learn and develop new types of machine learning algorithms. Machine learning algorithms can face challenge of proper, well managed dimensions scaling is, plus there are challenges of dealing with velocity, volume and many more machine learning algorithms. Here, in this paper , we are first see the sights of big data concept, bringing with an urgent need for advanced data acquisition, management, and analysis mechanisms, we have elaborated the concept of big data and draw attention to the four phases of big data that are engendering data, acquisition of data, storing this large data, and then analyzing data. The next segment of this paper, center of attention is on dealing with big data using machine learning (ML), and highlighted the three ML methods: supervised learning, unsupervised learning and strengthening learning and its large impact on Volume of data. One of the most essential problems of Big Data is the lack of security and privacy protection of information in the Big Data. In this paper we focus on reinforcing the security of Big Data platforms by proposing a blockchain-based access control framework.**

Keywords-**Big Data, Acquisition, Machine Learning, Blockchain, unsupervised learning. Reinforcement learning.**

I. INTRODUCTION

The emerging big-data paradigm, owing to its broader impact, has strongly transformed our society and will continue to attract diverse attentions from technological experts as well as the public in general. It is palpable that we are living a data deluge era that is evidenced by the absolute volume of data from a variety of sources and its growing rate of generation. An IDC report [1] determines that, from 2005 till 2020, the global data volume will enhance by a factor of 300, from 130 exabytes to 40,000 exabytes, representing a double growth every two years.

TheMachine learning can summarize as a "Study by which computer develop the ability to learn itself without having explicitly programmed". Science of algorithms says that, the algorithms "learn" from the dataset, rectifying patterns for instance, and then automates output- whether that's sorting data into categories or making predictions on future outputs.

Machine Learning is a era, defined any activity that involvesautomated learning from data or experience. At the core of machine learning is the ability of a machine to enhance the performance of particular tasks through being exposed to data. Machine learning gathers the knowledge from the data it is exposed to and then applies this knowledge to deliver predictions about the new data which is previously unseen data [1] [7].

The quality of the predictions delivered by machine learning model depends on a number of factors:

- ➢ Well the relevant knowledge is represented by the module.
- ➢ Affectivity to complete and learn data.
- ➢ Easier way to forecast the problem in general.

Most of the times, machine learning became a very good technology at certain recognition, identification or categorization tasks like fingerprint detection, voice or faces recognition [1][4]. Likewise recent clustering algorithms machine learning is very good at automatically grouping people according to their profiles, identify market segments, forming communities, and even segment images or distinguish genes [1], [6]. The successful application of machine learning models to these problems was possible not only thanks to the learning model ingenuity but primarily thanks to the verycareful data with the properties like transformative, identifiable, and generation of multidimensional that made the learning problem discriminative enough to make easy distinctions between the predicted targets based on the data [1], [2], [6].

Depending upon the depth of knowledge that is available for learning, machine learning models can be categorized into supervised, unsupervised and semi-supervised learning algorithms [1], [3], [5], [7].

Big data is categories into two types: Structured data and Semi structured data. Structured data are sequence of words and numbers that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like account balances, sales figures, and transaction data [18].

Unstructured data comprise more complex information such as reviews of customers and forums from commercial websites, photos and other multimedia, and comments on social networking sites. These data is very difficult to divide into categories or analyzed numerically.

Blockchain is a public ledger that keeps track of all online transaction securely and anonymously. It does this by keeping these actions recorded by all computers that are connected into the Blockchain. Essentially every user authenticates the transaction and approves it. This way no single person (unauthorized) can technically alter or "cheat" the system [19]. All computers must connected to the Blockchain, which are recording data, to alter its data for it to become "real". Blockchain is decentralized in nature, meaning that no governments, authorities and single locations are used. Being decentralized and because the data is scattered by many machines they keep Blockchain protected.

1.1 Big Data And Its Impact

a) Layered Architecture for Big Data

The big data system can be decomposed into a layered structure, as illustrated in Fig. 1. The layered structure is divided into three layers, i.e., Infrastructure layer, Computing layer, and the Application layer, in bottom up approach. This layered view only provides a theoretical hierarchy to underline the complexity of a big data system. The function of each layer is as follows.

