A Scalable Data Repository for Recording Self-Managed Longitudinal Health Data of Digital Patients

Xia Zhao, Youbing Zhao, Nikolaos Ersotelos, Dina Fan, Enjie Liu, Gordon J. Clapworthy, Feng Dong

Abstract— **This paper presents the proof-of-concept design of the data repository for 4D digital avatars in the MyHealthAvatar project. Taking account of the privacy and legal issues of patient health information, the research generates a set of synthetic data based on publicly available survey data. At the prototype stage, these synthetic data are used in the scenarios of data storage and management. The paper discusses the early proof-of-concept design of the technical stack which enables the storage and query of large scale patients' health data and empowers the future data mining and analysis for health care support. It provides the first stage implementation and the use of it for data analytics.**

I. INTRODUCTION

Personalised health care has become mainstream in modern health care systems. To achieve it, multi-dimensional information regarding the patient should be made available so that a personalised care decision can be made by the medical team. Further, recent technological advances now make it feasible to build a patient-empowered healthcare system. For instance, the rapid development of the Internet enables the general public to access, update, share, and manage personal data far more easily than in the past, and recent advances in Big Data provide the technological foundation for large-scale patient data to be collected, stored and analysed. Meanwhile, the ubiquity of Web 2.0 based social networks and smart mobile devices provides ready opportunities for people to share their feelings and conduct discussions in the digital world.

The general concept of using a digital avatar as a personal health information centralisation service has been proposed previously, and there have been several attempts at implementation including 3D Avatar from IBM [1], HealthVault[1] from Microsoft and Google Health [2]. The patientslikeme[2] platform aims to allow patients to share their experiences with other patients.

The 4D Avatar proposed within the MyHealthAvatar project[3] is intended to be a citizen's lifelong companion a complete digital representation of his/her health-related data, including both medical records and social or lifestyle information, to create a comprehensive picture of the person in the context of their individualised healthcare. All of the data will be useful at different levels in assisting health decisions to be made. For example, doctors will be better equipped with rich information about the patient when deciding upon the type of treatment the patient needs or will be able to make better predictions of the likely effectiveness of a drug prescribed for the patient. It will address the shortcomings of the existing highly fragmented resources in Europe and eliminate the barrier of accessing medical data when citizens migrate from one country to another. Most importantly, the Avatar will provide a unique interface to enable citizens to manage their own health data, to monitor their health status or make their own decisions on their preferred treatment.

Within MyHealthAvatar, collecting, accessing, managing and possibly sharing healthcare related data are not only important to individuals who can manage their own health, but also important for clinicians and other healthcare workers for patient monitoring and providing suitable in-time care. In this paper, we present the design of the infrastructure of the back-end data repository for MyHealthAvatar. The design takes into account the consideration that the data will be large and likely to fall into the category of Big Data, bearing in mind the number of users and the quantity and types of the information that will be stored on daily basis for the duration of a persons entire life. The system also anticipates multiple simultaneous data access. Other issues related to the system design, such as data security and visualisation are outside the scope of this paper.

The research and development presented in this paper is work in progress. The remainder of the paper is organised as follows. Section II provides the key user scenarios that guide the design of the current data model. Section III briefly discusses the method of generating synthetic data. Section IV proposes the proof-of-concept design of the whole data storage and analysis stack, and Section V provides the early implementation and results. Section VI finally draws conclusions and outlines future work.

II. ILLUSTRATIVE USER SCENARIO

In order to guide the design and development the data repository of the 4D avatar, we consider examples within the clinical context of Bipolar Disorder disease which is a mental illness classified by psychiatrists as a mood disorder. It is important to monitor the patient's information, such as personal history, daily diary and physical examinations.

