Applying NoSQL Databases for Operationalizing Clinical Data Mining Models

Marcin Mazurek

Military University of Technology
Warsaw 00-908, Kaliskiego 2,
marcin.mazurek@wat.edu.pl

Abstract. Access to data mining models built in clinical data systems is limited to relatively small groups of researches, while they should be available in real-time to clinicians in order to deliver the results at the point where it is most useful. At the same time, complexity of data processing grows as volume of available data exponentially rises and includes unstructured data. Clinical decision support systems based on relational and multidimensional technology lack capabilities of processing all available data because of its volume and format. On the other hand, NoSQL repositories offer great flexibility and speed in terms of data processing, but requires programming skills. A proposed solution presented in this paper is to combine both of the technologies in a single analytical system. Dual view of the data gathered in the repository allows to use data-mining tools, while Big Data technology delivers necessary data. Key-value style of querying a database enables efficient retrieval of input data for analytical models. Online loading processes guarantee that data is available for analysis immediately after it is produced either by physicians or medical equipment. Finally, this architecture can be successfully moved to the cloud.

Keywords: clinical decision support system, big data, architecture

1 Introduction

Hospitals collects more and more data about patients, both structured and unstructured. It can be than used to make better clinical decision, as doctor can rely not only on their own knowledge, but also take advantage of others experience [11]. The main benefit from clinical decision support systems is that general treatment guidelines can be customized to cases based on evidence. In order to make data available, data warehouses are built. Then, upon the relational structures exposed by the data warehouse predictive models are constructed. This shifts standard medical practice from ad-hoc and subjective decision making to evidence-based treatment. As more data is taken into consideration model has higher accuracy and eventually outperforms physician prediction [9]. What is unique in data mining in medicine is the need for robust operationalization of the models for staff engaged in treatment. Operationalizations is the process of execution of constructed models for operational data. In medicine, operational data are data describing patient. The real breakthrough in the treatment process can be achieved, when during patient examination, doctors would be able to score individual case against some of the published models or just find the most similar cases. With traditional data warehouse technology patients record has to be loaded to data warehouse and then become available for data analysis. As an alternative, doctors can manually input the patient record into the system and then run selected models on the data. None of this approaches is satisfactory. Cyclical Extract Transform Load (ETL) processes cause latency. Inserting data to analytical system by doctors is time-consuming and redundant operation, since the information anyway has to be earlier written to Hospital Information Systems (HIS).

The aim is to make use of unstructured data and facilitate execution of analytical models on data provided by physicians in real time. Before Big Data technology has changed the architectural landscape of decision support systems, such models where either moved to transactional systems as a calculation procedures (e.g credit scoring procedure) or their outcomes where shifted back as additional measures (e.g. churn probability). This causes latency which makes it useless in medicine applications. To boost usage of computer-aided decision making, such models has to be available online and give immediate answers.

To overcome these flaws, already existing data warehouse solutions might be supplemented with NoSQL repositories. They can be loaded online from Hospital Information Systems and directly from laboratory equipment. With their flexible data model it is easier to query database for attributes, which are specified as inputs data in predictive models. Although primary concern in this article is efficient operationalization of predictive data models, they also offer capabilities of efficient processing unstructural data, which is very common in medicine (scans, clinical notes and so on).

The proposed architecture is derived from evaluation of real-world implementation of data warehouse in cardiology clinic. The scope of project covered implementation of data warehouse based on relational and dimensional repository. The whole data warehouse solution has brought value for management processes, but the utilization of system as a clinical decision support system has been very limited. Accompanying data mining tools allow for building rather descriptive then predictive analysis. This is because usage of the model is restricted to historical records which has been already loaded to data warehouse. To be truly useful in clinical decision support, data mining model should be applied to new data. Models should be operationally deployed by moving them to operational systems and mobile devices.Another solution is real-time transfer of operational data to the analytical repository. The first choice, while being a common practise in business, is usually a lengthy process. Based on experiences from use cases of relational data warehouse I believe that the latter option is better, particularly in situation where input data for models are gigabyte images or video sequences.

