Modeling Stack Framework for Accessing Electronic Health Records with Big Data Needs

Jyotsna Talreja Wassan
Asstt. Professor
Maitreyi College
University of Delhi, India

ABSTRACT
In the recent years information technology has brought remarkable changes in various areas including health care services. The use of Electronic Health Records (EHRs) for maintaining and analyzing health care data online has set the clinical environment on the accelerating path of changes. These digitized health care records are form of Big Data, as they are voluminous, dynamic and heterogeneous. It is desirable to extract relevant information from EHRs and offer healthcare recommendations to novice users or various stakeholders of the clinical environment. In this paper, a sample modelling of general framework, based on extracting patterns and generating recommendations is reviewed for maintaining and accessing EHRs, which are form of Big Data. The main focus of the paper is to propose how data storage with Big Data stores and analytics with MapReduce paradigm, may be performed on simulated health data.

General Terms
Electronic Health Records (EHRs), Data Mining, Recommender, Stack, SQL

Keywords
Big Data, MongoDB, MapReduce, Sharding, HLQL

1. INTRODUCTION
Health care data are valuable resource that is useful for managing and planning clinical environment. The clinical environment deals with patient's profiling demographics (age, sex etc... information), treatment provided by a clinician, medical history of a patient, laboratory or radiology reports, billing or insurance claim data, etc. Electronic Health Records (i.e. EHR) support patient centric approach for storing and retrieval of online health data that is accessible across various hospitals. But the electronic storage, management and retrieval of health care data analytically are difficult tasks as the health data are complex, voluminous, distributed, dynamic, unstructured and heterogeneous [27].The collaborative large-scale management of health data is important for developing technology enabled health care systems. The aim of this paper is to review a trial on modeling of general stack framework that may be used to modularize, simplify and expedite the large-scale processing of electronic health data for various stakeholders.

2. ELECTRONIC HEALTH RECORDS
Health care is a data-intensive domain [6]. The practice of electronic health care generates, exchanges and stores large amounts of patient-specific information including diagnosis, medication, laboratory test results, radiological imaging data, insurance data etc. This data are to be distributed across various networked hospitals to provide more flexibility to the patient.

2.1 Variety of Health Data
The variety of health care data may be spatially, temporarily, physically, functionally or conceptually relate well to EHR systems. The medical care data could be classified as follows:

- Text Data: unstructured clinical notes by a doctor, pathology laboratory reports, etc.
- Imaging Data: various concepts related to medical imaging and complementary view into the structural or functional concepts of patient's organs. For example, ultrasound, computed tomography,, microscopy etc.[2, 10]
- Biological Data: data related to genetics i.e. genomic configurations like DNA sequence etc.
- Billing Data: Billing texts and details with discharge summary.
- Others: imprinted signals such as electro-cardiogram (ECG Graphs), electroencephalogram etc.

EHR is basically a longitudinal collection of electronic health care data. Users may access, enter, upload or update information (text or graphics) in EHR systems. The information deposits of medical domain are increasing day by day with increase in population and improved medical facilities for all, across the world. This variety and increasing masses of health data are stimulating demand for Big Data techniques in the landscape of online clinical environment.

2.2 Various Stakeholders of EHRs
There are different categories of stakeholders of the medical information with different needs and requirements. These are clinicians/physicians or doctors, medical students or researchers, patients, pharmacologists and medical insurance payers [6].

2.2.1 Physicians/Clinicians/Doctors
They are responsible for building treatment plan for a patient by providing medication or recommending various diagnostic procedures, if required. In stepwise treatment plan, physician may embrace different information needs at different steps to better conduct the diagnosis process. Their main aim is to take steps for improving patient's health. The pertinent approach they follow is to analyze the change in the symptoms over the period of time with the progress of disease in patients. If clinicians are provided with data in hand online, procedures will be faster. Various clinicians across the world are connected online for interactions and discussions to produce better results. They may be benefited from recommendations for prescribing medicinal drugs or treatment plans to cater to the needs of patients.

2.2.2 Medical students or researchers
Researchers and students doing medical training many times look upon to the sample medical cases as handled by their seniors or specialists of medical domain. They can interact

online and refer to sample databases of curing plans. They may also consult online articles in medical journals or textbooks to learn more about that information [6]. Medical practitioners are needed to be educated about the treatment plans and the analysis of results of diagnostic procedures conducted to cure the disease. It is beneficial for medical students to get recommendations about the information of interest to them that may allow them to do better medical practice.

2.2.3 Patients

A patient as a stakeholder can be described with parameters as indicated in Figure1.In today's world of technology, patients many times consult online available information sources to better understand their health condition and to gain knowledge about treatment measures for the diagnosed disease. They could be benefited from getting access to their own medical records available online so that they can better understand their medical conditions and they could be recommended best suitable medicines for diagnosing the disease symptoms being searched by patients online.