- *Personal history* not only helps clinicians assess the patient's past and current health situations, but also assists them to evaluate the development of the medical problems, produce corresponding treatment plans and

X. Zhao, Y. Zhao, N. Ersotelos, D. Fan, E. Liu, G. J. Clapworthy and F. Dong are with Department Computer Science and Technology, University of Bedfordshire, Luton, U.K. {xia.zhao, youbing.zhao, nikolaos.ersotelos, dina.fan, enjie.liu, gordon.clapworthy, feng.dong}@beds.ac.uk

[1]https://www.healthvault.com/gb/en

[2]http://www.patientslikeme.com

[3]http://www.myhealthavatar.eu

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

	pserial	age	sex	ethnici	eqvinc	topqual3	marstatb	wtval	htval	sysval	diaval	pulval
1	1	3	men	black british	9323.0	item not applicable	item not applicable	22.0	95.5	105.4	66.9	66.9
2	2	51	men	white	111642.0	no qualification	married	96.1	169.7	138.4	96.2	96.2
3	3	7	men	white	14962.9	item not applicable	item not applicable	49.6	139.2	116.6	67.5	67.5
4	4	59	men	white	15287.3	nvq4/nvq5/degree or equiv	married	62.1	183.3	125.1	67.5	67.5
5	5	73	men	white	6930.8	nvq3/gce a level equiv	married	104.9	181.1	143.5	74.9	74.9
6	6	3	women	black british	20765.9	item not applicable	item not applicable	12.5	105.6	120.0	59.3	59.3

	glwear	cigst2	drating	porfv	adtot30c	wlktot
1	item not applicable	item not applicable	NA	6.7	item not applicable	242.4
2	yes, heavy smokers, 20 or more a day	non-smoker	9.2	3.9	item not applicable	NA
3	no	item not applicable	NA	2.8	item not applicable	242.0
4	yes	non-smoker	7.0	2.1	item not applicable	NA
5	yes	non-smoker	8.3	0.8	item not applicable	NA
6	no	item not applicable	NA	0.5	item not applicable	241.3

Fig. 1. Samples of synthetic user data.

monitor the patient's health over time. Personal history information may include their age, gender, ethnicity, history of current conditions/symptoms, history of treatments and side effects, and so on.

- *Daily diary* allows clinicians to monitor closely the patient's mood changes and helps the patients themselves to track and alter their activities, if necessary. The information recorded includes physical, social, mental function levels, daily activities, any infections, etc.
- *Physical examinations* enable clinicians to assess the overall physical condition of the patient, covering height, weight, temperature, pulse, blood pressure as well as specific investigations thought useful in the patients context.

In addition to storing the above information, it is important for patients to be able to query their health record and, possibly, to share/compare their information with others; for clinicians, it is helpful to be able to perform real-time analysis in order to find early indicators of the likely onset of adverse aspects of the condition and, thereby, provide appropriate early interventions.

III. Synthetic Data Generation

As a proof-of-concept attempt for the digital representation of patient health status, we have built the infrastructure for patients to join in the community and build their own Avatar. However, due to ethical concerns, it is often difficult to obtain large amounts of health record data from clinics and hospitals even when the records are properly anonymized, as the patient can often be identified from the aggregated information in the health record. In order to explore the benefit and functionalities that the Avatar will bring to patients as well as clinicians, we collected patients statistical data and normalised patient data from the literature, and then synthesized these data into demo patient avatars.

For population synthesis, given a small sample of households and some global population statistics, a large number of cases can be generated that satisfy the global statistics. Data synthesis also has a large impact on health-related research. The data source for data synthesis was the Health Survey for England (HSE)[4] which ranges from the early 1990s to 2011. The most recent surveys have 10,000 records over 2,000 variables. In particular, we are using the Health Survey for England 2002 data as it comes with a teaching data set in which there is a reduced variable number of 60. The data are freely available from the UK Data Service.

