Enhancing EHR Systems Interoperability by Big Data Techniques

Nunziato Cassavia, Mario Ciampi, Giuseppe De Pietro, and Elio Masciari(✉)

ICAR-CNR, Rende, Italy
{nunziato.cassavia,mario.ciampi,
giuseppe.depietro,elio.masciari}@icar.cnr.it

Abstract. Information management in healthcare is nowadays experiencing a great revolution. After the impressive progress in digitizing medical data by private organizations, also the federal government and other public stakeholders have also started to make use of healthcare data for data analysis purposes in order to extract actionable knowledge. In this paper, we propose an architecture for supporting interoperability in healthcare systems by exploiting Big Data techniques. In particular, we describe a proposal based on big data techniques to implement a nationwide system able to improve EHR data access efficiency and reduce costs.

Keywords: Big data · Healthcare · Interoperability

1 Introduction

Nowadays, the availability of huge amounts of data from heterogeneous sources, exhibiting different schemes and formats and being generated at very high rates, led to the definition of new paradigms for their management – this problem is known with the name *Big Data* [3–6]. As a consequence of new perspective on data, many traditional approaches to data analysis result inadequate both for their limited effectiveness and for the inefficiency in the management of the huge amount of available information.

Therefore, it is necessary to rethink both the storage and access patterns to big data as well the design of new tools for data presentation and analysis. It is worth noticing that the problem of fast accessing relevant pieces of information arises in several scenarios such as world wide web search, e-commerce systems, mobile systems and social networks analysis to cite a few. Successful analyses for all the application contexts rely on the availability of effective and efficient tools for browsing data so that users may eventually extract new knowledge which they were not interested initially.

In this respect, also healthcare stakeholders have access to challenging knowledge integration and extraction problems. This information can be classified as big data, as they exhibit impressive volume and they are really heterogeneous and time varying. Moreover, pharmaceutical-industry experts are interested to analyze big data to obtain useful insights on their data. Although these efforts

M.E. Renda et al. (Eds.): ITBAM 2016, LNCS 9832, pp. 34–48, 2016.
DOI: 10.1007/978-3-319-43949-5_3

are still in their early stages, they could help for providing people better healthcare quality and reducing costs. As an example, it is possible to analyze patient data for understanding what treatments are most effective for particular conditions, identifying patterns related to drug side effects or hospital readmissions, and gaining additional important knowledge [18].

Indeed, a relevant research issue is the design and implementation of a distributed platform for accessing heterogenous Healthcare Information Systems (HIS) so as to enlarge their coverage and allowing the availability of medical data at heterogeneous, and geographically-sparse, healthcare providers by enabling data sharing in a seamless manner. In this paper we propose an architecture that overcomes the available architectures for federating HIS, as current solutions suffer the limitation of only allowing a communication style according to a pull-based data delivery, i.e., the user requests a clinical document knowing its unique reference. On the contrary, we take advantage of big data architectural advantages for offering a reliable solution that can be used also by non IT experts. In particular, we exploit the well know MapReduce framework in order to offer advanced querying capabilities for medical data.

2 Background and Related Work

2.1 Basics on Tools Supporting the Management of Big Data

Nowadays, dealing with a big volume of data is a very difficult challenge, since traditional technologies, like RDBMS, are not suited for this purpose. Many open source technologies were developed in order to handle massive amounts of data. The majority of these technologies are based on the MapReduce programming model. This paradigm make it easier to implement solutions based on the use of distributed systems for executing data mining tasks.

The MapReduce paradigm is based on the following steps:

- *Map*: Each node executes the *map* function on its local data, creating a set of pairs $\langle key, value\rangle$, and stores the results in a temporary storage.
- *Shuffle*: Pairs $\langle key, value\rangle$ are redistributed among nodes, in such a way that all the pairs with the same key are assigned to the same node.
- *Reduce*: Each node processes its group of pairs, independently of other nodes.

It is worth noticing that since each mapping operation does not depend on the others, mapping operations can be parallely executed. In a similar way, also the reduce step can be performed by multiple nodes at the same time, if the reduction function is associative.

The most widespread implementation of the MapReduce programming model is Hadoop MapReduce, part of the Hadoop framework [23]. Although Hadoop is a really pervasive technology, it has its drawbacks, especially with clustering (or in general with machine learning) algorithms based on iterative operations. This is because Hadoop MapReduce stores the results of intermediate computations on disk. The overhead to launch each job, moreover, is very high. MapReduce is well suited for large distributed data processing where fast performance is not an issue. Its high-latency batch model, instead, is not effective for fast computations or real data analysis.

