Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-707

Integrating Medical and Research Information: a Big Data Approach

Carlos M. TILVE ÁLVAREZ[a,1], Alberto AYORA PAIS[a], Cristina RUÍZ ROMERO[b], Daniel LLAMAS GÓMEZ[a], Lino CARRAJO GARCÍA[a], Francisco J. BLANCO GARCÍA[c] and Guillermo VÁZQUEZ GONZÁLEZ[a]

[a] *Avances en Telemedicina e Informática Sanitaria (ATIS). Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC). As Xubias, 15006. A Coruña, España.*

[b] *PRB2-ProteoRed/ISCIII, CIBER-BBN. Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC). As Xubias, 15006. A Coruña, España*

[c] *Servicio de Reumatología. PRB2-ProteoRed/ISCIII. Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC). As Xubias, 15006. A Coruña, España*

Abstract. Most of the information collected in different fields by Instituto de Investigación Biomédica de A Coruña (INIBIC) is classified as unstructured due to its high volume and heterogeneity. This situation, linked to the recent requirement of integrating it to the medical information, makes it necessary to implant specific architectures to collect and organize it before it can be analysed. The purpose of this article is to present the Hadoop framework as a solution to the problem of integrating research information in the Business Intelligence field. This framework can collect, explore, process and structure the aforementioned information, which allow us to develop an equivalent function to a data mart in an Intelligence Business system.

Keywords. Business Intelligence System, Big Data, Data Mart, Unstructured Information.

Introduction

In recent times, medical information compiled in hospitals has been growing exponentially on their databases. However, all of this information structured on databases did not provide, by itself, knowledge about the state of the hospitals. This situation has marked an inflection point in the development and implantation of Business Intelligence (BI) [1] systems in the medical field. This type of systems channels all the information and manage the possibility of extracting knowledge from data analysis of the medical field, which favours support systems such as information visualization as a whole, support systems for the decision making or the discovery of

[1] Corresponding author: Carlos.Manuel.Tilve.Alvarez@sergas.es

new models of information that cannot be extracted explicitly from this data. All the knowledge generated on these systems may affect immediately if it is applied in every field of an organization according to its different natures: increasing quality or optimizing different procedures developed in each field, minimizing costs, improving services in medical consultations and even providing suggestions for the patient's care.

Once the BI architecture has been implanted in the medical field of Complejo Hospitalario Universitario de A Coruña (CHUAC), the purpose of integrating information from research processes from different fields made by INIBIC was born in the wake of the DIPROA project, which is being developed at the present time by the companies Applied Mass Spectrometry Laboratory, SolidQ Global and Altia Consultores. Specifically information generated, on a daily basis, in the fields of genomics, proteomics, cell cultures and histomorphology. Most of this information comes from different instrumentation devices that can be classified as pure data, like in the case of mass spectrometers or a DNA sequencer, or as image information. The rest of the information corresponds to the record of patient's samples and processes developed, that can be stored in plain text files or have some sort of format (such as Excel).

Our estimation during the planning in the development of *ad hoc* tools, both for compiling and structuring as well as processing, obtained as a result a high cost and an important use of resources and time. Moreover, this option would suppose the imposition of using these developing tools since in the future the system will cover more fields of the INIBIC. We had to dismiss this development as it was too rigid and expensive.

1. Methods

After analysing the possibilities that the different options from the Big Data scope had to offer, Hadoop framework [2] showed up as the most complete option to fulfil the necessary requirements in order to manage unstructured information generated on the INIBIC fields.

We have chosen the open-source solution Apache Ambari [3] to deploy a prototype of a Hadoop cluster over four nodes with the operating system CentOS. Ambari helps deploying Hadoop framework, along with many other tools from its ecosystem, on a cluster and provides an interface quite intuitive to work with on managing and monitoring.

The central core of Hadoop is formed by two main components: the data storage system Hadoop Distributed File System (HDFS) [4] and the task processing engine MapReduce 2.0 (MRv2 o YARN) [5]. These components fulfil, respectively, requirements needed for storing and processing the data.

However, not all requirements can be achieved by the Hadoop core by itself, so it is necessary to install some tools from its ecosystem to complete the architecture we want to develop. In this way, and in order to give a structure to the information already registered and make it accessible for the BI system, we decided to install the Apache Hive [6] tool: a data warehouse focused on giving a structure to the information and that facilitates queries over the set of data inside the data storage system HDFS.

At the moment of the analysis and design of the phase of information compilation from different fields, in order to import it to HDFS, we have classified the information in two different groups: first information stored on local databases derived from the

acquisition of different devices and that can be accessed directly from the cluster; second, information compiled from different sources and that have to be processed before importing it. In the first case, Hadoop ecosystem provides a tool called Sqoop [7], which would be used from the master node of the cluster in order to store a wide range of gross information from different local databases. In the second case, we found it necessary to use a more complete ETL (Extract-Transformation-Load) tool as we had to cross information from different sources and types. We selected Pentaho Data Integration (PDI) [8] software which allows us to schedule information flows in order to gather and transform data before storing it.

Finally, and even when it was not absolutely necessary for our architecture, we saw that the possibility of creating knowledge on this specific component of the BI system would suppose an improvement regarding to the rest of data marts. We used the MLlib library [9] for the data process engine at large scale in Apache Shark. This library allows us to implant machine learning algorithms, specifically those related to classification and clustering. Knowledge created from these algorithms is added to the data group already accessible in the BI system.

2. Results

As result we have obtained the architecture shown in figure 1.