The infrastructure layer made-up of a pool of ICT resources, which can be organized by cloud computing infrastructure and enabled by virtualization technology. In this model, resources must be allocated to meet the big data demand while achieving resource efficiency by maximizing system utilization, energy awareness, operational simplification, etc.

Computing layer aggregates various data tools into a middleware layer that runs over raw ICT resources. In the era of big data, usual tools include data integration, data management, and the programming model. Data integration is a process acquiring data from disparate sources andintegrating the dataset into a unified form with the necessary data pre-processing operations. Data management refers to mechanisms and tools that provide persistent data storage and highly efficient management, such as distributed _le systems and SQL or NoSQL data stores. The programming model implements abstraction application logic and facilitates the data analysis applications. Map Reduce [12], Dryad [13], Pregel [14], and Dremel [15] exemplify programming models.

The application layer act as a interface provided by the programming models to implement various data analysis functions, including querying, statistical analyses, clustering, and classification; then, it combines basic analytical methods to develop various field related applications.

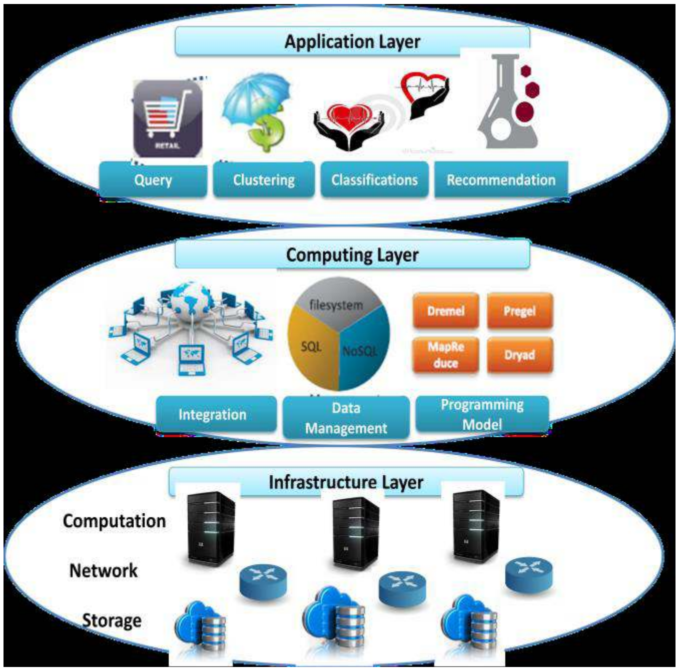

Fig. 1. Architecture for Big Data

1.2 Big Data Phases

A big-data is a complex system which provides functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction. At the same time, the system usually involves multiple distinct phases for different applications [16], [17]. In this section, we focus on the chain for big data analytics. Specifically, we describe a big data chain that consists of four stages (generation, acquisition, storage, and processing).

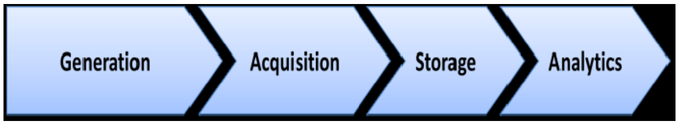

Fig. 2. Big Data Value Chain

a) Data Generation

The first and most important phase of Big data chain is Data Generation. Data generation rate can be characterized by the trends of big data generation. Specifically, the data generation rate is growing due to technological advancements. we roughly classify data generation patterns into three sequential stages:

Stage I : The first stage began in the 1990s. As digital technology and database systems were widely used, many management systems in various organizations were storing huge volumes of data, such as bank trading transactions, government sector archives and shopping mall records.

Stage II : The second stage starts with the rising popularity of web systems. The Web 1.0 systems, characterized by web search engines and ecommerce businesses after the late 1990s, generated large amounts of semi-structured and/or unstructured data, including web pages and transaction logs. Since the last decade, many Web applications created an abundance of user-generated content from online social networks, such as forums, online groups, blogs, social networking sites, and social media sites.

Stage III : The third stage is about the emergence of mobile devices, such as smart phones, sensors, tablets, Internet-enabled devices. The mobile oriented network has and will continue to create highly mobile, person-centered location-aware and context relevant data in the near future.

