Towards a Data Traceability Management for Big Data

Othmane RAHMAOUI, Kamal SOUALI, Mohammed OUZZIF

Abstract—Today traceability is a buzz word and it used in several domains like healthcare, food industry and transportation sectors. In Information Technology, traceability plays a very important role and it can be defined in various ways, depending on the environment and process under consideration. In this paper we build upon a prior body of word to highlight the need in traceability management in general and especially its use in large volume of data storage. For big data systems, traceability means knowing where your data are located, how information is derived and from which data. Also knowing who should have access to it and for what reasons.

Index Terms—Big data, data traceability, trace, traceability, traceability management;

1 Introduction

TRACEABILITY plays an important role in different sectors; in ISO (ISO 9001:2000), it is defined as "the ability to trace the history, application or location of that which is under consideration". For some people, it is merely a tool to keep a history over something important that happened in the past. For others, is has no added value to their actual processes or product but in fact, it is becoming more and more valued. A global traceability uses two techniques in general, backward traceability (The tracing) aims to identify the origin and characteristics of a product and descending traceability (The tracking) aims to find the location of a product [1].

In food industry, traceability is a well-coordinated and well documented movement of product and documented activities associated with the product, from producer, through a chain of intermediaries, to the final consumer [2]. It is considered as a mechanism used to keep the history of a row or semi-finished unit during manufacturing and until unit is delivered. It has a great potential to improve food safety as well as to promote consumer protection, by providing quality information [3].

In healthcare, traceability is becoming increasingly important to pharmaceutical, biotechnology and medical device companies, it also introduces the technology aspect to manage the clinical information system and to be more effective to consider the basic principles of how data, information and knowledge are generated and how these can traced throughout the lifecycle of a product [4]. Today we talk also about the blood traceability like the perfect solution for a safe blood transfusion; means the ability to trace each individual unit of blood or blood component derived thereof from the donor to its final destination [5].

In transportation sectors, traceability is necessary to improve customer satisfaction and to establish a fluent communication throughout logistics process. It refers to the capability for tracing goods along the distribution chain on a barcode or series number basis.

For information technology (IT), traceability plays an important role in the development and assurance of software systems, it helps to assure that an as-build system correctly implements all requirements by supporting change impact assessment, re-engineering of applications and other critical software engineering activities. Requirements traceability means the capability to show, follow the life and expiry of the requirements are properly designed and well tested [6]. Software traceability has long been recognized as an important quality of a well-engineered software system defined by the Center of Excellence for Software and systems Traceability (CoEST). Traceability is successfully implemented in some projects within some organizations while the majority of projects fail to achieve affective traceability or incur excessive costs in so doing [7]. Software traceability is an essential element of the software development process.

Today traceability is one of the basic tenets of all software safety standards and a key prerequisite for certification of software. The traceability research is constantly gaining attention and in the future it will be a serious research topic.

This work introduces, in the first section several researches that are an important value to data traceability management. In the second section, data traceability management for big data is exposed. The third section describes our future work as a solution will be implemented on Hadoop Environment.

2 Data Traceability Management

2.1 Software Traceability System

According HongWu Bai & al. 2017 [8], the traceability has a very important role for farm animals and their products by using technologies to manage it and to control it. A traceability system is composed of the identification of Traceable Resource Units (TRUs), a database that provides needful data with TRUs and information flow for associating the TRUs with their respective codes by collection and inquiry.

- *Othmane RAHMAOUI is currently a Phd Student in Computer Science in Hassan II University, ENSEM, Casablanca, Morocco, E-mail: othmane.rahmaoui@gmail.com*
- *Kamal SOUALI is currently a Phd Student in Computer Science in Hassan II University, ENSEM, Casablanca, Morocco, E-mail: kml.souali@gmail.com*
- *Mohammed OUZZIF is a Phd Professor in Computer Science in Hassan II University, Casablanca, Morocco, E-mail: ouzzif@gmail.com*

In China, Traceability legislation imposes an important development for monitoring products evaluation during its production, process, storage and distribution; however it is necessary to develop new technologies and realistic approaches in order to provide a fundamental traceability system.