A widespread tool for Big Data application design is Apache Spark [25], a framework optimized for low-latency tasks. Spark caches data sets in memory and has a very low overhead in launching distributed computations. As stated in Spark website, Spark can "run programs up to 100 times faster than Hadoop MapReduce in memory, or 10 times faster on disk." In multi-step jobs, moreover, Hadoop MapReduce blocks each job from beginning until all the preceding jobs have finished. This can lead to long computation times, even with small data sets. There are other ways to schedule tasks, one of which is *Directed Acyclic Graphs (DAG)*. A graph is used, where the vertices represent the jobs and the edges specify the order of execution of the jobs themselves. Since the graph is acyclic, independent nodes can run in parallel, resulting in a much lower overhead compared to the traditional MapReduce. Spark offers capabilities for building highly interactive, real-time computing systems using DAGs and so is very suitable to implement applications which require an high level of parallelism. Spark is built against Hadoop in order to access *Hadoop Distributed File System (HDFS)*. The key concept beyond Spark is called *Resilient Distributed Dataset (RDD)* [24]. An RDD is a read-only, partitioned collection of records. Data are partitioned across many nodes in the cluster. Fault tolerance techniques are used to avoid data loss due to node failures. Given an RDD we can manipulate the distributed data through operations called *transformations* and *actions*. Transformations consist in the creation of new data set from an existing one, and actions in running a computation on the data set and returning the results to the driver program. We recall that, in the MapReduce paradigm, map is a transformation, reduce is an action. We point out that, in our architecture we will leverage MapReduce for effectively indexing data.

2.2 Evolution of HIS

A HIS [16] is an information system with the aim of capturing, storing, managing or transmitting information related to the health of individuals for contributing to a high-quality and efficient healthcare. Three generations of HIS have evolved in the last decades. The first generation consisted in HIS limited within small facilities, such as departments of hospitals. This type of HIS manages the digitized form of medical documents, such as images or reports created by means of editing programs. A practical example is Radiology Information System (RIS) [17] for storing and managing radiology-related documents. The second generation, born in the 1990s, concerned the integration of such departmental information systems so as to support combined information processing in the hospital. A first example of such integrations was the so-called Electronic Medical Records (EMR), which are legal records created in hospitals and ambulatories, including documents and images, which are consulted by healthcare professionals from a single organization. Another examples is the Picture Archiving and Communication System (PACS) [17], a system for managing and communicating medical images, often integrated with the systems of different departments within the hospital, such as RIS. Such evolution contributed to increase the size of the HIS and the amount and diversity of the exchanged data. In fact, the transition from the first two generations imposes the resolution of technical and syntactical interoperability, i.e., technological and protocol compatibility and diversity in formats of medical data. The DIOGENE project [11], which integrates all patient-related information so as to obtain a seamless communication between hospital actors, is an example of this transition.

The current third generation consists in integrating the hospital-wide HIS so as to form regional HIS, and in federating the these ones so as to have national and trans-national HIS. In this context, Electronic Health Records (EHR) represent a subset of the EMRs issued by each healthcare provider that took care of the given patient during his/her clinical history. These systems permit to share medical information about patients and to have patient-related information following him/her through the various healthcare providers in a given region or country. Practical examples are the Clinical Data Repository/Health Data Repository (CHDR) [12] and the epSOS project. The first one consists of interconnecting all the offices belonging to the Department of Defence and the Veterans Affairs over the overall territory of United States. The second one aimed at designing and developing a service infrastructure supporting the interoperability among every national HIS in several European countries. epSOS is connected to similar initiatives running in the European countries participating to the project, for integrating their regional HIS. The evolution towards HIS of third generation has the consequence of increasing system size in terms of number of interconnected components and amount of exchanged data, but it also exacerbates the interoperability issues to be addressed. Specifically, the transition towards the third generation adds also semantic and business Interoperability, that are common information models/terminology, and common business processes. Figure 1 summarizes the described features.