2.2.4 Pharmacologists or Diagnostic Laboratory Attendants

They are responsible for providing medicines or conducting diagnostic tests. Physicians may directly send the details of patients and requirements online to the pharmacologist or lab attendants making it more convenient process for a patient.

2.2.5 Medical Insurance Payers

Payers, such as insurance companies are responsible for managing costs during patient's medication and treatment procedures and they need to verify the same for preventing misuse. They may aim to minimize costs, control risks, and provide the quick services to the patient.

3. BIG DATA PARADIGM FOR EHRs

The ubiquitous Big Data has brought revolution in the age of Internet and has attracted big attention in last few years. Big Data are heterogeneous and large data sets; difficult to be handled by traditional database systems. It has become important to choose a right platform for capturing, organizing, searching and analyzing the context of voluminous data. The volume of data is also augmented by variety and velocity i.e. time related variations in data. The wide variety of areas like online businesses, social networking, demographics, geographic information systems, online education, physics (e.g. astronomy), biology (e.g. genomics), chemistry, text linguistics etc. are gaining insight from Big Data paradigm. The electronic health management is a progressing area. Various recent platforms supporting Big Data management mainly focus on data storage, management, processing, and distribution and on data analytics.

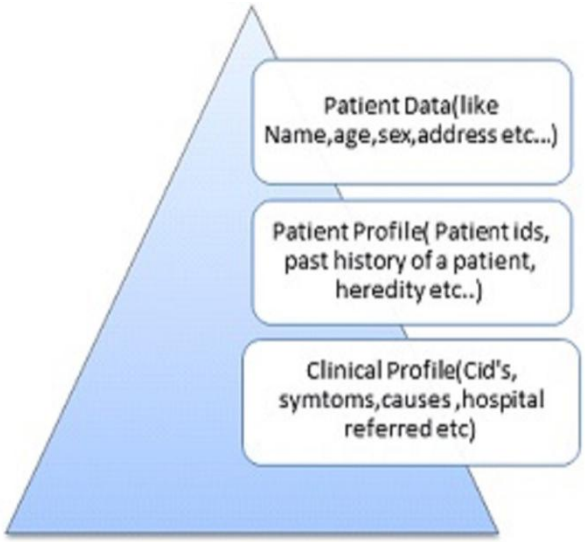

Fig 1: Patient's Description

Various NoSQL data stores such as Cassandra, MongoDB and Hadoop HBASE etc. are in use today to acquire, manage, store and query Big Data. NoSQL databases are schema-less and flexible [7, 8, 15]. Various hospitals dealing with online health records deliver near real time data. Streams are the manifestations of the same. Mostly clinically relevant medical data is unstructured and heterogonous in terms of variety like visuals from imaging labs, EMRs, graphs, textual clinical notes, diagrams, X-rays, medical correspondence from insurance claims etc. Thus three V's: Volume; Variety and Velocity as depicted in Figure 2, have impacted the overall horizon of Big Data in healthcare industry. Veracity and Value as two new dimensions have been added in today's Big Data world [32]. Big Data is useful if it could be turned into value. The volumes often lead to lack of accuracy, trustworthiness and quality. Thus, it is important to add the feature of veracity to Big Data exploration. Both are relevant for usage of Big Data in healthcare industry.