[4]https://discover.ukdataservice.ac.uk/series/?sn=2000021

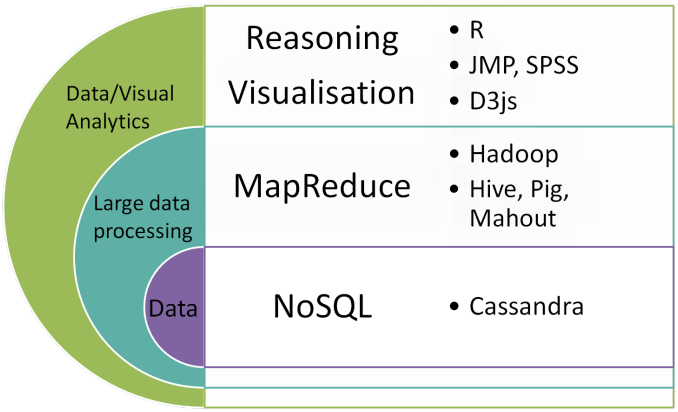

Fig. 2. The overall technical stack for real-time digital patient data analytics.

Our current method converts all continuous variables into discrete variables and generates joint probability tables followed by conditioning based on the dependent variables, for example, the systolic blood pressure is dependent upon "age", "sex" and "wtval" (weight). Apparent dependencies will be preserved but a disadvantage is that hidden dependencies will not. However, the number of dependencies is usually very limited as the dependencies increase, so the computation becomes slower due to multi-dimensional joint probability table generation and conditioning. To reduce computation time, some of the variables are predicted with linear regression. For example the diastolic blood pressure is predicted based on linear regression of the diastolic-systolic dependence.

The program is written in R [3]. Figure 1 shows a number of sample records with 18 fields extracted from 1 million synthetic records generated with R. There are also some existing problems to be improved during data synthesis. The R code currently uses a for loop to create data for each record, which is slow for the generation of a large number of records. In addition, as the number of data records in the data source is limited, there is an insufficient number of records for certain data types to make realistic data prediction.

IV. Proof-of-concept Design

In MyHealthAvatar, historical health data are imported and real-time data collections are updated. The research employs cloud computing and NoSQL approaches in order to enable digital patients' data to scale well in large volume and to allow real-time analysis of the collected data. The results of the analytics may be processed and returned to the data store. The primary technical stack has three layers: *scalable data store*, *large distributed file systems and data warehouses*, and *statistical and visual data analytics*.

A. Scalable NoSQL Data Store

The lifelong patients' data to be stored is complex, with hundreds of attributes per patient record that will continually evolve as new types of calculations and analysis/assessment results are added to the record over time. As shown in Figure

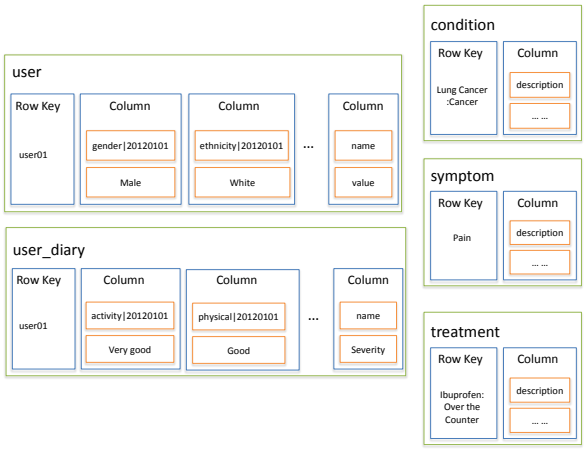

Fig. 3. Cassandra column families implemented in MyHealthAvatar

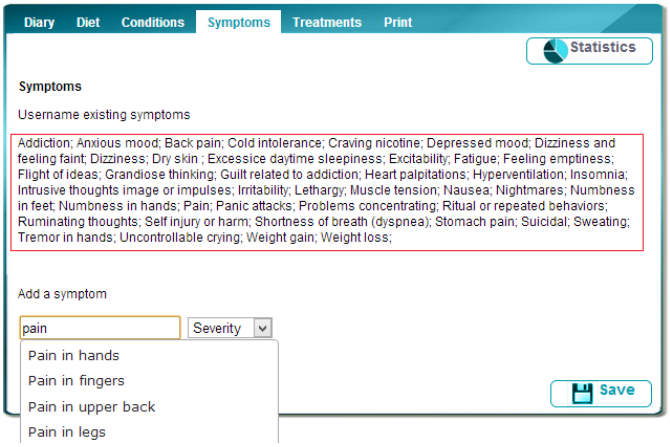

Fig. 4. Existing symptoms and adding a new symptom.