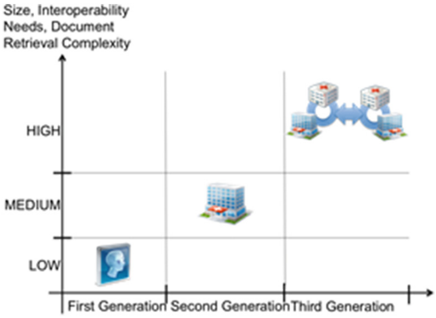

Fig. 1. Systems classification

Interoperability has always been considered as a key challenge to be faced in the described evolution of HIS. since the transition from the first to the second generation, there has been the need of defining standard formats for medical imaging storage and transmission, bringing to the specification of Digital Imaging and Communications in Medicine (DICOM). With the the third generation, as mentioned, there has been the need of specific solutions and guidelines for driving an interoperable HIS interconnection. In the last years, different solutions and guidelines have been proposed and formalized. The first example is represented by the international not-for-profit Foundation called openEHR, which issued a detailed and tested specification for an interoperable HIS platform. Such a vision of openEHR had a significant influence on the development of the emergent healthcare industry standards, such as Health Level 7 (HL7) and CEN EN13606, with recommendations for an interoperable interconnection of HIS. HL7 has defined numerous specifications for enabling interoperability among health applications: among others, two different relevant specifications are Clinical Document Architecture (CDA), based on HL7 v3, which defines the XML schema (format) of exchanged medical documents; the recent Fast Healthcare Interoperability Resourse (FHIR), based on the evolution of HL7 v3, specifies a large, pictorial, representation of medical data in resources. EN13606 represents a subset of openEHR [21], with a specification of the data exchange issues and not for a full federated HIS of the third generation, which is contained in openEHR. HL7 CDA and FHIR support instead syntactic and semantic interoperability by introducing a common model for exchanged medical data. Nevertheless, the history of healthcare interoperability of the last three decades has shown that healthcare standards are not sufficient alone to ensure interoperability. Indeed, they include many if not all the possible situations, thus suffering from various ambiguities and offering many choices that hamper interoperability [13]. To address these issues, the Integrating the Healthcare Enterprise (IHE) initiative has specified some integration profiles, like Cross-Enterprise Document Sharing (XDS). IHE XDS aims at facilitating the sharing of clinical documents within an affinity domain (a group of healthcare facilities that intend to work together) by storing documents in an ebXML registry/repository architecture. In a similar way than the other standards, IHE XDS needs to be localized by specifying an affinity domain, that is the formalization of the set of policies, codes and rules shared by the facilities working together [14]. HIS solutions of the first generation were basically in-house and ad-hoc programs for archiving and retrieving medical documents, tailored on the peculiarities of the given system in terms of type of managed data and hardware configurations. When moving towards HIS of the second generation, the implementation involved the use of well-assessed and -established middleware technologies such as CORBA or DCOM, due to their ability of resolving issues related to technological interoperability. A practical example is represented by the CORBAMed initiative [2], which presents a set of domain-specific services expressed within a specific CORBA domain for the medical environment. Also WS technology is used within the context of the second generation, such as WebCIS described in [22], thanks to the ability of XML-based communication to deal with syntactical interoperability, as proved in [1]. For the third generation, the preferred solution is to use Web Service technology since it has demonstrated a high interoperability capacity and the flexibility to integrate already-existing legacy systems. The previously mentioned standards do not demand a specific technology for implementing federated HIS; however, they recommend SOAP communications for exchanging medical data. This has brought key stakeholders to drive the technological choice towards WS. As a concrete example, we can cite the mentioned epSOS project and CHDR, which are implemented by means of WS. However, the current research has also moved towards different kinds of middleware solutions, such as the Tuple Space-based infrastructure described in [19]. At the moment, the products implemented by these recent research efforts have been scarcely applied in concrete real usage. Although XML-based communications resolve syntactic interoperability, they represent only a pre-requisite to the semantic one, and proper additional mechanisms are needed. In the current literature, Semantic Interoperability is typically addressed by means of proper ontologies integrated within the communication system to provide the defining concepts of the given domain [8]. Such a solution has been investigated within the context of healthcare [9], and practically adopted in [19]. Although the foundation ontologies for healthcare have been developed by academic research, none of them have been adopted in concrete applications. In fact, [10] arguments that ontologies for healthcare are not mature solutions, yet. This can be also seen if we study the previously-mentioned standards: only openEHR has specified an ontology to be used in medical data sharing. The most common solutions, i.e., the one adopted by all the other standards and in practical use cases such as epSOS, is to adopt a reference common model for the communications among HIS, as the one specified in HL7 RIM, and a set of mediators for translating from/to such a common model towards the one adopted by each specific HIS.

