EasyChair Preprint

№ 1593

Characterization of Big Data Platforms for Medical Data

Hicham El Alaoui El Hanafi, Nadia Afifi and Hicham Belhadaoui

October 6, 2019

Characterization of Big Data Platforms for Medical Data

Hicham EL ALAOUI EL HANAFI
CED Engineering Science, ENSEM
Lab. RITM/ESTC. Hassan II University
Casablanca, Morocco
mr.elalaoui.hicham@gmail.com

Nadia AFIFI
CED Engineering Science, ENSEM
Lab. RITM/ESTC. Hassan II University
Casablanca, Morocco
afifinadia@yahoo.fr

Hicham BELHADAOUI
CED Engineering Science, ENSEM
Lab. RITM/ESTC. Hassan II University
Casablanca, Morocco
belhadaoui_hicham@yahoo.fr

***Abstract*—In recent years, technology has seen a growth in the use of big data, which helps to the decision, to find the needs of people, to know their desires, and certainly it is beneficial to the evolution of our life. On the other hand, medicine and people's lives are a very sensitive area, and in some cases the doctors have to follow their patients periodically such as chronic diseases, pregnant women, etc.**

This paper presented a big data platform for medical domain especially for the cases that require Remote Patient Monitoring in real time, which helps the doctors to follow-up their patients remotely via a smart systems thanks to new technologies using big data and its advantages.

***Index Terms*—Big Data, Hadoop, Sensors Data, HDFS, healthcare.**

I. Introduction

In the last decades, big data has become a serious topic in the world of Information Technology. this large amount of data comes from various sources : social media, databases, sensor data, entreprise data, etc. The most important thing that these big data have a very interesting value which have to be exploit.

Big Data is a generic term used to characterize the strategies and technologies used to collect, organize, process and analyze large datasets. It is the art of managing and exploiting large volumes of data.

Big Data has some new challenges compared to the classical or small data resumed In 3Vs [2], then it goes up to 7Vs and more.

The essential 5 challenges (5Vs) are:

Volume: describes the size of data, it represent the amount of data. here we talk about Terabytes and Petabytes while in the traditional databases we didn't have that amount of data.

Velocity: is also a big challenge required, velocity refers to the speed of the access to data in real time, considering the large volume of data

Variety: refers to the different types of data (structured, semi-structured, and unstructured).

Veracity: is uncertainty and imprecision of data.

Value: This refers to the ability to make this data beneficial

Fig. 1. THE 5V'S OF BIG DATA.

II. Big data in medical domain

Behind these data there is very important information that can be deduced. in the medical field can be also more important, in the fact that can simplify the tasks of doctors, they will be alerted of all patient's body reaction via smart system. it will help to predict the patient's condition, it will also help to medical researches. Thats the big evolution in medical domain.

A lot of studies has been done in this subject. Recent researches [3-4] indicate that big data and health data analysis techniques are used to adress the challenges of health efficiency and quality. For example, by collecting and analyzing health data from disparate sources, such as clinical data, treatment outcomes can be tracked by using various resources. This grouping, in turn, contributes to improving the effectiveness of care. In addition, identifying high-risk patients and predictive models that lead to proactive patient care will improve the quality of care.

W. Raghupathi and V. Raghupathi described the promise and potential of big data analytics in healthcare[5]. They

deduced that Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

In our case of following up patients with chronic and sensible diseases , data will come from sensors via a smart textile, its about different types of data (temperature, cardiac and respiratory, movement) In real time, so the Doctor can receive all information data about his patients in distance and be notified in case of troubles or emergency.

In the next chapter we will talk about the different forms of data and where it will be stored.

III. Types of Data and Data sources

Big medical data are so wide, therefore there is different forms of data; structured as well as unstructured, in the majority of cases we will have Digital imaging like MRI or X-ray scan copies, documents and sensor data coming from the smart textile. Here is another challenge to implement this types of files in a wide storage.

A. Data lake[6]

A data lake is a data storage method used by big data. These data are kept in their original formats without any transformation. In a lake of data, therefore, we find data of different natures: structured data, notably from relational databases (rows and columns), semi-structured data (CSV, newspapers, XML, JSON ...), and unstructured data (emails, documents, PDF) and blob files (images, audio, video in particular).

Here the data are stored without any schema required, without any transformation, they are kept in their original forms. the transformation comes in the last step. Data lake works according to the processing approach of Schema on read, differently to the classical alternative adopting Schema on write.