The rest of the paper is organized as follows. The first section describes architectural principles and describes how different users will use the system. Next section introduces nonrelational repositories. Then the concept of architecture is discussed.

2 Usage Scenario and Design Principles

The target group of the system are varied. Access to advanced analytics will benefit doctors, researchers, patients and system of education. We distinguish different usage scenarios for the system:

1. Knowledge extraction from big volumes of both structured and unstructured data.
   (a) Actors: Data scientists team.
   (b) Input: All available data about patients (including unstructured data like medical images), drugs, treatment procedures and clinical paths. Patients data should undergo process of de-identification.
   (c) Output: Tables with extracted features, algorithms, visualisation of patterns.
   (d) Tools: Statistical packages and data analysis languages like Python[1] or R[2] .
2. Predictive modelling.
   (a) Actors: Medical researchers.
   (b) Input: Relational repository with anonymised health records.
   (c) Output: Various classifiers (e.g. decision trees, regression models), prediction models. clusters, similarity measures, association rules.
   (d) Tools: Data mining tools and statistical packages.
3. Operational: using models to make better decision during patients examination.
   (a) Actors: Physicians.
   (b) Input: Models and data describing particular case.
   (c) Outcome: Prediction.
   (d) Tools: Mobile and desktop custom application.
4. Education.
   (a) Actors: Students and academical staff.
   (b) Input: Models and data describing hypothetical case.
   (c) Outcome: Prediction for hypothetical case, aggregated statistics for selected population.
   (d) Tools: Desktop custom application.

Usage scenarios described above can be used to verify completeness of the vision. The main principles of architecture work are as follows.

[1] http://www.python.org/

[2] http://www.r-project.org/

Single Point of Data Entry All data that is manually inserted to information systems by doctors is inserted once. Currently, doctors fills in required forms to enter data into Hospital Information Systems (HIS). They should not be forced to repeat this step to use decision support models. Once the data is loaded, they should be automatically available in analytical repository.

Online Loading The latency between entering the data about patient case and the moment, when they are available as input for analytical models should not exceed single minutes. This is due to the limited amount of time doctors spend on examining each patient.

Ease of Models Dissemination Models should be easily accessible to all doctor. Using it as a supplementary tool in their practise does should not enforce additional effort. Real-time learning from clinical data and delivery of results to point of meeting patient and doctor may become a methodology for transforming healthcare [12]. This also imposes requirements on end user devices and appplication interfaces.

Capability to Process all Available Data In addition to clinical data stored in EHR, there are others sources of data which may contribute to better understanding of relations between patients and treatment outcome. Examples are pharmaceutical R&D databases and social networks that can be mined for patients activities and preferences.

Providing User with Proper Tools and Interfaces This is particularly important principle, as doctors are not supposed to be specialist in information technology. Each group of users mentioned earlier has different functional requirements. Data scientists with programming background should be provided with more powerful tools than physicians and access data in their native formats.

3 NoSQL Databases

NoSQL databases cover a range of data storage solutions, which differs significantly from each other. However, they have common properties which make them fit the proposed solution:

1. They consume data coming in any of digital forms.
2. Data is loaded real-time. Unlike to ETL processes in data warehouse, cleaning and transformation is deferred. Batch processing is directed by analytical goals.
3. Distributed storage and processing based on Map Reduce paradigm allows for processing huge volumes of data. This architecture is very scalable.
4. Flexible output data model. Key-value databases allows for efficient retrieval of features. Fig 1. shows representation of simplified laboratory procedure outcomes.

```
{
Patient:{
           ID: "20123312344" ,
           "DOB - Date of birth" : new Date ("Oct 23, 1967"),
           "Patient sex" : "Female"
           } ,
"Date of procedure": new Date ("Jun 23, 2005"),
"Complete blood count":
          {
          "RBC - Red blood cell count": 5.20,
          "WBC - White blood cell count": 8.1,
          "Haematocrit - PCV level": 37.5,
          "Erythrocyte volume":90,
          "MCH - Mean corpuscular hemoglobin": 30
         }
}
```

Fig. 1. Procedure outcomes represented in key-value document data model

Idea of applying NoSQL database to medical data storage is widely discussed in bibliography. As a key driver of change, architects point out capability of processing constantly growing volumes of medical data, often unstructured [6]. While this remains unquestionable, NoSQL databases can also fill the gap between medical data mining researchers and practitioners. With means of key-value databases interfaces, preparation of input data for analytical models is much easier than in relational structures. This task can be automated - it is sufficient for doctor to choose case identification and model. The rest is performed by system.