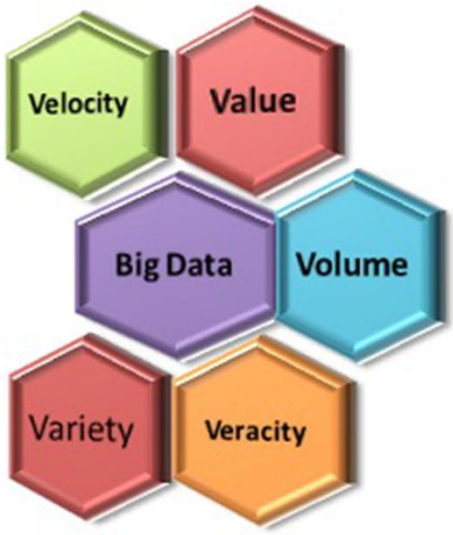

Fig 2: V's Related to Big Data

Advanced data analytics is crucial for improving health care outcomes, incentivizing health management models with efficiency. It is desirable to develop a model based on big data paradigms, to infer valuable aspects, from health care data with some combination of variables already existing in the data. Various healthcare organizations are leveraging big data technology to harness the potential for supporting medical paradigms to build consistent and easily accessible health care systems and to collaborate for improving health care. **Premier** analyzes data from more than 86,000 healthcare providers and enabling its members with trusted integrated information in hand [30]. Researchers at **SUNY Buffalo** are using big data analytics for dealing with multiple sclerosis patients [30]. **University of Ontario Institute of Technology (UOIT)** is using IBM big data technology to capture and analyze real-time data from medical monitors, alerting stakeholders to potential health problems for patients [7, 30]. Various repositories such as **Google Health, Microsoft Health-Vault, and Dossia** etc. have emerged to support health data analytics. **TransCelerate Biopharma** collaboration by pharmaceutical companies globally is accelerating drug development. Various technology driven applications based on big data paradigm are emerging online [7]. **Asthma-polis** developed for asthmatic patients for monitoring via GPS enabled tracker [7, 30]. **mHealthCoach** focusses on providing treatment plans to patients or identifying higher-risk patients based on clinical trials. The **dashboard** technology is also gaining importance for exchange of online health management scenarios. **RiseHealth** is one such example of dashboards [7, 30]. In the clinical sphere, various stakeholders have started to embrace these Big Data platforms supporting storage and analytics of health data. It is required to make health care data available in real time. It is important to efficiently handle large amounts of medical data and leveraging the correlation between data and stakeholders in longitudinal records. The relevant information is to be extracted from complex heterogeneous health data sources. Also with increasing amounts of health data across the hospitals; it is flexible to add computing nodes storing information instead of replacing big powerful servers which is the facility provided by many Big Data platforms including MongoDB[1,29]. The section 3.1 discusses about one such platform "MongoDB" for storing medical health data.

3.1 MongoDB for Storing Health Data

MongoDB (from "hu**mongo**us") is a NoSQL, open source document-oriented database system developed by 10Gen Company. MongoDB stores structured data as JSON-like heterogeneous documents with dynamic schemas. MongoDB has the capability to scale horizontally through sharding. It also has a functionality of querying database [29]. The platform such as MongoDB is suitable for storing EHR data due to its scalability and flexibility in structural format for storage.

3.1.1 Using MongoDB for EHRs

JSON objects can be passed to MongoDB for storage. The patient profiling could be passed as a JSON to Mongo server from a mongo client with just a simple interface command for inserting. MongoDB supports basic CRUD (create, read, update, delete) operations on documents subject to maximum of 16MB size [29] (Figure 3).

A JSON objects are exemplified as follows that stores patient information for a disease registry.

SAMPLE CASE A:

```
Patient_profile_A = {
  "Pid": "12356"
  "Name": "John",
  "Age": 35,
  "Contact": [
   "Address": "105, Park Street, LA venue, Mumbai, India."
   "Email": "w_john@gmail.com"
   "Telephone no": "091-98760569346"  ] }
Clinical_profile_A = {
  "Cid": "CA12356"
  "Name": "John",
   "Symptoms": [
          "Cold", "Cough", "temperature", "headache"
         ]
   "Drugs Recommended": ["Paracetamol","ciplox"]
   "Tests recommended": ["Blood Test]}
```

SAMPLE CASE B:

```
Patient_profile_B = {
  "Pid": "12358"
  "Name": "Rohan",
  "Age": 40,
  "Email_id" : "rohan_kar@yahoo.com"}
Clinical_Profile_B = {
  "Cid": "CB12358"
  "Name": "Rohan"
 "Symptoms": ["Coughing, especially at night", "Shortness of breath", "Chest tightness, pain, or pressure"]
 "Radiological Tests": "Chest X-Ray"}
```

To create the database for storing the above profiles, following commands on the Mongo Client are to be issued.

db.healthrecord.save (Patient_Profile_A);

db.healthrecord.save (Clinical_Profile_A);

db.healthrecord.save (Patient_Profile_B);

db.healthrecord.save (Clinical_Profile_B);

The above sample cases reflect that MongoDB has flexibility for storing documents and thus have dynamic schemas. Documents are not needed to have the same number of fields and the same basic structure. This helps in aggregating and storing hospital information in dynamic form enhancing portability and accountability. The CRUD operations on EHR data stored in Mongo database can be performed easily as showcased in Table1. They are comparable to traditional SQL formats. MongoDB is agile in case of schema design.

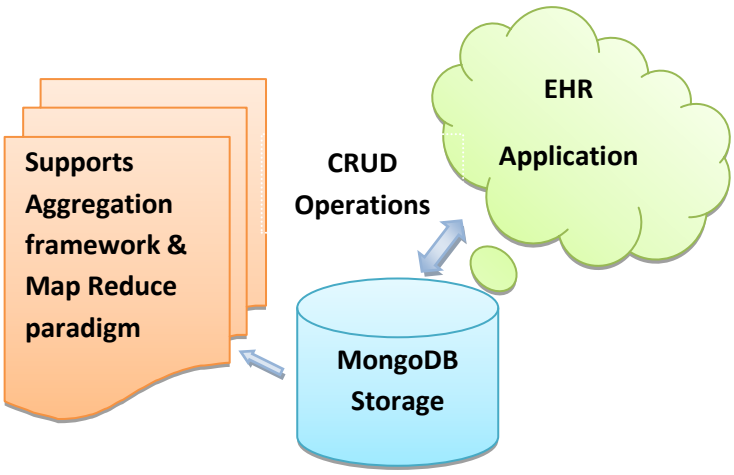

Fig 3: MongoDB for EHRs

Also the Mongo supports following features which makes it a good environment for Big Data sets such as health data [29]:

i. Data are stored in the form of JSON style documents and uses simplified Java Script engine.
ii. It has GridFS for storing data.
iii. It has built-in aggregation framework for simple operations and MapReduce paradigm functionality which can be integrated with platforms like Hadoop for complex data aggregation using Map/Reduce.
iv. It is Cloud-friendly.
v. No complex joins, as in traditional database systems.
vi. MongoDB supports dynamic queries on documents using a document-based query language.
vii. MongoDB is easy to scale.

Table1. Sample CRUD operations on patient data

Traditional SQL's in Relational World	**MongoDB CRUD Queries in NoSQL world**

CREATE TABLE PATIENT (Name String, age Number, Contactid Number)
CREATE TABLE CONTACTS(Id Number, URL String);
INSERT INTO PATIENT VALUES("John",35,105);
INSERT INTO CONTACTS VALUES(105,"j_uod@gmail.com")
Schema Design for RDBMS
patient= {
id: 100,
Name: 'John'
age: 35
Contactid: 105
}
Contacts = {Id: 105,
URL: 'j_uod@gmail.com'}
	db.createCollection(" patient")
	
	Schema Design for MongoDB
	patient = {
	Name: 'John',
	age: 35,
	Contact: [
	Id:105
	URL: 'j_uod@gmail.com'
	]
	}
	
	db.patient_entry.save(patient)
	
	* MongoDB supports Embedded Objects unlike RDBMS and hence there is no need of join queries.

INSERT INTO PATIENT VALUES("Joy",38,106); INSERT INTO CONTACTS VALUES(106,"j_kumu@gmail.com")	db.patient.insert({'Name':"Joy", 'age':38,{Contact: ['Id':106, 'URL':'j_kumu@gmail.com'}]);
SELECT * FROM patient	db.patient.find()
SELECT * FROM patient WHERE age>33 AND age<=40	db.patient.find({'age':{$gt:33,$lte:40}})
SELECT * FROM patient ORDER BY Name DESC	db.patient.find().sort({Name:-1})
SELECT COUNT(*) FROM patient	db.patient.count()
UPDATE patient SET age=40 WHERE Name='John'	db.patient.update({Name :'John'}, {$set:{age:40}}, false, true)
DELETE FROM patient WHERE Name ="abc"	db.patient.remove({Name :'abc'});

3.1.2 *Distributed Access to EHRs*

The physical architecture of MongoDB supports processing of huge amounts of health care data which expects information regarding various stakeholders of clinical environment like patients, clinicians etc. as listed in section 1.2 .MongoDB scales horizontally through sharding. Sharding uses range-based partitioning to distribute *documents* based on a specific index known as *shard key*. Sharding automatically balances data and load across machines. The concept of sharding in MongoDB is illustrated in Figure4. mongos acts as a router and directs the queries from user application to the actual data storage elements known as mongod [29]. In EHR systems, health data could make use of sharding concept and be divided amongst various computer nodes/machines according to say practice location of a physician or hospital and the metadata about clinical profiling could be stored in config servers making sharding an effective phenomenon. This can support both local and centralized access where hospital wise responsibility is delegated among multiple peripheral servers (as reflected in Figure 4, 5). mongos acts as a router and directs the queries from user application to the actual data storage elements known as mongod [29]. In EHR systems, health data could make use of sharding concept and be divided amongst various computer nodes/machines according to say practice location of a physician or hospital and the metadata about clinical profiling could be stored in config servers making sharding an effective phenomenon.

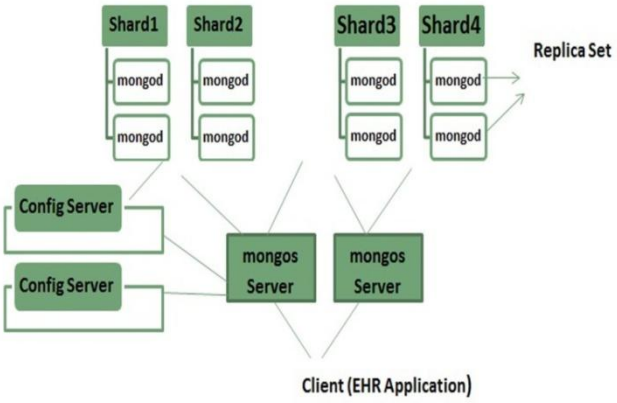

Fig 4: Sharded Model for MongoDB

This can support both local and centralized access where hospital wise responsibility is delegated among multiple peripheral servers (as reflected in Figure 4, 5). The concept of distributed EHR can be realized with sharded cluster as instead of putting up the whole data in one big server, it is being distributed across several computing server nodes known as shards. The replication instances within shard enable protection from automatic failover and also the fault tolerance is enhanced. When a patient is referred to another hospital, his records are distributed or uploaded to the respective hospital server from the main server. This topology favors efficient data storage useful for clinicians across the hospitals without the concern of storing and migrating huge amount of information [17].