- Schema on write is the classical way to process the data, the terminology used is ETL (Extract, Transform, Load) the most used process for storing and analyzing data like in data warehouses and classical databases.
- Schema on read is different from schema on write because schema is created only when we want to read the data. here we use the process ELT (Extract, Load, Transform) the first step is extracting data from one or more remote sources, then loading them into the target data warehouse without changing the format, and the transformation comes in the data reading part, here isd the principle of the Data Lake.

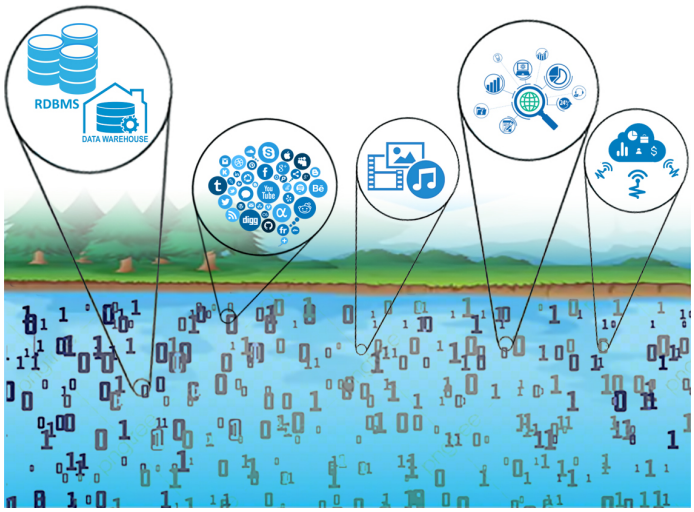

Fig. 2. Representation of the data lake.

B. Digital Imaging and COmmunications in Medicine DICOM

DICOM (Digital Imaging and Communications in Medicine) is the international standard to transmit, store, retrieve, print, process, and display medical imaging information.[7] It also defines the file format, which is for the most part either a DCM or DCM30 file extension.

DICOM has been executed in pretty much every restorative imaging methodology, for example, MRI, CT, X-beam, Ultrasound, Camera and so on. Consequently, DICOM guarantees that all the medicinal gear introduced in facilities, medical imaging centers, and emergency clinics will cooperate and distribute the medical images correctly.

Many medical imaging softwares work in this field, to manage and view this documents:

1) DicomBrowser[8]: DicomBrowser is one of the most popular and used DICOM viewers on Linux,it is an open-source software developed at the Washington University by the Neuroinformatics Research Group. This medical imaging tool can stack a huge number of pertinent medical images simultaneously.

2) Mango: Another powerful medical imaging software that supports both JAVA and Python API developments, developed with analytics tools for medical professionals.

3) Ginkgo CADx: One of the excellent medical imaging software, in addition to its functionalities it can convert different file formats(PNG, JPEG, BMP, PDF, TIFF) into DICOM files.

4) NextCloud: NextCloud is a cloud-based medical imaging software for Linux. The highlight of NextCloud that it's machine independent and works on the cloud entirely.

To store and manage all this volume of data with its various types, certainly there is another kind of platforms that can provide new characteristics and new functionalities.

IV. Hadoop ecosystem

Big data needs a new kind of technologies that can maintain all its requirements, many solutions are existing and we are

choosing Hadoop Platforme as a solution to manage and to process our big data coming from various sources (sensor data, documents, databases, doctor's information).

Hadoop is an Apache open source framework developped on Java that allows distributed processing of large dataset across cluster of computers using simple programming model. Hadoop creates cluster of machines and coordinates work among them. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage Hadoop consists of various components (HDFS, YARN, MapReduce) and services (ingesting, storing, analyzing, and maintaining) inside of it.[8]

A. Hadoop distributed File System HDFS

It is the core of Hadoop Ecosystem, HDFS being the storage layer, it allows to store different types of data sets (i.e. structured, unstructured and semi structured data). here the data are stored across various nodes; NameNode as a table of content or a log file and DataNodes which store the actual data.

- NameNode and also named Master Node
  Generally, we have one in a cluster. Its role is to maintain and to manage the DataNodes, and knows every block of data where it is stored in these Datanodes.
- DataNodes and also named Slave Nodes
  Here the actual data files are stored divided into blocks; the default size of each blocks is 128MB in Apache Hadoop 2.x (64MB in Apache Hadoop 1.x) this blocks are duplicated in 3 others Datanodes to avoid data loss in case the server goes down.

B. YARN

Refers to Yet Another Resource Negotiator, it has been projected as the resource management framework of Hadoop 2, Hadoop Yarn manages the efficiently and effectively of the resources.