Among examples of open-source solutions that can be used are MongoDB [2] integrated with Hadoop [1]. Big data technology, which NoSQL databases are part of, grows very dynamically so selection of platform should not be done in much advance. Besides comparing the database management system functionality and query language, special consideration is required when moving solution to the cloud [3].

4 System Architecture

The key concept of architecture is enabling dual view of the data: relational and key-value. Data from these repositories partly overlaps: attributes and measures describing history of treatment is available from relational and nonrelational repository. The latter stores also semistructured or unstructured data, from which features are extracted. The structures of the repositories are linked by common definitions stored in metadata repository. The logical components of the system are presented on Fig. 2.

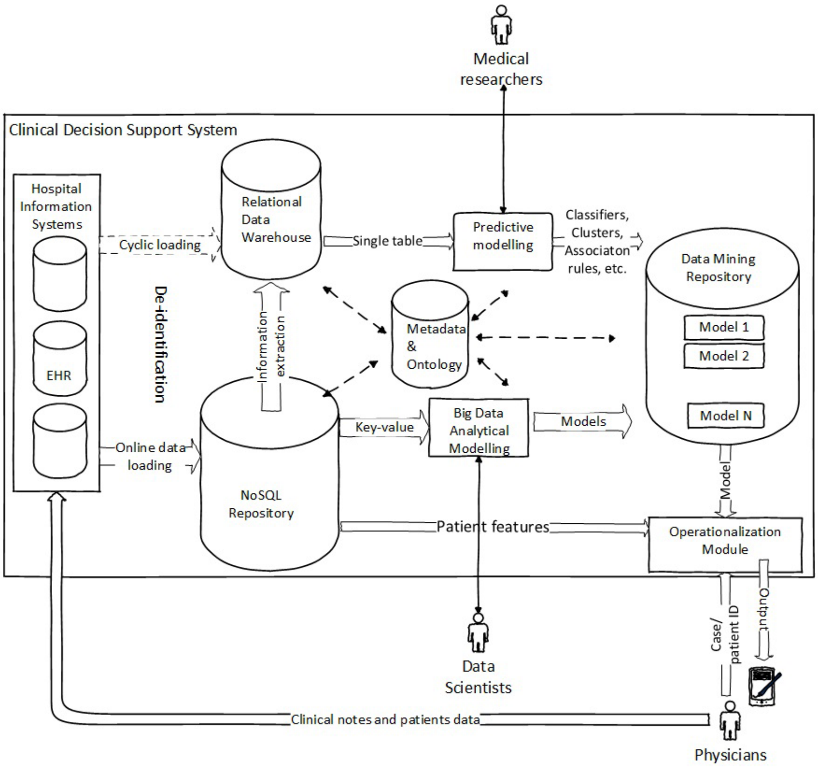

Fig. 2. System architecture

Data Sources Electronic Health Records (EHR) stored in Hospital Information Systems are primary source of data for decision algorithms. The problem is that they cover only a subset of digitalized information about patient. The remaining sources are medical images, laboratory tests outcomes, health monitoring equipment, remote sensors, pharmaceutical databases.Additional potentially valuable data may be gained from social networks. The Big Data technology with NoSQL repositories is trying to give an answer how we can process such volume of unstructured data constantly flowing into the system.

From operational point of view a most important source of data is a physician examining the patient. Data from clinical notes is input in decision process. These notes includes patient's complaint, symptoms, social circumstances, etc. Currently, doctors are obliged to enter some part of it into HIS. Much have been done to set standards of treatment description and patients state. There are international dictionaries like ICD-10. Because this data is used solely for statistical analysis and operational management its quality and scope is limited, and so is the possibility to treat them as complete source of knowledge about patient. This situation may change with implementation of decision support tools, that will be fueled by this data.