Considering this classification, we can see that the data generation pattern is evolving rapidly, from passive recording in Stage I to active generation in Stage II and automatic production in Stage III. These three stages are primary sources of big data, which will help to produce automatic production pattern which will contribute the most in the near future.

b) Data Acquisition

As illustrated in the big data value chain, the task of the data acquisition phase is to aggregate information in a digital form for further storage and analysis. Intuitively, the acquisition process consists of three sub-steps, data collection, data transmission, and data pre-processing, as illustrated in Fig. 3.

Data collection refers to the process of retrieving raw data from real-world objects. The process needs to be well designed. Otherwise, inaccurate data collection would impact the subsequent data analysis procedure and ultimately lead to invalid results.

Data transmission refers to once we gather the raw data, we must transfer it into a data storage infrastructure, commonly in a data center, for subsequent processing. IP backbone transmission and data center transmission are the two procedures of transmission.

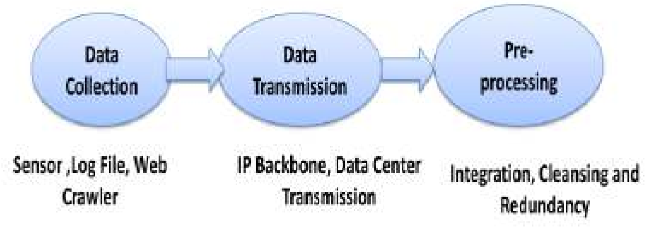
Data Collection
Data Transmission
Pre-processing
Sensor ,Log File, Web Crawler
IP Backbone, Data Center Transmission
Integration, Cleansing and Redundancy

Fig. 3. Subtasks of Data Acquisition

Data pre-processing is an important process in data acquisition phase. Collected data sets may have different levels of quality in terms of noise, redundancy, consistency, etc. On the demand side, certain data analysis methods and applications might have strict requirements on data quality. As such, data preprocessing techniques that are designed to improve data quality should be in place in big data systems.

c) Data Storage

The data storage organizes the collected information in a well structured format for analysis and value extraction purpose. For this data storage system provide two sets of features: Data storage must accommodate information persistently and reliably. The data storage system must provide a scalable access interface to query and it will analyze a huge amount of data..

d) Data Analysis

The last and most important stage of the big data chain is data analysis. Goal of data analysis is to extract useful values, suggest conclusions and/or support decision-making.

II. MACHINE LEARNING FOR BIG DATA

Well structured step by step execution is very important to process big data. As our traditional systems are not capable of processing this huge data, so we must take access of big data processing machine to avoid waste of time and importantly resources. Analyzing and processing big data this challenge came from all sides: accessing large amounts of data, learning the models from it and carry out mass predictions all in a reasonable time.

In fact all machine learning algorithms with computational complexity of O(n2) immediately become intractable when faced with billions of data points. Depending on the depth of knowledge that is available for learning, ML models can be categorized into supervised, unsupervised and semi-supervised learning algorithms [1], [3], [5], [7].

Supervised Learning

Supervised learning algorithms use as training examples. As our traditional systems are not capable of processing this huge data, so we must take access of big data processing machine. In the era of processing big data, huge data is first categorize into some equal section, where input data and the target outputs are given to particular intelligent machine to process allotted data. This process is known as mapping function as we are mapping huge data across different intelligent machine for processing purpose.

All machines required equal amount of time to process given data and to generate output. Data aggregation will be used to combine the output generated by all machines to generate common output. This process is known as reduce function.

Support vector machines (SVM) were also reported as very powerful in terms of achieved performance on the numerical data thanks to their explicit effort to maximise the margin of misclassification along the boundaries among the classes of data [1]- [5]. Traditional SVMs are quadratically complex O(n2), yet there are successful attempts to cleverly reduce their complexity to linear complexity [15]. We intend to expand on these developments.