Securing web services has been an active research topic in the last decade and has been standardized in the OASIS specification called WS-Security [20], which is a composite standard made by combining other different specifications and methods, and specifies two different levels of mechanisms to enforce the provided security level: (1) the first is implemented at the message level by defining a SOAP header that carries out extensions to security; (2) the second is realized at service level to perform higher-level security mechanisms such as access control or authentication. In particular, at the message level we can find two main XML security standard techniques that can be introduced in the mentioned SOAP header extensions: XML Signature and XML Encryption. The former aims at having a small portion of the XML content digitally signed (such element is called digest) so as to provide integrity and non-repudiation for the overall XML content. On the other hand, the latter has the goal to encrypt a part of the overall XML content by using a certain key, which can be public or private according to the chosen encryption strategy. In the case of WS-Security, the SOAP header has a given field, called DigestValue, to contain the digest with indications of the adopted signature method. If encryption is used, the SOAP header has to contain the adopted key, which is itself encrypted by using a proper public key. Besides these two important message-level methods, we have an additional one: Secure

Socket Layer (SSL), which realizes a secure form of the TCP transport protocol, by offering mechanisms for the key agreement, encryption and authentication of the endpoints in a connection-oriented communication. On top of these message-level mechanisms, we can find service-level ones: (i) Security Assertion Markup Language (SAML) is a framework to exchange authentication and authorization information in a request/respond manner when the communication participants do not share the same platform or belong to the same system. The core of this framework is the assertion, expressed with XML constructs, containing the identity of the requestor, and the authorization decisions or credentials. (ii) Extensible Access Control Markup Language (XACML) is used to specify roles and policies used by an access control mechanism to infer the access decisions for users. Different HIS can adopt their own access roles and grants, and XACML is used to exchange such decisions among HIS and to orchestrate their access decisions. (iii) Last, we can find two other specifications: Extensible Rights Markup Language (XrML) and XML Key Management Specification (XKMS). The first is used to express rights and conditions related to the access control (such as expiration times); while the second defines interfaces for the distribution of keys used in XML Signature and XML Encryption.

Security is a key issue in HIS, and the review in [7] provides a complete view of the research efforts spent and achievements obtained so far. As a matter of fact, few works focus on architectures and frameworks, but more focus has been given to qualitative research, modeling and economic studies. Based on these research activities, few prototypes have been realized [15].

3 Advanced Data Search

3.1 Complex Search

Nowadays, the availability of huge amounts of information calls for proper solutions to the complex search problem. To this end, search engines have been proposed since the early stage of Internet. However, results returned by search engines are often quite far from the expected query answers from a user viewpoint. Indeed, search results can be improved by building a custom map that, based on the initial query results, tries to learn additional knowledge about data being queried by iterative refinement of search dimensions and parameters. Figure 2 shows the above mentioned scenario.

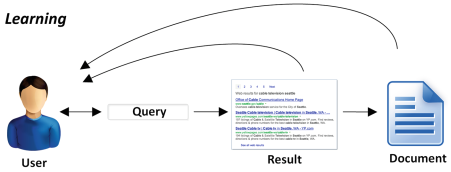

Fig. 2. Learning by results

In this scenario, the type of query being performed plays a crucial role. Unfortunately, this process is suitable only for simple search of well-defined terms. On the contrary, dynamic learning by exploratory research cannot be performed by this naive process. Obviously enough, for well defined queries, a search engine like Google, is able to provide correct results in a few milliseconds[1].

However, in some cases users do not know exactly how to find the desired information about an object or a service (e.g. a book or a restaurant). In this case, the model depicted in Fig. 3 is more suitable.

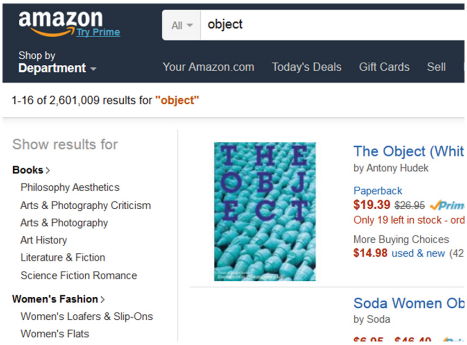

Fig. 3. Amazon search

More in detail, Amazon-like search tools, feature product categorization and recommender systems, thus making the user search experience quite interactive and iterative. In a sense, intermediate results guide users to a better definition of target information. Furthermore, search engines usually allow non-structured queries (referred as "ranked retrieval") whose results are sorted according to some relevance criteria w.r.t. the target search. As a matter of fact, these queries are easier to pose by users compared to boolean expressions, but they can produce low quality results.