A fundamental traceability system should unify the retroactive protocols and data interfaces in order to link different traceability data. Patrick Rempel & Al. on May 2016 are proposing an automated traceability assessment approach called TRUST that can be used to assess software traceability [9], it consists of four components; Traceability Store, Traceability Planner, Traceability Collector and Traceability Assessor. TRUST comprise four steps as shown in Figure 1, Planning traceability information (A), Collecting traceability data (B), Assessing traceability data (C) and reporting assessment results (D).

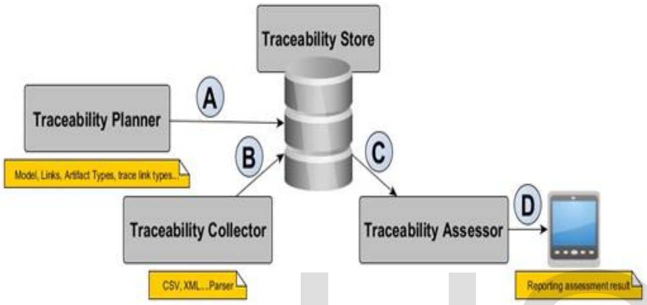


Fig. 1. Software Traceability Assessment Approach (TRUST)

In (A) a target state of a traceability implementation is specified as a reference for assessing the actual state of a project's traceability, specifying all artifact types, trace link types and trace path types that are required to enable all software development activities that require traceability. The result of the planning step is a traceability information model (TIM). The resulting TIM is stored in the Traceability Store.

In (B), the Traceability Collector collects and parses data traceability of a software development and it also stored in the Traceability Store.

In (C) a traceability quality model is proposed to define assessable traceability data properties with respect to the qualities completeness, correctness and appropriateness. Based on this model the Traceability Assessor analyzes each traceability data property within the Traceability Store for deviations.

In (D) the Traceability Assessor aggregates the traceability assessment results into a report.

This approach is not limited to specific artifact types and can be applied to any software project.

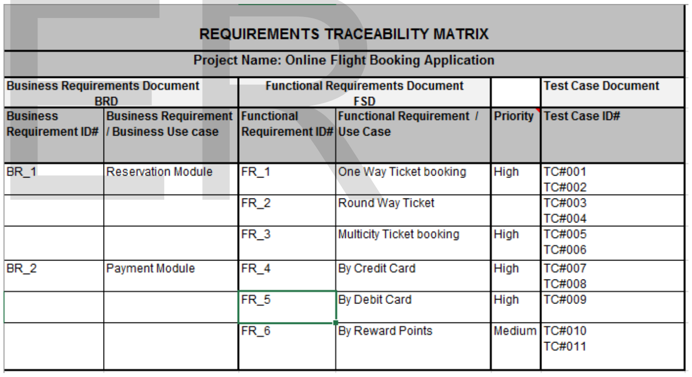
Software traceability is a required component of many software development processes. It is an indispensable tool for the success of information system development projects, in reality, it is a powerful strategy to ensure and to be able to prove what has been requested as a need to verify and validate the exploitation of the developed systems.

2.2 Requirements Traceability

On software quality, requirements traceability has been recognized as an important quality of a well-engineered system. It is broadly recognized as a critical factor for software quality according Patrick Rampel and al. 2016 [10], in this research, authors are focusing on the completeness of requirements traceability in software projects and how it impacts the implementation quality of requirements. A tool was developed to automatically collect and analyze project specific software development artifacts can be used by practitioners to automatically calculate their degree of traceability implementation completeness. The results provide for the first time empirical evidence that improving the degree of traceability completeness indeed decreases the defect rate to be expected, and thus helps to raise implementation quality of the developed software.

Requirement traceability means the links between requirements to another requirement and that means also tracking all requirements in Software Requirements Specification (SRS) to confirm that all requirements are properly designed and well tested.

According Swathine. K and Al. (January 2017), Software have many artifacts, multiple version of artifacts are existing, Artifacts are continuously changing and updating regularly, which need proper management in order to software work properly. Traceability links provide good understanding relationships between artifacts and it represented by traceability matrix [11].