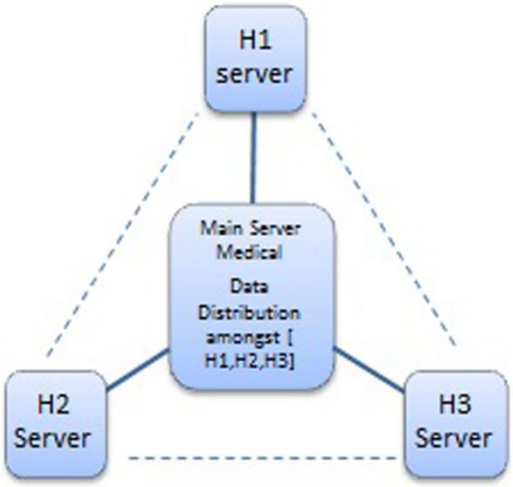

Fig 5: Sharding Data Hospital Wise (H1, H2 and H3 are hospital indicators)

4. MAP REDUCE BASICS

Map Reduce is a powerful programming model that aids in easy development of scalable applications.The MapReduce programming model is describes as follows [4]:

Map Function: Takes an input pair and produces a set of intermediate key/value pair's e.g.

Map: (key_1, $value_1$) → list (key_2, $value_2$)

Reduce Function: This function accepts an intermediate key and a set of values for that key.

Reduce: (key_2, list (key_2, $value_2$)) → $value_3$

The MapReduce supporting library routines try to group all intermediate values associated with the same key. The basic idea behind MapReduce as illustrated in Figure6; is to divide the problem into a set of smaller problems that perform the same operations on a subset of the data in parallel in Map phase and subsequently results from multiple Map functions are synthesized to get intermediate results which are given to Reduce Phase. Parallelism is achieved by running many Map tasks and many Reduce tasks in parallel [4].Health care data being big; may also make use of MapReduce framework in various activities and achieve significant improvement in processing performance by dividing the processing across a cluster of computing nodes.

5. THE PROPOSED STACK FRAMEWORK FOR EHR PROCESSING

Health care data analytics research increasingly involves the construction of prototyping models for clinical centric stakeholders using electronic health records (EHRs). To facilitate this process, framework of pipelined tasks of extracting useful patterns from EHR data via first applying data mining techniques and then generating recommendations for various stakeholders is proposed as depicted in figure 7. Both stages can be implemented with the help of Big Data paradigm. The proposed approach focuses on mining of electronic health records (EHRs) for establishing new patterns in health management and for studying patient, physician and health correlations. Various data mining techniques such as clustering, classification or association rule mining could be used. The generated patterns may be used in generating recommendations for different stakeholders.