In order to overcome this limitation, some categorization service like Yahoo!Directory, exploits context information[2]. More in detail, directory contents are hierarchically organized in order to guide users through a subset of documents potentially related to information being queried, thus limiting the possibility to input free text queries. In this respect, users re-think and refine their needs by learning the adjustments to the search being performed by exploiting the available choices. To better understand how directory navigation works, we resurge to accommodation booking portals analogy. Indeed, those portals offer a hierarchical navigation systems, i.e. from the home page, user can choose

[1] As a matter of fact, due to its quick result presentation, many users go through Google even if they exactly know the URLs of the resources they are interested in.

[2] Yahoo!Directory is no longer active since 2014, however it is worth mentioning as it was one of the first services for massive assisted browsing.

the desired country, then s/he can specify the city and finally the type of structure s/he is interested in. This navigation model suffers a great limitation due to taxonomy specification. Indeed, taxonomy specified by the service designer may not meet user needs. A solution to overcome the above mentioned limitations is the implementation of *faceted* navigation that helps users in the information “surfing” process.

4 A Big Data Architecture for Supporting EHR

In this section, we describe the overall architecture of our proposal for assisting medical data search. Our goal is to provide users a flexible tool for assisted text search, that is interactive, scalable and dynamic in order to easily connect and integrate all the nodes of the healthcare infrastructure. To this end, we exploited several indexing and data management strategies that allow us to cope with high volume, heterogenous and burst information. Figure 4 shows our system architecture.

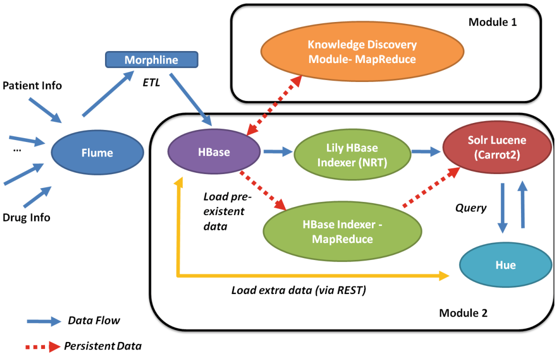

Fig. 4. System architecture for big data search (Color figure online)

In order to guarantee maximum implementation flexibility, we exploited several open source tools as Apache Hadoop[3], Flume[4], HBase[5], Solr[6], Lily HBase Indexer[7] and Hue[8].

Data are collected in our system from heterogeneous sources and they arrive in a streaming way. In order to properly manage these data, we implemented some specialized *Crawling* services that are closely tied to the data set being

[3] http://hadoop.apache.org.
[4] https://flume.apache.org/.
[5] https://hbase.apache.org/.
[6] http://lucene.apache.org/solr/.
[7] http://ngdata.github.io/hbase-indexer/.
[8] http://gethue.com/.

collected. As data are crawled, we collect them by *Flume* module (note that the blue arrows in Fig. 4 refer to data flows arriving at different rates from multiple sources). This module will host our staging area, as it is a reliable and distributed service designed to efficiently collect, aggregate and forward huge amounts of data for later storage in a permanent repository. Furthermore, it is well suited for dealing with data streams, as it provides fault tolerance by an easily configurable reliability mechanism that are mandatory for managing healthcare data. The latter feature has been profitably exploited for dealing with data inconsistency due to null or missing values that could arrive from multiple data sources.

Once data are gathered by Flume module, they undergo through an “on the fly” ETL (Extraction, Transformation and Loading) process performed by *Morphline* module. This module is devoted to data cleaning and data mapping on the column set in the datastore. The output of this step is a cleansed data flow on top of which our analysis takes place. Collected information are sent to our big data storage layer, implemented by *HBase*, that is devoted to data storage.