REQUIREMENTS TRACEABILITY MATRIX					
Project Name: Online Flight Booking Application					
Business Requirements Document BRD		Functional Requirements Document FSD			Test Case Document
Business Requirement ID#	Business Requirement / Business Use case	Functional Requirement ID#	Functional Requirement / Use Case	Priority	Test Case ID#
BR_1	Reservation Module	FR_1	One Way Ticket booking	High	TC#001 TC#002
		FR_2	Round Way Ticket		TC#003 TC#004
		FR_3	Multicity Ticket booking	High	TC#005 TC#006
BR_2	Payment Module	FR_4	By Credit Card	High	TC#007 TC#008
		FR_5	By Debit Card	High	TC#009
		FR_6	By Reward Points	Medium	TC#010 TC#011

Fig. 2. Example of simple Requirements Traceability Matrix (RTM), (Opencodez Website).

Requirements traceability is an important mechanism for managing verification, validation and change impact analysis challenges in system engineering.

2.3 Traceability Management

In order to manage and adopt the most appropriate traceability scheme for the project, we should answer the following questions [12]:

- What information is recorded in the artifact?
- Who has created or updated the artifact?
- Who are the potential users of the artifact?
- What is the source of information (recorded in the artifact)? (whether it is policies, telephone calls, documents, standards, etc)

- How is the information represented? (whether it is documented as formal or informal text, or as graphics, or documented as audio or video recordings)
- When was the artifact created or modified?
- Why was the artifact created or updated?

The traceability links between different artifacts generated by different tools have to be stored and maintained.

Today many enterprises use some techniques and technologies to manage data traceability in several domains like RFID (Radio Frequency Identification), barcodes, ERP (Enterprise Resource Planning), etc and in recent years, management software vendors have added more features to help facilitate traceability.

3 Data traceability Management For Big data

3.1 Big data and challenges

Big data is a term that describes the large volume of data (structured and unstructured), the term "big data" is being increasingly used almost everywhere on the planet. The concept has grown in the early 2000s when expert industry analyst Doug Laney articulated the definition of big data as the three Vs:

- Volume: is the V most associated with big data because, well, volume can be big. What we're talking about here is quantities of data that reach almost incomprehensible proportions. Volume refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, photographs, video, etc. The vast amounts of data have become so large in fact that we can no longer store and analyze data using traditional database technology.
- Velocity: refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every day the number of emails, Social network messages, photos, video clips, etc. increases at lighting speeds around the world. Every second of every day data is increasing.
- Variety: is defined as the different types of data we can now use. Data today looks very different than data from the past. We no longer just have structured data (name, phone number, address, financials, etc) that fits nice and neatly into a data table. Today's data is unstructured. In fact, 80% of the entire world's data fits into this category, including photos, video sequences, social media updates, etc. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers.

Today Big Data and the collection of the large volumes of data classified, structured and stored, allows to quickly analyzing it to propose more intelligent decisions to evolutions.

Indeed, to have some flexible systems, reconfigurable and adaptable to the changes requested, traceability plays a very important role in order to allow the reuse and the enrichment of the models concerning the process modeling of the volumes of data.

The quantity of data that is being stored is not always one hundred percent useful. To become useful, specialists must sort and clean up data. This process is time consuming and the costs are proportional to the volume of data [13].

Big data has great potential to produce useful information for companies which can benefit the way they manage their problems. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. These massive data sets are too large and complex for humans to effectively extract useful information without the aid of computational tools.

3.2 Traceability Management For Big data

The processing of Big data becomes the center of interest in science and industry.

In 2015, Richard McClatchey and al. are presenting an approach that has been adopted to provide detailed traceability to support research analysis processes in the study of biomarkers for Alzheimer's disease, but is generically applicable across big data medical systems [14]. This study is based on a virtual laboratory (N4U VL) which offers neuroscientists tracked access to a wide range of datasets, algorithm applications and to computational resources and services in their studies of biomarkers for identifying the onset of Alzheimer's disease, and a workflow and process tracking system called CRISTAL, that was developed to track and manage the evolution of data and workflow usage over time in N4U project (see https://neugrid4you.eu). neuGRID is a web portal aimed to help neuroscientists do high-throughput imaging research, and provide clinical neurologists automated diagnostic imaging markers of neurodegenerative diseases for individual patient diagnosis.