Neural networks (NN) are yet another example of very powerful, flexible and robust predictors for both classification and regression problems [1]- [5]. Its generic structure allows to learn virtually any function successfully with enough hidden layers of processing nodes and training data. Like for SVM, NNs are typically O(n3) or at best quadratic ally complex with the number of examples. Complex structures further expand computational demands of learning NN. There have been recent attempts to redesign NN into more scalable architectures [16], further work is required to make NNs fully scalable for big data processing.
Fig. 4 captures a brief demonstration of a supervised learning applied to the famous Iris species detection problem based on just 2-dimensional data of petal width and petal length:

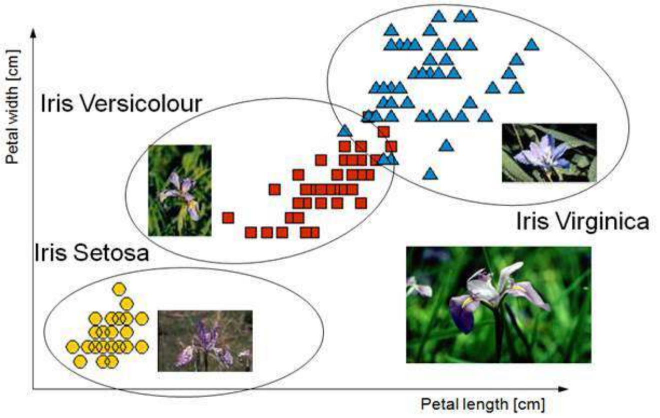

Fig.4 Demonstration of a supervised learning applied to the famous Iris species detection problem

Machine Learning is knowledge of creating algorithms and program to perform intelligent task. Once algorithm designed, there is no need of human interaction to become better. Machine learning applications include following: Web Search, stock trading, recommender systems, spam filters, credit scoring, ad placement, fraud detection, computer vision and drug design.

An easy way to understand machine learning- It is not possible for we human-beings to create models for every possible search or spam, so make the machine intelligent enough by writing algorithms or programs to learn by itself. This process is known as machine learning.

Machines learn to execute tasks that are not exclusively programmed to perform. Many techniques are put into observations like supervised clustering, regression, naive Bayes etc.ML is an important part of data science. Data science is a big era covering each and every aspect of data processing. Machine Learning covers statistical as well as algorithmic aspects.

To mention, data science includes

- data visualization
- data integration
- dashboards and BI
- distributed architecture
- automated, data-driven decisions
- automating machine learning
- deployment in production mode
- data engineering

Machine learning is very helpful technique for data science by making a provision for data analysis, data preparation, decision making like real time testing, online learning. In order to provide a solution, Data science clubs collectively algorithms derived from machine learning. Data science taking a lot of ideas right from basic mathematics, statistics and domain expertise to carries out this activity.

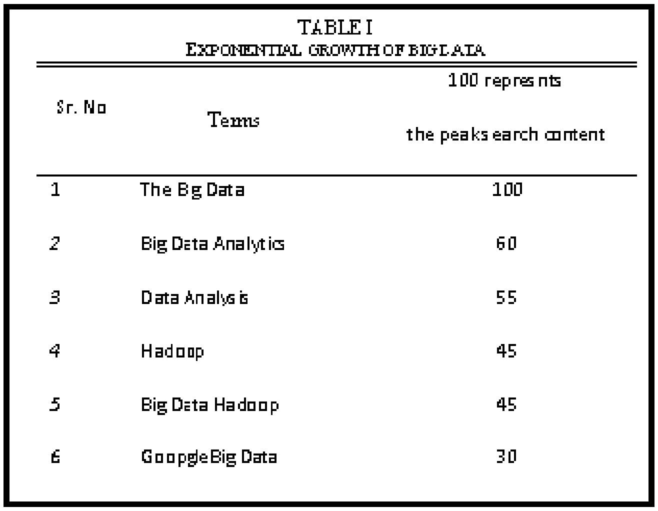

TABLE I
EXPONENTIAL GROWTH OF BIG DATA

Sr. No	Terms	100 represents the peak search content
1	The Big Data	100
2	Big Data Analytics	60
3	Data Analysis	55
4	Hadoop	45
5	Big Data Hadoop	45
6	Goopgle Big Data	30

This table shows the exponential growth of the ―big data‖ google trends, along geographical distribution and the most popular related terms, source http://www.google.com/trends.