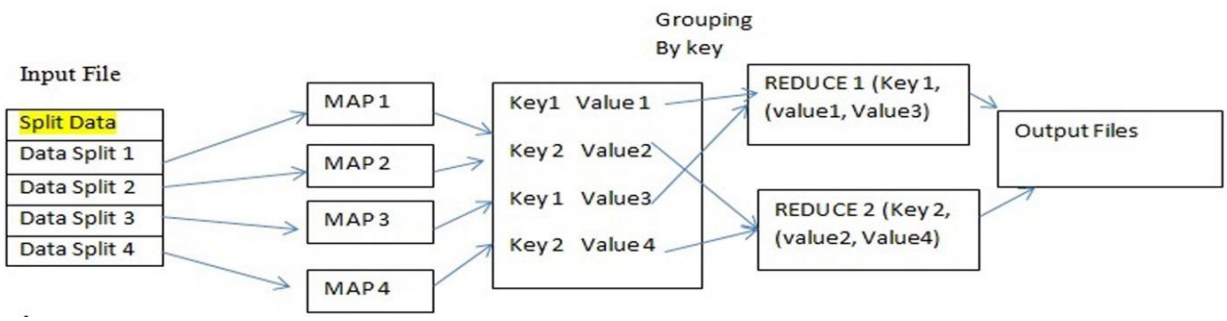

Fig 6: Map Reduce Basics

5.1 Mining EHRs

Data mining techniques are designed to extract meaningful and useful patterns from high-dimensional, heterogeneous and dynamic data sets. EHR datasets are one example of such data sets. The patient's clinical profile serves as the basic entity for EHR dataset which could be represented by a feature vector, having any combination of various data items being stored in EHR systems. Data Mining also deals in finding associations and correlations between elements of feature vector and metadata, such as linkage with patient demographic information. Data mining methods can be characterized as supervised and unsupervised learning [11, 20]. A supervised learning approach deals with a data set of classified labels/classes from which a model is derived to predict future labels/classes from the features. Examples are: classifiers such as naive Bayes; artificial neural networks etc. Unsupervised methods, such as clustering algorithms, take unclassified data set and try to group vectors on the basis of similarity features. Both the variations could prove themselves useful for EHRs in their respective forms [11, 20].Various kinds of mining could be correlated with EHRs; for example *Text Mining* for clinical notes, *Image Mining* for graphical lab reports etc. The mining of EHR databases generate clinical knowledge that could be used for providing treatment plans to individuals and for various other purposes like medical claims. The main motive of EHRs is that "clinician must provide best treatment plan to the patient with clinical information in hand".

Big Data analytics done with Map Reduce phenomenon may support various data mining techniques on large scale distributed environment. For Example K-Means a clustering algorithm has been implemented with MapReduce approach [31].Recommenders (as discussed in section 4.2) may aid clinicians and other stakeholders programmatically in implementing clinical guidelines to process all of the recorded EHR data [22].

5.2 Recommenders

Today in the world of information overload, where there are plenty of offers in almost every domain, it is important to have a system to analyze the information mass and reach to a relevant and fruitful decision in an effective manner. The systems which identify and evaluate items (products and services) of interest for users from vast amount of available information are known as Recommender Systems [22].The two main approaches dealing with generating recommendations are collaborative and content-based filtering. Collaborative Filtering (CF) [16, 19] aggregates view points from various users in the form of explicit or implicit ratings on various items and tends to construct a rating matrix. The recommendations produced are based on the opinions of users similar to the current active user. Content-based approach [3, 19] tries to find out items of interest, which are most similar to the preference of current user in the past. Many advanced recommender systems combine the concepts of collaborative and content-based filtering, taking advantage of both and trying to mitigate the drawbacks of each other. These systems are known as the hybrid systems [3].

EHR systems also deal with large and complex data for finding useful medical patterns and generating recommendations. For example, Comorbidity, i.e. disease co-occurrence in clinical practice (which is based on co-occurrence of patterns of symptoms in patients), could be investigated easily by conceptual recommender engines [11].

It is fruitful to link data mining platforms to the recommender engines which may be used by various stakeholders. Few of the simple recommendations are discussed as follows:

i. The symptoms may be extracted as features from a patient's clinical profile and recorded in a matrix structure. Consider an example in Table 2.

Table2. Sample matrix record for patients with disease symptoms

Pids ⇓ cids ->	Fever	Cold	Cough
001	No	No	Yes
002	Yes	Yes	Yes
003	Yes	Yes	Yes
004	No	No	No
005	Yes	Yes	Yes
006	No	Yes	No
007	No	Yes	No

The table depicts that patients with identifiers 002,003 and 005 are of similar type (forms a cluster on similarity basis); therefore similar kind of medication is provided to all the patients of a cluster (i.e. those having similar kind of symptoms. Similar is the case with patients with identifiers 006 and 007. Assuming a new patient comes with Pid 008 and having symptoms as follows:

Table3. New patient and his symptoms

Pids ⇓ cids ->	Fever	Cold	Cough
008	Yes	Yes	Yes

It is most similar to the cluster of patients having identifiers as 002,003 and 005; therefore similar kind of medication/physician or hospital or tests could be recommended to a new user (patient) as was being provided to patients with identifiers 002,003 and 005.This is an example of collaborative recommendations.

i) If a medical drug is being used and referred by many physicians/clinicians for a disease then whenever a new patient enters the system with the same disease symptoms, recommenders may recommend that drug to the new patient. TOP N recommendations can be generated and may be represented to the patient online for curing a disease.

ii) Considering the past history profile of a patient, match the current disease symptoms of a patient with the past historical symptoms. If a drug proved to be a successful remedy in the past; then the same drug may be recommended for the current cure.

iii) Medical Researchers or students may be recommended with knowledge of treatment plans or drugs that are being followed by the active physician specialists or practitioners for a particular disease.

iv) Recommendations to the insurance payers regarding which kind of coverage care to be provided to a particular patient.

v) Recommendations may also prove to be effective in PHR's i.e. Personal Health Records which aims to provide a complete and accurate summary of an individual's medical history accessible online in real time. The health data on a PHR might include patient-reported outcome data, lab results, and data from devices such as mobile phones etc. [18, 28].An online recommender system for PHR's could be proven to be most effective scenario. The knowledge for recommendations may be gathered from analysis of user interaction with the health management system (user ratings, usage patterns etc...).

Recommenders using Big Data paradigm are emerging and may prove beneficial for above scenarios.