The N4U VL is composed of different components as shown in the following figure 3:

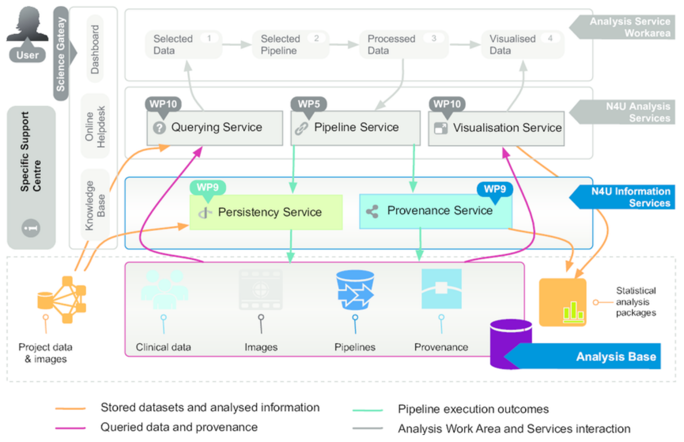


Fig. 3. The N4U Virtual Laboratory

The N4U Analysis Service provides access to tracked information (images, pipelines and analysis outcomes) for querying/browsing, visualization, pipeline authoring and execution. Its Workarea is a facility for users to define new pipelines or configure existing pipelines to be run against selected datasets and to be dispatched to conduct analyses.

CRISTAL captures the provenance data resulting from specification and execution of the stages in analysis workflow. Its provenance management service also keeps track of the origins of the data products generated in an analysis and their evolution between different stages of research analysis.

The final result of completed workflow/pipeline will be presented to the user for evaluation. There are further plans to enrich the CRISTAL kernel (The data model) to model not only data and processes (Products and activities as items) but also to model agents and users of the system (whether human or computational).

Researchers require systems that provide traceability of information through provenance data capture and management to support their analyses.

4 Overview of the Proposed Traceability Solution for Big Data

After having studied several researches, data has become an indispensable part of any domain (Industry, organization, Business, etc). Data are collected and analyzed to create information suitable for making decisions.

Emerging technologies such as the Hadoop framework and MapReduce offer new and exciting ways to process and transform big data defined as complex and unstructured into a meaningful knowledge.

4.1 Hadoop

Hadoop is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop services provide for data storage, data processing, data access, data governance, security, and operations.

Hadoop was born out of a need to process an avalanche of big data. The web was generating more and more information on a daily basis, and it was becoming very difficult to index over one billion pages of content.

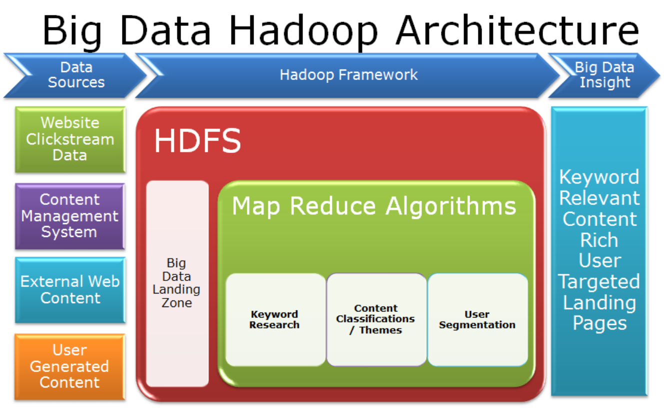


Fig. 4. Hadoop Architecture

Some of the reasons organizations use Hadoop is its' ability to store, manage and analyze vast amounts of structured and unstructured data quickly, reliably, flexibly and at low-cost.

- Scalability and Performance – distributed processing of data local to each node in a cluster enables Hadoop to store, manage, process and analyze data at petabyte scale.
- Reliability – large computing clusters are prone to failure of individual nodes in the cluster. Hadoop is fundamentally resilient – when a node fails processing is re-directed to the remaining nodes in the cluster and data is automatically re-replicated in preparation for future node failures.
- Flexibility – unlike traditional relational database management systems, you don't have to create structured schemas before storing data. You can store data in any format, including semi-structured or unstructured formats, and then parse and apply schema to the data when read.
- Low Cost – unlike proprietary software, Hadoop is open source and runs on low-cost commodity hardware.