NoSQL Repository. NoSQL repository is loaded online with content stored in HIS and all other sources of data mentioned earlier. The loading infrastructure has to guarantee that raw data will become available in the system within minutes from being produced. The next step in the process, which is executed in batch is feature extraction. These features extend set of attributes and metrics available through relational repository. Identification of essential features and algorithms of its extraction are job of data scientist team. This is closed-loop process as available information scope constantly spreads. Data from this repository is indexed with concepts from knowledge base.

NoSQL database systems may also provide user with relational view of the stored data, and thus relational repository described below may be actually a part of Big Data ecosystem. This makes cyclically executed ETL processes redundant in the architecture. However, in many medical centers there is data warehouse already deployed with set of customized application.

Relational Data Warehouse and Data Mining Tools Relational data warehouse allows for immediate answers for ad-hoc queries about either features of individual patients or some population statistics. The repository is loaded with features extracted form unstructured content processed in Big Data part of the system.

Relational data warehouse delivers data to data mining tools used by medical researchers. Data mining tools are used to perform all kind of exploratory analysis and predictive modelling. Various techniques and algorithms are usually delivered to the researchers as black-boxes modules run in graphical user interfaces. The tools support researchers on all stages of the process: data sampling and transformation, modelling, assessment. Ease of use allows to engage wide range of medical experts to build and verify models. They do not have to posses knowledge of data processing techniques and programming, like data scientist teams.

Relational data mining requires that all input data for data mining tasks is in attribute-value form. A training examples have to be stored in single-table, which requires rather extensive preprocessing of relational data model. This includes:

1. Calculation of derived measures (ratios, frequencies of occurrence, flags, levels according to norms).
2. Flattening of structures like time-series (repeated lab-test).
3. Variable transformation: bining, removal of outliers etc.
4. Handling multiple-instance learning problem - patient may have more than one diagnosis.

Some of transformations are used solely for model building, some has to be performed also on input vector. To perform tasks mentioned above, medical researchers should be supported, as transformations like these requires at least familiarity with logical structures of data warehouse. This is a bottleneck of data mining system in medicine. In proposed architectures, those transformation are performed in nonrelational repositories.

Repository of Models The content of analytical repository is made of calculation procedures and algorithms constructed either on relational repository with data mining tools or directly on NoSQL database. Content include but is not limited to:

1. Predictive models, which come in variety of forms: regression equations, neural networks, decision trees, Bayesian networks, complex classifiers.
2. Similarity measures between clinical cases, clusters and their parameters.
3. Association rules.
4. Visualization of patterns.

Each model is accompanied with set of metadata, describing its purpose, set of inputs, accuracy measures. The latter is constantly evaluated in order to withdraw models that performs poorly on new cases. Models may be ranked based on their performance.

Metadata and Ontology. Metadata allows for semantical linking data from relational and non-relational repository. This is accompanied by ontology module which is responsible for providing semantics to the diverse medical information stored in repositories [8]. It includes the ontologies used to define and classify different types of health care data. Ontology concepts serve as keys for non-relational repositories and enables users to communicate with the system with domain terms. Fig 1 is an example of representing laboratory test outcomes with keys which are taken from SNOMED CT ontology [7]. For sake of clarity instead of concept keys the names were presented.

Operationalization Module. This component of the system is responsible for querying database for all relevant attributes required by the predictive models and concerning a particular medical case, which is being evaluated. An examining doctor provide ID of the case. Then a calculation procedure has to be executed with parameter values retrieved from repository. The outcome is presented online.

Implementation of this module will rely on scoring engine executing predictive models. Scoring engine will be decoupled from modelling tools by means of using standardized form of models like Predictive Modelling Markup Language (PMML) [4]. It allows for using any of of data mining tools as long as they provide possibility of export models to PMML. The key point is mapping data dictionary part of PMML model with operational data stored in repository. The assumption is that both data dictionary elements in model and keys in NoSQL database refer to common ontology concepts.