As stated above, our goal is to improve full-text search. To this purpose, we exploit *Apache Solr* tool for our discovery task (more precisely we used the *SolrCloud* implementation). The rationale for exploiting *Solr* is twofold: 1) it allows effective and efficient searching for keywords appearing in any column that has been previously indexed and 2) it allows to faster display documents ranked by their relevance w.r.t. the query being issued. Moreover, *Solr* provides several useful presentation features as: field facets, range queries and pivot facets that allow a proper organization of the results to be shown to the user. Those operators can be also fruitfully exploited for providing users the classical on line analytical processing operators as slice & dice, drill-down, roll-up and pivoting. In this respect, *Solr* has been proven to be an excellent real-time analysis engine for text documents (like user queries and suggestions exploited in our system). Consider, as example, web site logs: *Solr* can easily indexed them in order to execute (time-stamped) range queries for a given (set of) keyword(s). Moreover, it is possible to build the information graph containing aggregate information, such as the growth over time of registered users or transactions grouped by type. We exploit these information in our system for providing better suggestions. As users interact with our system, e.g. by searching new information or by posting new documents related to a disease, new data are collected by the storage layer. Based on data arrival rate, we schedule offline clustering of the whole dataset (e.g. after a burst of tuples is collected) in order to better organize data and for boosting the indexing strategy assisted by ad-hoc MapReduce functions.

In order to properly display search results, we exploit *Hue* features. The latter, is a software offering a customizable user friendly interface.

5 Case Study: A Nationwide System

In this section we present our case study, i.e. a nationwide system. In particular, we will first show the law requirements to be met.

5.1 Law Requirements

A recent Italian decree establishes security and organizational requirements that regional EHR systems (EHR-S) must provide:

- Implementation of a set of functions based on a shared functional model. Specifically, a common functional model for EHR-S, obtained localizing the HL7/ISO EHR-S FM standard, has been defined by an interregional initiative, comprising regional representatives, some government agencies and associations (i.e. HL7 Italy), and research institutions. The functional model defined specifies, in a structured and integrated way, a set of business functions for the EHR, delegating implementation details about the realization of interoperable EHR-Ss. The profile, published by HL7 Italy, has been defined through an analysis of the existing laws, rules, work processes, and actors involved in the use of an EHR-S.
- Development of an enabling platform able to connect all the healthcare facilities distributed on the regional territory, in order to enable users to search, insert, and retrieve documents within their purview.
- Realization of services aiming at collecting health documents generated from health professionals. The mandatory documents that each EHR-S has to be able to handle are Patient Summary (PS) and Laboratory Report. To this scope, the software applications used by the health professionals have to be integrated with the regional platform.
- Implementation of a service that enables a health professional to identify a patient before he/she requests the system to access clinical documents.
- Integration of consent management services with the regional platform, in order to satisfy the legal constraint according to which health documents can be uploaded and consulted only if the patient has provided, respectively, two kind of informed consents: one aimed at making the patient able to express his/her intention to allow health professionals registering documents into his/her EHR; another allowing the patient enabling the consultation of his/her documents to all the health professionals that have the roles he/she authorized.
- Implementation of access policy management integrated with all the business services. EHR-Ss must satisfy the will of the patient, which is expressed through policy policies. It is therefore necessary to establish strict authentication and authorization policies for documents access.
- Implementation of interoperability services integrated with the regional platform in order to make this one able to interact with other regional platforms for (i) searching, (ii) retrieving, and (iii) registering health documents.

5.2 Experimental Infrastructure

An experimental interoperability infrastructure conformed to the Italian norms and technical specifications has been implemented with the scope of using the proposed approach for accessing EHR document managed by some Italian Regions (referred in the following as Region 1, 2 and 3).

The experimentation consists in enabling regional HIS to exchange medical documents related to some patients, available at the various healthcare facilities. In particular, the HIS of Region 1 and Region 2 have a similar architectural model that we cannot show in detail for regulatory issues. The operations experimented enable physicians to query and retrieve health documents of a patient which are available in another region, e.g. because in the past he/she has benefited from a health service in this region. The interconnection of the interoperability infrastructure with the regional HIS of the three regions has required a set of actions, described below.

The platforms of the Region 1 and Region 2, which share the same architectural model, have been integrated at the same way with the big data infrastructure. First, an Access Interface and an Indexing Strategy components have been deployed at each regional node of the infrastructure; in particular, the Indexing Strategy component interacts with the storage layer. Second, several instances of the Document Manager components have been deployed at a set of healthcare facilities. Such components have been integrated with the information systems of such facilities by means of wrappers able to translate the standard protocols with the ones used by the legacy systems. The actions performed for the integration of the HIS of the Region 3 with the big data infrastructure are: (1) development of a wrapper able to interconnect the Index Strategy component with the registry of the regional HIS, with the aim to translate the language used to represent the metadata of the shared information model with the one used by the local HIS; (2) implementation of a wrapper capable of interacting the Access Interface component with the legacy repositories of the healthcare facilities.