4.2 MapReduce

A framework for writing applications that process large amounts of data

MapReduce is the original framework for writing applications that process large amounts of structured and unstructured data stored in the Hadoop Distributed File System (HDFS). MapReduce is useful for batch processing on terabytes or petabytes of data stored in Apache Hadoop.

MapReduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The MapReduce concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions.

4.3 Hadoop Distributed File System (HDFS)

HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters. Developed by Apache Hadoop, HDFS works like a standard distributed file system but provides better data throughput and access through the MapReduce algorithm, high fault tolerance and native support of large data sets.

When HDFS collects data, the system segments the information into multiple blocks and distributes it across multiple nodes in the cluster, allowing for parallel processing. The file system copies each data brick multiple times and distributes the copies on each of the nodes, placing at least one copy on a separate server in the cluster. As a result, data stored on failed nodes can be found elsewhere in the cluster. Treatment can continue despite the failure.

Because HDFS is written in Java, it has native support for Java application programming interfaces (API) for application integration and accessibility. It also may be accessed through standard Web browsers.

4.4 Our Vision

Depending on the data traceability requirements, the need for a solution describing the operations performed on all the data by the different systems and actors involved becomes a necessity.

This solution will be based on a set of procedures and activities involved in planning Big Data architecture as shown in figure 5:

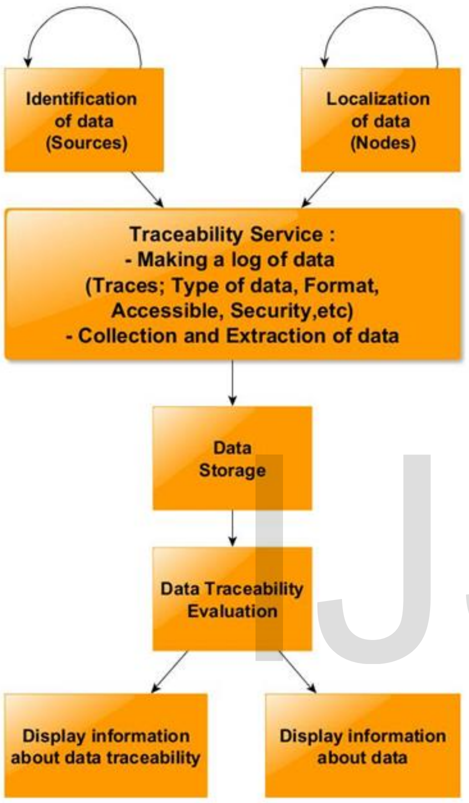

Identification of data (Sources)
Localization of data (Nodes)
Traceability Service :
- Making a log of data
(Traces; Type of data, Format, Accessible, Security,etc)
- Collection and Extraction of data
Data Storage
Data Traceability Evaluation
Display information about data traceability
Display information about data

Fig. 5. The proposed traceability management approach for Big Data

1. Process and inputs
   - Identification of data
   - Localization of data
   - Traceability Service (Mechanism process)
   - Making logs of data (Traces)
2. Features during the process
   - Data Store
   - Data Traceability Evaluation
   - Automatic
   - Real-time
   - Accessible
   - Security
3. Outputs
   - Information about data traceability
   - Information about data to make decision

Data will be identified to know what are sources and also who has created it or update it and who will be use it from users.

After the identification, Data will be stored and our mechanism process took information about the localization of data, however, the traceability service will play a very important act to collect all data knowledge (Type, Format, accessibility, security, etc.). All that information will be stocked to be evaluated and verified, and therefore we can have results (Information about data traceability) into reports or on a graphical user interface let us make best decisions.

This approach can be applied automatically and in the real-time.

4.5 Operating Scenario on Hadoop environment

In Hadoop environment we can configure our proposed traceability service with a MapReduce Jobs in a heterogenic cluster.

All information will be identified and stored by using HDFS.

In the cluster, the NameNode have information about data (Name of file, Volume, etc.), the DataNode have data themselves fragmented into Blocs.

The idea is to let the Traceability Service add more information, according to the requirements and the needs, about data by making a new node will called 'TraceabilityNode'.