2, the current design uses Apache Cassandra[5] to store these large-scale data. Cassandra is an open-source, peer-to-peer and key value based store, where data are stored in key spaces. One key space can match to a list of column families, each of which is mapped to a list of columns, while the column is mapped to a list of timestamps, then the timestamp is mapped to the value. The unique row key identifies one record; it is mapped to a list of column families. Once a data store is set up, the key space is typically fixed, but the column families and the columns can be added during data loading.

Cassandra supports row-key based query, all-row query and range query on both row and column. Column name can be of composite type, which supports slice based range query on columns. For example, the name of a column to store patients' symptoms can be saved in the form "username|date", which will enable a column range query with a slice of either patient's user name or the record time.

B. Distributed Data Processing and Warehousing System

Since MyHealthAvatar requires large-scale statistical data mining in order to discover patterns within the data concerning medical conditions, treatments, etc., the technical design adopts Apache Hadoop[6] to perform distributed data processing using MapReduce [4]. Both MapReduce and NoSQL are evolving to meet the demand of managing large amounts of Internet-based data. The NoSQL data model emerged for storing and querying large volumes of data on clusters, while the MapReduce programming model performs well for processing these large data sets on large commodity clusters.

Cassandra has built-in support for the Hadoop implementation of MapReduce [5], together with high-level scalable data warehousing such as Apache Hive and Pig on top of Hadoop. The typical combination with Hive is frequently considered for real-time data analysis.

[5]http://cassandra.apache.org/
[6]http://hadoop.apache.org

C. Statistical and Visual Data Analytics

In MyHealthAvatar, data are generated, accumulated and updated continuously. One objective of MyHealthAvatar is to perform data analysis on the overall patients information collected in order to extract clinically meaningful data from the heterogeneous data of individual/shared avatars, such as the patterns of symptoms, experience of treatments, medicines, and risk factors, etc. Visual analytics is the field of making data analysis via interactive user interfaces.

MyHealthAvatar uses a web interface, accordingly web-based visual analytics is needed. Unlike standalone software that has to be installed on each computer, web-based programs are downloaded and interpreted by the web browser and the most common language used is Javascript. At the data analytic layer, two types of analysis will be conducted: one is off-line data analysis using tools such as R [3] and SPSS [6], to provide in-depth analysis on available avatar data; the other is online visual rendering using libraries such as D3.js[7], which is a very popular and powerful JavaScript library for online data visualisation and analysis and combines visualisation and data-driven analysis.

V. Implementation and Early Results

This section discusses the implementations of the back-end data repository based on Apache Cassandra and the initial analysis results shown in the client side prototype.

A. Back-end data store

The data store and query with Cassandra are implemented in Java with the Astyanax[8] client library. A list of column families is created; these store user account, personal information, all collected medical conditions, symptoms, treatments as well as their information for each user. Figure 3 shows some examples of these column families.

For user-related data, the username is used as the row key, all timeline-based data are stored in composite column type with time data in the column name. For example, the

[7]http://d3js.org/
[8]https://github.com/Netflix/astyanax

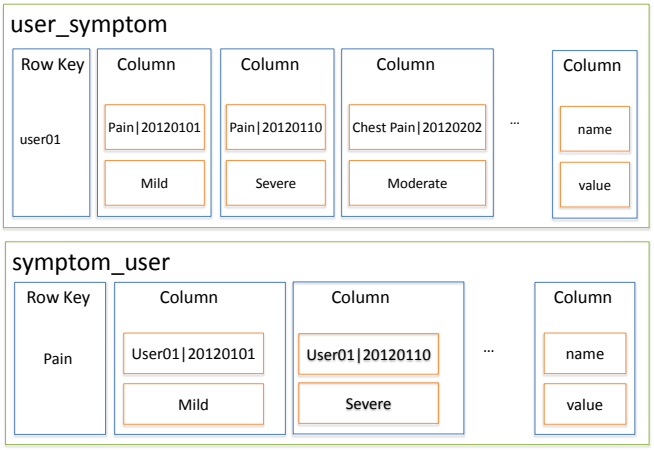

Fig. 5. Column families for storing symptoms of users and vice versa.

patient's activity data stored in the *user_diary* column family is identified by the username "user01", and the column name is "activity|20120101". In this way, a Cassandra column slice query can be used to retrieve the data by diary type and list them in chronological order.