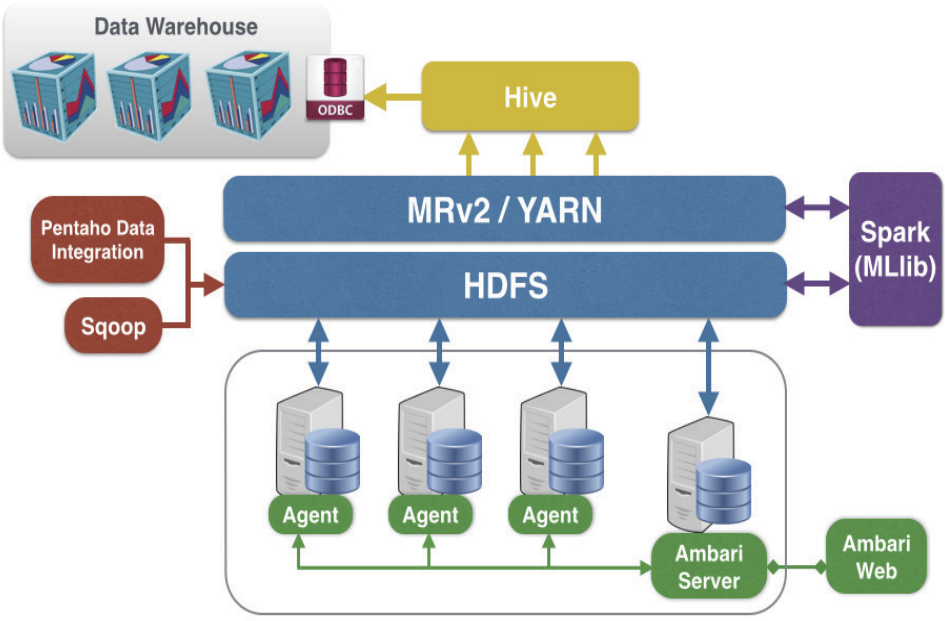

Figure 1. Final configuration of the Hadoop architecture.

By using Sqoop tool (to import information directly from databases in different devices) and PDI, as a more complete ETL tool, we would cover all the possible options to add information into the HDFS system.

The implementation of this architecture offers a data storage system with the following features:

- Distributed.
- Low latency.
- High fault-tolerance, as the information is replicated between the nodes.
- Horizontal scaling; when a new node is added, HDFS rebalances the charge which implies that more nodes can be added in order to enlarge its storage capacity on demand.

The processing component of Hadoop offers the possibility of generating an information process on a large scale over the cluster, by using the strategy "divide and conquer" corresponding to the programming paradigm MRv2. Another available option of processing is the use of the MLlib library from the Spark engine, which can be collocated in the cluster deployed in Hadoop with a better performance than the one in MRv2. All the algorithms related to data transformation or generation have been implemented using one of these two alternatives.

With the last component, Hive, we have structured the information: from data storage file system provided by HDFS to a table structure similar to the entity-relationship model from relational databases. By means of this component, all the information that we want to load from BI is made accessible. In order to perform the load, Hive allows establishing the connection through an ODBC driver (Open Database Connectivity), which has been previously installed in our BI system.

With this environment the information flow is defined, from the various sources, where it is generated, to its structure to bestow accessibility to the BI system.

Discussion

As we have mentioned before, the development of ad hoc tools to manage information from different fields of the INIBIC was a problem due its high cost and lack of flexibility. The running of the Hadoop framework and the tools from its ecosystem as used in the environment to process information presented in this article, fulfils the goal of making available to the BI system all the unstructured information generated on the different fields of the INIBIC.

Once this component is added to the BI system, presented as an equivalent to a conventional Data mart, it will allow us to apply techniques of Data Mining (DM) over the whole group of data from the medical-research field for the first time. In this way, we have created the possibility of analysing and visualizing information from both sides as a single block of information.

As our main objective was to incorporate research data into the BI system in order to exploit it, gathering data research from different fields in just one point favours the aforementioned exploitation by the researchers. Hadoop framework offers even the possibility of using DM libraries, as the previously mentioned over Spark to perform machine learning, and also the connection of HDFS systems as data source to other environments in order to create statistics researches or graphics, as it can be seen in the

statistics language R [10]. These options are accepted by researchers as a support tool to back up with data at the beginning of new lines of investigation.

Finally, in regards of a future expansion, it should be pointed out that Hadoop framework is flexible enough and it can be easily adapted to any field from the INIBIC that could be incorporated, as it can define new workflows or add new information to the ones already existing.

Acknowledgments

DIPROA has been funded by CDTI (Centro para el Desarrollo Tecnológico Industrial) and the ERDF (European Regional Development Fund), thus being backed by the Ministerio de Economía y Competitividad of Spain and by the Consellería de Economía e Industria of the Xunta de Galicia through the GAIN (Axencia Galega de Innovación).

References

[1] Olszak, C & Ziemba, *Approach to Building and Implementing Business Intelligence Systems*, Interdisciplinary Journal of Information, Knowledge, and Management, vol. 2, 135-148.
[2] Apache Hadoop [Online]. Available: http://hadoop.apache.org/index.html [2014, October 22]
[3] Apache Ambari [Online]. Available: http://ambari.apache.org [2014, October 22]
[4] Hadoop Distributed File System [Online]. Available: http://hadoop.apache.org/docs/r2.5.1/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html [2014, October 22]
[5] MapReduce 2.0 [Online]. Available: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html [2014, October 22]
[6] Apache Hive [Online]. Available: https://hive.apache.org [2014, October 22]
[7] Apache Sqoop [Online]. Available: http://sqoop.apache.org [2014, October 22]
[8] Pentaho Data Integration [Online]. Available: http://community.pentaho.com/projects/data-integration [2014, October 22]
[9] Apache Spark – MLlib [Online]. Available: https://spark.apache.org/mllib [2014, October 22]
[10] Project R [Online]. Available: http://www.r-project.org [2014, October 22]