5.3 Query Interface

Big Data platforms provide an inbuilt query engine (like in MongoDB) or support HLQL (higher-level query language) layer to simplify both the specification of the MapReduce operations and the retrieval of the result. Several of these HLQL's like HiveQL, Pig Latin, and JAQL etc. have emerged either as open source projects or commercial products [23].

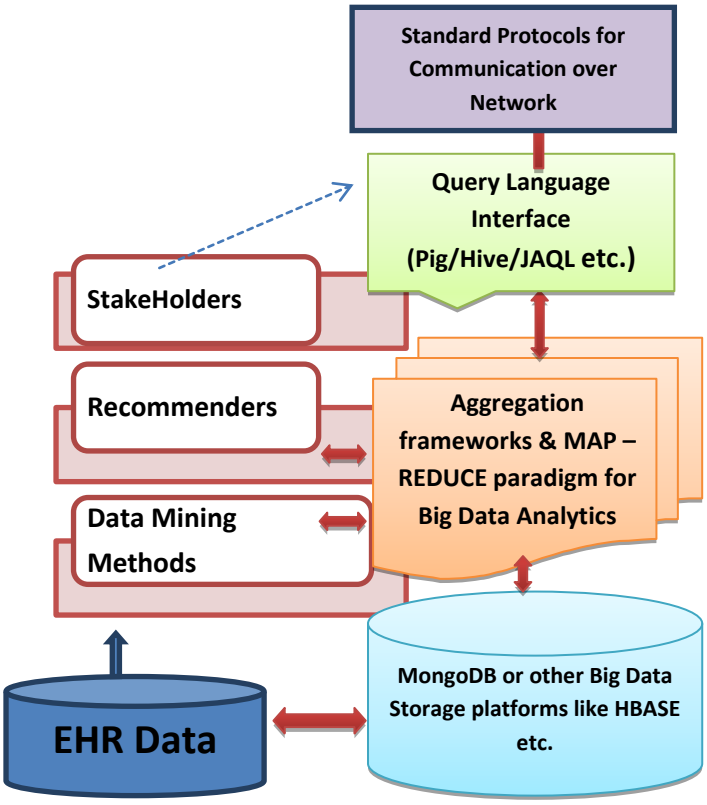

Fig 7: The Proposed Stack Framework

6. A DISCUSSION ON BIG DATA PARADIGM IN EHR WORLD

MapReduce can effectively be used for EHR data analytics and functionality. For example, a pattern could be extracted from large scale EHR data stating that how many times a hospital is being referred by various patients for a particular disease (say heart diseases). Then the hospital being accessed maximum number of times could be recommended to the new patient having similar symptoms for a disease. The count of number of visits to the hospital can be extracted by MapReduce analytics as follows:

map(key, value):

// key: hospital visited; value: patient's clinical profile

for each disease d in value:

for each hospital h in value:

emit(h, 1)

reduce(key, values): // key: hospital visited; value: an iterator over counts, patient's clinical profile

result = 0

for a particular disease in patient's clinical profile

for each count h in values:

result += h

emit(result)

Considering another recommending functionality; the EHR data can also be pre-processed and transformed to the following structure in the form of Matrix (Pi X Hi where P represents a patient and H represents Hospital Accessed).

Table4. The Sample Matrix

	H1	H2	H3	H4	…..	Hn
P1	1	0	0	1		0
P2	1	1	0	1		1
P3	0	0	1	1		0
….						
Pn	1	0	0	1		0

Here weight 1 depicts hospital visited by a patient and 0 depicts not visited. Hospitals having similar preferences by the patients in the EHR data can be grouped together. K means clustering could be used to determine similarly accessed hospitals as follows:

i) Initially choose random k column vectors representing corresponding hospitals as k-means of k clusters.
ii) The data is scanned to assign each column accessed to a cluster which has the closest mean.
iii) Calculate the new mean of each cluster.
iv) Above steps are repeated till the convergence criteria is met.

This type of clustering can be implemented in MapReduce framework as follows:

In the map step:

i) Read the initial cluster centers into memory from an input data file
ii) For each input key/value pair, iterate over each cluster center.
iii) Measure the distances.
iv) Store the nearest center that has the lowest distance to the vector
v) Write the new cluster-center with its vector.

In the reduce step:

i) Iterate over each value vector and calculate the average vector. (Sum each vector and divide each part by the number of vectors received).
ii) This is the new center, save it into a data file.
iii) Check the convergence between the cluster-center that is stored in the key object and the new center. If it they are not equal, increment an update counter.

Run this whole thing until nothing was updated anymore.

Once the clusters are formed, recommendations could be generated for a new user (patient) on the basis of similarity (e.g. cosine similarity) of visiting patterns of a patient with the centers of clusters formed. Also it can be interpreted that the hospital (column), which has maximum number of non-zero entries, is the most popular hospital.