The purpose of this resource management and task scheduling technology is to allocate system resources to different applications running in a Hadoop cluster. It is also used to schedule the execution of tasks on different parts of clusters.

also, additionally to MapReduce, YARN allow to other applications and processing engines to run Hadoop application such as Apache Spark.

C. MAPREDUCE

Being the processing layer of Hadoop. It process huge amount of structured and unstructured data stored in HDFS by using the Map and Reduce functions. It allows to divide the tasks of data processing on the various nodes of the cluster (Map function), then organizes them and gathers the results provided by each node in only one answer (Reduce function). These programs can be written in different languages : java, Ruby, Python or C++.

D. Other Framworks

Different components work in parallel with Hadoop, we are listing which we will use in our architecture:

- Apache Flume, come in the Data Collecting and Ingesting part. Its role is to collect unstructured and semi-structured data from its origin and send it to HDFS.
- Apache Sqoop, has the almost same role as Apache Flume but it work only on structured data such as RDBMS or Enterprise data warehouses.
- Apache Hbase is a Non-relational, distributed column oriented database built on top of HDFS system written in Java. This database has a significant fault tolerance, since the data is replicated to the different servers in the Data Center.
- Apache Hive is a data warehouse tool built on top of Hadoop. It allows to write HQL queries which is very similar like SQL to manage and process the data stored in HDFS. and it's mainly used for creating reports.
- Apache Pig, another component of Hadoop Ecosystem uses PigLatin language which is very similar to SQL.

We will bring together the different frameworks in the different phases in order to put in place a rigorous architecture answering to all the requirements of the new system.

V. ARCHITECTURE PROPOSED

Below we present our proposal architecture, including all the phases of acquisition and manipulation until the final utilization. Using the different technologies and tools inside Hadoop Ecosystem.

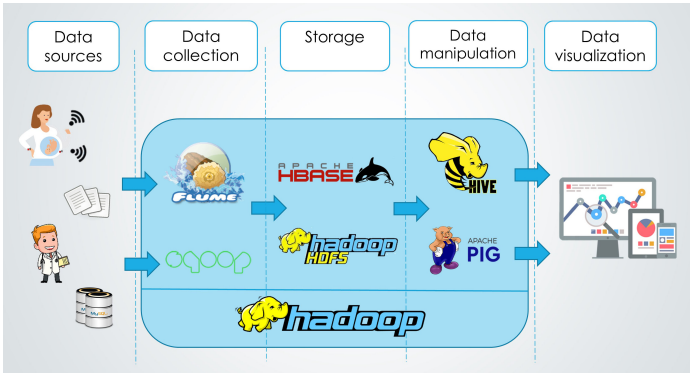

Fig. 3. Architecture proposed.

- Data sources: the first step for our flow of data, here we have different sources of data which are data sensors, documents, doctor's information, and different databases.
- Data collection: the phase of collecting and transforming data from different sources into HDFS.
- Storage: the core layer where all types of data will be stored.
- Data manipulation: Hive and Pig are the component chosen to manipulate and process the different types of datasets.

- Data visualisation: the last but not least, the important step of our work, here we have all important information in form of graphs and reports.

VI. CONCLUSION

This paper describes the big data environment with the different components based on Hadoop Ecosystem that contributes in the realization of a system of Remote Patient Monitoring based on sensor data, medical history, Medical Imaging, the data environment (climate, season, city) etc., using different layers and different frameworks.
This system is able to generate graphs and reports, also alerts and notifications in case of emergency. It makes it easier for doctors to follow their patients.

REFERENCES

[1] R.Mazhar, A.Awais and P.Anand, "The Internet of Things based Medical Emergency Management using Hadoop Ecosystem"2015 IEEE
[2] Gartner Research Cycle 2014, http://www.gartner.com
[3] Frost and Sullivan, U.S. Hospital Health Data Analytics Market, 2012
[4] P. Groves, B. Kayyali, D. Knott and S. Van Kuiken. The big data revolution in healthcare. M cKinsey & Company. 2013
[5] J. Dixon "Pentaho, Hadoop, and Data Lakes" Blog Entry: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-anddata-lakes/
[6] Digital Imaging and COmmunications in Medicine DICOM https://www.dicomstandard.org
[7] Washington University Neuroinformatics Research Group. DicomBrowser, version 1.5.2. Washington University website. nrg.wustl.edu/software/dicom-browser/. Published February 23, 2012. Accessed December 19, 2013
[8] Apache Hadoop: http://Hadoop.apache.org
[9] Hadoop Distributed File System, http://hadoop.apache.org/hdfs
[10] YARN Yet Another Resource Negotiator: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
[11] MapReduce: http://hadoop.apache.org/docs/r2.9.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html