The TraceabilityNode will be contacted automatically, in real-time, by the Traceability Service, and thus data will be to stored and evaluated.

The results (information) will be displayed on a graphical user interface (Figure 6).

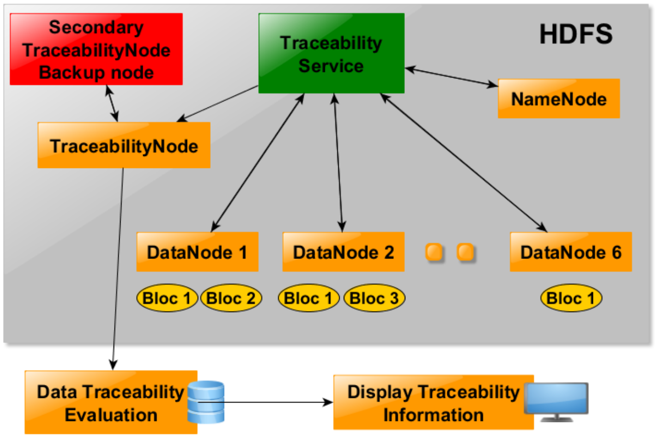


Fig. 6. The Operating Scenario of the proposed traceability management for big data on Hadoop Environment

In the future work, after studying the hadoop architecture, we aim to propose an implementation of a service or a system that will allow us to query traces via MapReduce operations. In order to generalize the solution by proposing a model that can be used in any big data project to manage the traceability in an intelligent way.

5. Conclusion

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, media sharing sites, news sources, etc. in this paper we discussed about the issues and challenges related to data traceability management for Big Data like a first step that initiates a set of future steps aiming to implement a solution to manage traceability into Big Data Systems.

References

[1] K. Souali, O. Rahmaoui and M. Ouzzif: "An overview of traceability: Definitions and techniques". CIST 2016: 789-793.

[2] A. F. Bolle and J. P. Emond, "Traceability in Postharvest Systems", Postharvest Handling, May 2014, pp.485-504.

[3] M. Garcia Martinez and F. M. Brofman Epelbaum, "The role of traceability in restoring consumer trust in food chains", Food chain integrity, March 2011.

[4] GS1 Standard, issue 1.2.0; October 2013, "Global Traceability standard for healtcare", Business process and system requirements for supply chain traceability.

[5] Directive 2005/61/EC of the European Parliament and of the Council 30 September 2005, "Certain technical requirements for blood and blood components".

[6] Parastoo Delgoshaei, Mark A. Austin and Daniel A. Veronica, "A Semantic Platform Infrastructure for Requirements Traceability and System Assessment", ICONS 2014 : The Ninth International Conference on Systems.

[7] Jane Cleland-Huang, Orlena C. Z. Gotel, Jane Huffman Hayes, Patrick Mäder, and Andrea Zisman, "Software traceability: trends and future directions", FOSE 2014 Proceedings of the on Future of Software Engineering, Pages 55-69, Hyderabad, India — May 31 - June 07, 2014.

[8] Hongwu Bai and al., "Traceability technologies for farm animals and their products in China", Food Control, Volume 79, September 2017, Pages 35-43.

[9] Patrick Rampel and Patrick Mader, "Continous Assessment of Software traceability", May 14-22, 2016 IEEE/ACM 38th international Conference on Sofware Engineering Companion (ICSE).

[10] Patrick Rempel and Patrick Mäder, "Preventing Defects: The Impact of Requirements Traceability Completeness on Software Quality", IEEE Issue No. 08 - Aug. (2017 vol. 43), ISSN: 0098-5589, pp: 777-797.

[11] Swathine. K and Al., "study on requirement engineering and traceability techniques in software artefacts", IJIRCCE, Vol 5, Issue 1, January 2017.

[12] Dr. vinay Kumar and Reema Thareja, "Managing Traceability in Data Warehouse development Projects", IJCTA May-June 2014, Vol 5 (3), 1001-1011.

[13] Alexandru Adrian TOLE, " Big Data Challenges", Database Systems Journal Vol 4, N°3/2013.

[14] Richard McClatchey and Al., "Traceability and Provenance in Big Data Medical Systems", 2015 IEEE 28th International Symposium on Computer-Based Medical Systems.