7. CONCLUSION

Big Data paradigm may help greatly in generating new knowledge in health care industry by analyzing unstructured and schema less data efficiently. Also it is useful in sharing information across hospitals and various stakeholders via distributed computing. The layer-wise modular framework for accessing EHRs is suggested in this paper to cope up with the complexities involved in dealing with large scale digitized health data. The paper reflects that mining components and recommendation engines working above them could be useful in effective utilization of EHRs for various stakeholders with the growing number of recorded features in EHRs. The framework gathers information about consumers of health data from EHRs stored in a scalable fashion and tries to calculate similarities between different concepts to produce recommendations. The Big Data platforms available in today's world can effectively be used for storing health data and analyzing it with MapReduce paradigms.

8. REFERENCES

[1] Adrián, Gómez, et al. "MongoDB: An open source alternative for HL7-CDA clinical documents management."
[2] Beutel, Jacob, et al. *Handbook of Medical Imaging, Volume 3. Display and PACS*. Washington, DC: SPIE Press, 2002.
[3] Burke, Robin. "Hybrid recommender systems: Survey and experiments." *User modeling and user-adapted interaction* 12.4 (2002): 331-370.
[4] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.
[5] Duan, Lian, W. Nick Street, and E. Xu. "Healthcare information systems: data mining methods in the creation of a clinical recommender system." *Enterprise Information Systems* 5.2 (2011): 169-181.
[6] Ebadollahi, Shahram, et al. "Concept-based electronic health records: opportunities and challenges." *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006.
[7] Groves, Peter, et al. "The 'big data revolution in healthcare." *McKinsey Quarterly* (2013).
[8] http://nosql-database.org/, Retrieved May 2014.
[9] http://www.mongodb.com/presentations/partner-webinar-electronic-health-records-ehrs-and-mongodb-advancing-data-platform
[10] Jan, J. "Medical Image Processing, Reconstruction and Restoration: Concepts and Methods. 2005."
[11] Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* 13.6 (2012): 395-405.
[12] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."*Journal of Healthcare Information Management—Vol* 19.2 (2011): 65
[13] Lee, Choon-oh, et al. "A framework for personalized Healthcare Service Recommendation." *e-health Networking, Applications and Services, 2008. HealthCom 2008. 10th International Conference on*. IEEE, 2008.
[14] Lee, Ken Ka-Yin, Wai-Choi Tang, and Kup-Sze Choi. "Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage." *Computer methods and programs in biomedicine* 110.1 (2013): 99-109.
[15] Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).
[16] McCrae, John, Anton Piatek, and Adam Langley. "Collaborative Filtering."*http:// www. imperialviolet. org* (2004).
[17] Patra, Debkumar, et al. "Achieving e-health care in a distributed EHR system."*e-Health Networking, Applications and Services, 2009. Healthcom 2009. 11th International Conference on*. IEEE, 2009.
[18] Personal health records Retrieved May 2014, from http://en.wikipedia.org/wiki/Personal_health_record
[19] Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2012.
[20] Ramakrishnan, Naren, David Hanauer, and Benjamin Keller. "Mining electronic health records." *Computer* 43.10 (2010): 77-81.
[21] Romero, Francisco P., et al. "An Ontology-based Recommender System for Health Information Management."
[22] Schafer, J. "The Application of Data-Mining to Recommender Systems." *Encyclopedia of data warehousing and mining* 1 (2009): 44-48.
[23] Stewart, Robert J., Phil W. Trinder, and Hans-Wolfgang Loidl. "Comparing high level mapreduce query languages." *Advanced Parallel Processing Technologies*. Springer Berlin Heidelberg, 2011. 58-72.
[24] Strauch, Christof, Ultra-Large Scale Sites, and Walter Kriha. "NoSQL databases." *URL: http://www. christof-strauch. de/nosqldbs. pdf (дата обращения 07.11. 2012)* (2011).
[25] Talreja, Jyotsna, et al. "Using Web Mining to Generate Recommendations: A Website Recommender." *IICAI*. 2009.
[26] Wang, Fei, et al. "Large-scale multimodal mining for healthcare with mapreduce." *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 2010.
[27] Wasan, Siri Krishan, Vasudha Bhatnagar, and Harleen Kaur. "The impact of data mining techniques on medical diagnostics." *Data Science Journal* 5.19 (2006): 119-124.
[28] Yina, Wan. "Application of EHR in health care." *Multimedia and Information Technology (MMIT), 2010 Second International Conference on*. Vol. 1. IEEE, 2010.
[29] 10gen, Inc: *mongoDB*. Retrieved May 2014, from http://www.mongodb.org.

[30] Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on mapreduce." In *Cloud Computing*, pp. 674-679. Springer Berlin Heidelberg, 2009.

[31] Marr, B. A Talk on Big Data- the 5 Vs Everyone must know (Feb. 2014).

[32] IBM Industry supporting Healthcare Information Retrieved 2014 from http://www01.ibm.com/software/data/bigdata/industry-healthcare.html