III. BIG DATA REVOLUTION

Big Data initiated the true revolution in the way we store, manage and process the data. Coming now in Exabyte per day BD and its popularity continue to grow exponentially as it can be seen from the Google Trend chart presented in the table 1.

Big data are linked to the explosive growth in mobile devices and their ability to generate, collect, share and access huge amounts of numeric, textual, imagery and video data. This data combined with a connected development of internet resources, networked services and the cult of sharing on ever growing social networks what we are experiencing is a true data revolution.

IV. BLOCKCHAIN BASED SECURITY

Blockchain technology is a new technique of combining existing technologies, such as public and private keys, digital signatures, peer to peer networks and distributed ledgers. Considering security of Blockchain, information can be shared, produced and maintained completely decentralized.

Everyone who joins the network gets a copy of the network via the distributed ledger. In this large network everything is accepted or denied using a consensus procedure. These procedures are conducted by "nodes" which control the quality of data. It also accepts or denies information depending if it is according to the rules of the Blockchain. When the network agrees that data is accepted it is then put into a block. The peer to peer network decides between certain intervals if a block is added to the chain.

Blockchain is a technology based around everyone in the network. For example, in figure 1 below, if Person A wants to send money to a person B, they create a "transaction block", which then is shown to everyone who belongs to the network. When all the users approve the transaction, the block is then added to the previously entered blocks that other users have made doing these kinds of transactions. Blocks contain a digital signature, timestamp and other important relevant information.

The blocks are added and create a long chain, therefore technology name "Blockchain". The root of the chain dates to the very first block ever created.

Block is not allowed to join the chain unless either block is not approved by the network or does not have a corresponding number to the previous block.

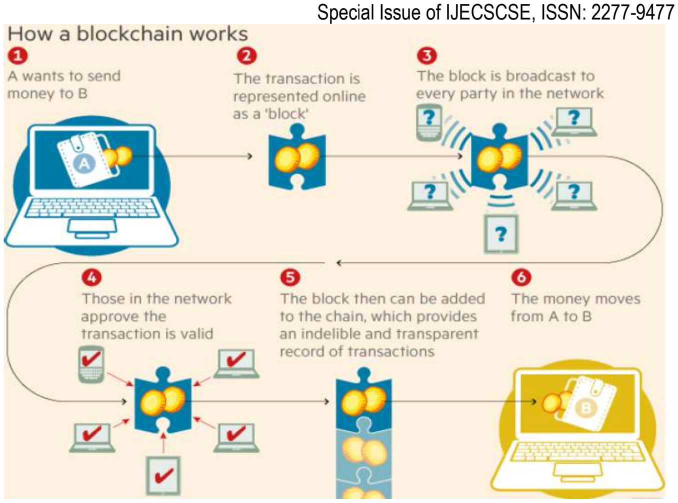
How a blockchain works
1
A wants to send money to B
A
2
The transaction is represented online as a 'block'
3
The block is broadcast to every party in the network
?
4
Those in the network approve the transaction is valid
✓
5
The block then can be added to the chain, which provides an indelible and transparent record of transactions
6
The money moves from A to B
B

Fig. 5 Working of Blockchain

a) Industry Uses for Blockchain

Right now, the biggest use of Blockchain is for the virtual currency Bitcoin, Blockchain can be used for a variety of different applications. They can be used in many industries, such as cyber security, voting systems, shipping, banking, real estate etc. Today, Blockchain technology is not yet used for many of these commercially, but it has created a lot of interest on how it could make these industries easier in terms of security. One of the biggest, banking, has sparked interest due to possible billions in savings, voting could reduce fraud and violence, shipping could reduce money, time and easiness. All in all, saving money, time, increasing safety are the three biggest factors.

b) Banking and other Contracts

Japanese banks start working on a blockchain startup called Ripple to facilitate money transfers between bank accounts using blockchain. The ultimate reason behind the progress is to perform real-time transfers at a significantly low cost. Traditional real-time transfers were expensive because of the potential risk factors. Double-spending is a real problem with real-time transfers. Using Blockchains this risk is largely avoided [18].