De-identification of patient's data and possibility to query particular case by doctors at the same time can be achieved by reversible coding of id attributes [5]. This form of de-identification goes with mapping tables, which maps artificial keys, which are used in repository with real ones. Nevertheless, security architecture of such system remains a challenge, because medical data often carry an unremovable identification footprint.

5 Conclusions

In this paper I presented a concept of NoSQL repository in architecture of clinical decision support systems. It allows for a shift from monitoring to prediction and simulation [6]. Data warehouses will begin to play substantial role not only as a tool for better health management system, but also will support doctors in their practise. Doctors point out that even visualizing patient's case against population and comparing case to most similar cases would be helpful, especially for less-experienced ones, who begin their carrier.

Key characteristics of the presented architecture is real-time loading of all available data to non-relational repository, from which processed data is published. Data is exposed in relational structures and key-values stores. The first model is best for data-mining tools, while the latter simplifies querying databases for input attributes of predictive models.

This architecture might be easily scaled to cover data from more than single medical institution. The quality of prediction rises with bigger set of evidence. Finally, already built classifiers might be published in 'Analytics-As-a-Service' model in cloud, and it is not technological difficulties which are worst to overcome. Except for data privacy matters, the concept of common repository and data sharing itself seems to be troublesome. Those concerns should be opposed to better outcomes of treatment procedures at possibly lower costs, thanks to rejecting the paths that somewhere proved to be unsuccessful. The portals like

PatientsLikeMe[3] proves that evidence-sharing may be valuable tool in decision making even by non-professionals.

The feasibility of the architecture is still to be proved by set of prototypes. Some of them are already implemented: scoring engine based on Python written Augustus [10] and another one operating in Amazon Cloud based on ADAPA [13]. Future work will concentrate on structures and efficiency of loading data to NoSQL database from operational data sources.

References

1. Hadoop, 14/02/2014, http://hadoop.apache.org
2. Mongo DB, 14/02/2014, http://www.mongodb.org
3. Bajerski, P., Augustyn, D.R., Bach, M., Brzeski, R., Duszeko, A., Aleksandra, W.: Databases vs. cloud computing. Studia Informatica 33(2A), 9–25 (2012)
4. Data Mining Group: PMML 4.1. Specification, 14/02/2014, http://www.dmg.org/v4-1/GeneralStructure.html/
5. Emam, K.E.: Guide to the De-Identification of Personal Health Information. CRC Press (2013)
6. Groves, P., Kayyali, B., Knott, D., Van Kuiken, S.: The 'big data' revolution in healthcare. Tech. rep., McKinsey & Company (Jan 2013)
7. International Health Terminology Standards Development Organisation (IHTDSDO): SNOMED CT, 13/12/2013, http://www.ihtsdo.org/
8. Khan, A., Doucette, J., Jin, C., Fu, L., Cohen, R.: An ontological approach to data mining for emergency medicine. In: 2011 Northeast Decision Sciences Institute Conference Proceedings 40th Annual Meeting. pp. 578–594. Montreal, Quebec, Canada (April 2011)
9. Oberije, C.: Mathematical models out-perform doctors in predicting cancer patientsŕesponses to treatment (Apr 2013), retrieved December 13, 2013 from http://www.sciencedaily.com/releases/2013/04/130420110651.htm
10. Open Data: Augustus. PMML model producer and consumer.Scoring engine., 14/02/2014, https://code.google.com/p/augustus/
11. Savage, N.: Better medicine through machine learning. Commun. ACM 55(1), 17–19 (Jan 2012), http://doi.acm.org/10.1145/2063176.2063182
12. Yu, S., Rao, B.: Introduction to the special section on clinical data mining. SIGKDD Explor. Newsl. 14(1), 1–3 (Dec 2012), http://doi.acm.org/10.1145/2408736.2408738
13. Zementis: ADAPA Scoring Engine, 14/02/2014, http://www.zementis.com/adapa.htm/

[3] www.patientslikeme.com