General medical condition, symptom and treatment information including the name, category and short descriptions are collected and imported into the data store. This is used on the client side in the form of a drop box in order to guide patients to add new medical data into their records. For example, as shown in Figure 4, a drop-down list of the *pain* symptom is provided from the data store while the user is trying to add a new symptom.

In order for queries to be fast, user-related medical data are stored in two types of column family: one type is identified by the username row key, the other is identified by the medical information. For example, as illustrated in Figure 5, the user's medical symptoms are stored in two column families: "user_symptom" and "symptom_user". The implementation of the "user_symptom" column family enables fast retrieval of the complete history of the symptoms over the entire period of the record or within a period of a specified time, and the "symptom_user" enables us to acquire efficiently the distribution of the users for a given symptom.

B. Client-side Data Management and Visualisation

On the client side, a series of management tools are being developed to enable synthetic data generated earlier from the health survey to be imported into the data store as the base data for testing of the prototype. These tools include automatic generation of Web user accounts, importing synthetic user information and diary, importing lists of medical conditions, symptoms and treatments, etc.

Further, a prototype Web application has been developed to allow users to login as a synthetic user to view and update the record for the user. Figure 4 shows a snapshot of the page in which users can record their medical symptoms, and

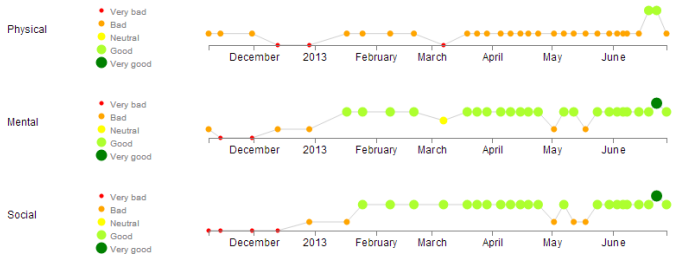

Fig. 6. Examples of charts showing a patient's diary in MyHealthAvatar

Figure 6 depicts a list of charts that show the time line data of the patient's diary.

VI. Conclusion and Future Work

This paper has presented the proof-of-concept design of a data repository for managing digital patients data and supporting large scale analysis on these data. The prototype described in this paper has shown a promising vision of making a 4D health avatar available online to both patients and clinicians.

A three-tier technical stack is proposed, which facilitates the large volume data store, distributed data processing and in-depth data analysis. The implementation of the technical stack is still at the work-in-progress stage, and our future work includes:

- integrating Apache Cassandra and Hadoop and performing both data storage and processing on a cloud platform;
- generating more synthetic data related to the patients medical records and importing them into the data repository;
- providing a data validation mechanism for use when storing the synthetic data; and
- performing data analysis on various scenarios created within the MyHealthAvatar project.

VII. Acknowledgments

This work is supported in part by the European Commission under Grant FP7-ICT-9-5.2-VPH-600929 within the MyHealthAvatar project.

References

[1] IBM, "IBM research unveils 3D avatar to help doctors visualize patient records and improve care." Retrieved June 2013 from http://www-03.ibm.com/press/us/en/pressrelease/22375.wss.

[2] Google, "An update on Google Health and Google Powermeter." Retrieved July 2013 from http://googleblog.blogspot.co.uk/2011/06/update-on-google-health-and-google.html.

[3] S. Stowell, *Instant R: An Introduction to R for Statistical Analysis*. Jotunheim Publishing, 2012.

[4] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[5] E. Hewitt, *Cassandra: The definitive guide*. O'Reilly, 2010.

[6] M. Norusis, *SPSS 16.0 guide to data analysis*. Prentice Hall Press, 2008.