The healthcare metadata and documents related to a patient available in the regions different from those where a patient resides can be performed in two steps. The first step consists of a simple search: a user (e.g., a general practitioner) sends a query to the regional HIS, which propagates it to the Indexing module; it (i) makes the query to its own data store, (ii) interacts with the overall Indexing module (that is automatically maintained by our big data infrastructure) of the other regions, which executes the query to their registries, (iii) aggregates all the metadata results, and (iv) returns the results to the user. The second step is a document retrieval: the user selects a document s/he wants to obtain and sends a request to the regional HIS, which forwards it to the Access Interface component of the region containing the document; this one retrieves the document by communicating with the HIS of the region where it is deployed. It is worth noting that the regional platforms receiving the query and retrieve requests from the other regions have to be adequately processed and verified. The verification process consists in two phases: access control and information availability.

The first phase (access control) has the aim of verifying that the requesting user has the right to access the health documents he/she demanded. The access control system, based on an XACML architetcure, performs the following steps:

1. a Policy Enforcement Point (PEP) intercepts the message sent from the interoperability service of another region;
2. the PEP analyzes the claims transmitted in the messages in order to verify that they have all the attributes necessary (like the role of the user, the purpose of use, etc.). In particular, the claims are represented as assertions according to the SAML standard;
3. the PEP verifies the validity of the digital signatures of the SAML assertions;
4. the PEP controls if the patient has provided the opportune consents;
5. the PEP forwards the request to the Policy Decision Point (PDP);
6. the PDP provides the final decision (that is, permits or denies the access to the service). The decision is taken on the basis of the privacy policies established by the patient: the most important aspect to verify is that the role of the user contained in the assertions has the right to access the service. The final decision is transmitted to the PEP;
7. if the decision is a PERMIT, the PEP forwards the message to the appropriate service on the basis of the operation requested (that is, query or retrieve); if the decision is a DENY, the PEP returns an error message to the user.

The second phase is realized in case of a PERMIT response from PDP and is realized by the service invoked. In case of query, the service interacts with the registry in order to verify if there are entries that meet the search criteria indicated by the user: all the metadata satifying such criteria are returned to the user. In case of retrieve, the service interacts with the appropriate repository to obtain the health document satisfying the request: the document, if available in the repository, or an error message is sent to the user.

The quality of the search and retrieval operations have been tested with an experiment scenario consisting in a user in a given region that wants to obtain a document of a patient from a different region. Another experimentation has been made considering an intra-regional scenario, i.e., the user in a given region intends to obtain a document of a patient from the same region. With regards to the first scenario, we have performed from one region about 125 requests of search and retrieval operations of document identifiers randomly chosen among those hosted by the other two regions. The system returns two versions for each medical document requested: one according to the XML-based HL7 CDA Rel. 2.0 format of about 50 KB and the other one in PDF/A format of about 270 KB. We had the 100 % success rate (in terms of appropriate results provided to the user) for searched documents. With regards to the second scenario, the results have shown that almost half of the time is needed to retrieve the searched documents as we avoid the communication cost among regional nodes.

6 Conclusions and Future Works

This work presents a big data based architectural model aiming at enabling access to regional EHR systems, which have to be developed or revised according to recently issued specific Italian laws. The work represents an important first step in the process of digitizing the national EHR system. The proposed model is turning out to be successful for both Regions that have already started an e-health process and Regions that are still in a start-up phase. The efforts which have been made so far help the organizations to overcome the main difficulties to treat large amount of health data, highlighting the benefits that automated processes could bring in terms of time efficiency and care effectiveness. The solutions described in this paper are quite flexible as, on the one hand, they provide a standardized approach to ensure interoperable access to health data for processing. Anyway, due to regulatory issues we are still at the early stage of our experimental assessment as we need to overcome some of the above mentioned legal problems in order to fully exploit the potential of the proposed system. Thus, we are planning to define specific architectural components with the aim of performing anonymization operations when necessary.