Big data analytics makes it possible to recognize patterns in consumer spending and identify uncertain transactions a lot quicker than they can be done currently. This reduces the cost with real-time transactions.

In Industries, the main drive for adoption of Blockchain technologies has been security. Healthcare, retail and public administration, establishments etc. these organizations started experimenting with blockchain to handle data to prevent hacking and data leaks.

Use of Blockchain technology in banking would abolish middlemen costs and other handling fees from the user and from the bank. However, for Blockchain use to succeed in banking, it would have to be taken into use around the world to help reduce the risk of failure. Transactions generally can take up too many days before being completed today, but with the help of Blockchain they could be completed within hours or even minutes/seconds and without any huge costs for the user or bank.

V.Conclusion

In this paper we have surveyed the relevant work of big data, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In this paper, we have explained the concept of big data and the big data value chain, which covers the overall big data lifecycle. The big data value chain is aggregation of four stages: Generation of Data, data acquisition, data storage, and data analysis. Further this paper presents approach which will focus on dealing with big data using machine learning(ML), and highlighted the three ML methods: supervised learning, unsupervised learning and reinforcement learning and its impact on big data.
Blockchain can be used to save costs in transactions and getting rid of middlemen, make data storage easier, increasing cyber security, creating new jobs, saving time and money, etc. Theoretical uses for the Blockchain are essentially endless and more uses for it will grow in the future.

References

1. T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

2. R.O. Duda, P.E. Hart and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.

3. C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

4. S. Marsland. *Machine Learning, An Algorithmic Perspective*. Chapman and Hall / CRC Press, Boca Raton, 2009.

5. A. Mohri, A. Rostamizadeh and A. Talwalker. *Foundations of Machine Learning*. The MIT Press, Cambridge, 2012.

6. C.C. Aggarwal and S.K. Reddy. *Data Clustering, Algorithms and Applications*. Chapman and Hall / CRC, Boca Raton, 2014.

7. O. Chapelle, B. Cholkopf, A. Zien. *Semi-Supervised Learning (Adaptive Computation and Machine Learning Series)*. MIT Press, Cambridge, 2006.

8. J. Liebowitz. Big Data and Business Analytics. CRC Press, Boca Raton, 2013.

9. V. Mayer-Schonberger and K. Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think. Houghton Mifflin Harcourt Publishing, New York, 2013.

10. YC Kwon, D. Nunley, J.P. Gardner, M. Balazinska, B. Howe, S. Loebman. Scalable Clustering Algorithm for N-Body Simulations in a Shared-Nothing Cluster. Scientific and Statistical Database Management 6187: 132-150, 2010.

11. J. Gantz and D. Reinsel, ``The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,'' in *Proc. IDC iView, IDC Anal. Future*, 2012.

12. J. Dean and S. Ghemawat, ``Mapreduce: Simpli_ed data processing on large clusters,'' *Commun. ACM*, vol. 51, no. 1, pp. 107_113, 2008.

13. M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, ``Dryad: Distributed data-parallel programs from sequential building blocks,'' in *Proc. 2nd ACM SIGOPS/EuroSys Eur. Conf. Comput. Syst.*, Jun. 2007, pp. 59_72.

14. G. Malewicz*et al.*, ``Pregel: A system for large-scale graph processing,'' in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2010, pp. 135_146.

15. S. Melnik*et al.*, ``Dremel: Interactive analysis of web-scale datasets,'' *Proc. VLDB Endowment*, vol. 3, nos. 1_2, pp. 330_339, 2010.

16. E. B. S. D. D. Agrawal *et al.*, ``Challenges and opportunities with big data_A community white paper developed by leading researchers across the united states,'' The Computing Research Association, CRA White Paper, Feb. 2012.

17. D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, ``Interactions with big data analytics,'' *Interactions*, vol. 19, no. 3, pp. 50_59, May 2012.

18. MikkoHuhmo" Blockchain Technology, Bitcoin as a case" Spring 2018.

19. Hamza ES-SAMAALI, Aissam OUTCHAKOUCHT, Jean Philippe LEROY, "A Blockchain-based Access Control for Big Data" IJCNCS VOL. 5, NO. 7, JULY 2017, 137–147