References

1. Synapses/SynEx goes XML. Studies in Health Technology and Informatics, IOS press (1999)
2. (2001). http://healthcare.omg.org/Roadmap/corbamed_roadmap.htm
3. Big data. Nature., September 2008
4. Data, data everywhere. The Economist., February 2010
5. Drowning in numbers - digital data will flood the planet - and help us understand it better. The Economist., November 2011
6. Agrawal et al., D.: Challenges and opportunities with big data. A community white paper developed by leading researchers across the United States., March 2012
7. Appari, A., Johnson, M.E.: Information security and privacy in healthcare: current state of research. Int. J. Internet Enterp. Manage. **6**(4), 279 (2010)
8. Bittner, T., Donnelly, M., Winter, S.: Ontology and semantic interoperability. Large-Scale 3D Data Integration. CRC Press, London (2005)
9. Blobel, B., Oemig, F.: What is needed to finally achieve semantic interoperability? IFMBE Proc. **25**(12), 411–414 (2009)
10. Blobel, B., Kalra, D., Koehn, M., Lunn, K., Pharow, P., Ruotsalainen, P., Schulz, S., Smith, B.: The role of ontologies for sustainable, semantically interoperable and trustworthy ehr solutions. In: Medical Informatics in a United and Healthy Europe - Proceedings of MIE 2009, The XXIInd International Congress of the European Federation for Medical Informatics, Sarajevo, Bosnia and Herzegovina, Agust 30 - September 2, 2009, pp. 953–957 (2009). http://dx.doi.org/10.3233/978-1-60750-044-5-953
11. Borst, F., Appel, R., Baud, R., Ligier, Y., Scherrer, J.: Happy birthday diogene: a hospital information system born 20 years ago. Int. J. Med. Inf. **54**(3), 157–167 (1999)
12. Bouhaddou, O., Warnekar, P., Parrish, F., Do, N., Mandel, J., Kilbourne, J., Lincoln, M.J.: Exchange of computable patient data between the department of veterans affairs (va) and the department of defense (dod): terminology mediation strategy. J. Am. Med. Inf. Assoc. **15**(2), 174–183 (2008)
13. Dogac, A., Laleci, G.B., Aden, T., Eichelberg, M.: Enhancing ihe xds for federated clinical affinity domain support. IEEE Trans. Inf. Technol. Biomed. **11**(2), 213–221 (2007). http://dx.doi.org/10.1109/TITB.2006.874928

14. Dogac, A., Laleci, G.B., Kabak, Y., Unal, S., Heard, S., Beale, T., Elkin, P.L., Najmi, F., Mattocks, C., Webber, D., Kernberg, M.: Exploiting ebxml registry semantic constructs for handling archetype metadata in healthcare informatics. Int. J. Metadata Seman. Ontol. **1**(1), 21–36 (2006). http://dx.doi.org/10.1504/IJMSO.2006.008767
15. Esposito, C., Ciampi, M., De Pietro, G., Donzelli, P.: Notifying medical data in health information systems. In: Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems. pp. 373–374. DEBS 2012, NY, USA. ACM, New York (2012). http://doi.acm.org/10.1145/2335484.2335528
16. Haux, R.: Health information systems past, present, future. Int. J. Med. Inf. **75**(3–4), 268–281 (2006)
17. Huang, H.K.: PACS and imaging informatics. Wiley-Liss, Hoboken (2004)
18. Masciari, E., Mazzeo, G.M., Zaniolo, C.: Analysing microarray expression data through effective clustering. Inf. Sci. **262**, 32–45 (2014). http://dx.doi.org/10.1016/j.ins.2013.12.003
19. Nixon, L.J.B., Cerizza, D., Valle, E.D., Simperl, E., Krummenacher, R.: Enabling collaborative ehealth through triplespace computing. In: Proceedings of the 16th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 80–85. WETICE 2007, IEEE Computer Society, Washington, DC (2007). http://dx.doi.org/10.1109/WETICE.2007.140
20. Nordbotten, N.A.: Xml and web services security standards. IEEE Commun. Surv. Tutorials **11**(3), 4–21 (2009). http://dx.doi.org/10.1109/SURV.2009.090302
21. Schloeffel, P., Beale, T., Hayworth, G., Heard, S., Leslie, H.: The relationship between cen 13606, hl7, and openehr (2006)
22. Sittig, D., Kuperman, G., Teich, J.: Www-based interfaces to clinical information systems: the state of the art. In: Proceedings of the AMIA Annual Fall Symposium, pp. 694–698. CRC Press, London (1996)
23. White, T.: Hadoop: The Definitive Guide, 1st edn. O'Reilly Media Inc, Sebastopol (2009)
24. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, pp. 2–2. USENIX Association (2012)
25. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: Cluster computing with working sets. In: Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing, pp. 10–10. HotCloud'10, USENIX Association, Berkeley, CA, USA (2010)