Scalable and Secure Electronic Health Record (EHR) System in BigData Environment

S. Shanmuga Priya

Senior Assistant Professor
Department of Computer Science and Engineering
New Horizon College of Engineering
Bangalore, Karnataka, India
priya_soundarajan@yahoo.co.in

Abstract

Cloud computing services offer several flavors of Virtual Machines (VMs). As a result, the centralized approaches are difficult in handling large-scale data sets well on cloud using just one single VM even if the VM has the highest computation and storage capability. Maintaining the linkage structure and updating the statistic information will be huge when the data sets become large. Centralized approach probably suffers from low efficiency and scalability when handling large data sets. All data processed should fit in memory for centralized approaches, but theassumption fails to hold data-intensive on cloud. A large memory should be implemented in order to hold the records and registrations, but the total process is not fit in the centralized system. Also the scalability of the system is not up to the mark. To implement all the features there should be large memory. The centralized Top Down Specialization (TDS) approaches exploits the data structure Taxonomy Indexed Partition Structure (TIPS) to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information in TIPS. The data structure speeds up the specialization process because with the help of indexing structure that helpsin avoiding frequently scanning entire data sets and storing statistical results, thus, thwart re-computation overheads. The amount of metadata that is retained for maintain the statistical information and other linkage related information of a record partitions is comparatively huge set. This consumes considerable amount of memory as like the original data set. The overhead incurredby maintain and updating these linkage structure, and the statistical information maintained will also be huge and becomes a larger data set. Hence, centralized approaches probably suffer from low scalability and efficiency while handling large-scale data sets.

Keywords: big data, scalable, TIPS, TDS, electronic health record

1. INTRODUCTION

Cloud computing is one of the most pre-dominant paradigm in latest traits for computing and storing purposes. It poses a significant impact on the current IT industry as well as on research communities [1] [2] [3]. Data safety and privateness of information is one of the predominant situation within the cloud computing. Data anonymization has been notably studied and broadly followed approach for privacy preserving in information publishing and sharing methods. Data anonymization is displaying up of sensitive information for owner’s information document to mitigate unidentified risk. The privacy data of person may be correctly maintained even as a few mixture records are shared to information person for information evaluation and information mining.

The rest of the paper is organized as follows. Section 2 discusses the related works, section 3 brings out the proposed work, section 4 gives the results and section 5 conclusion.

2. RELATED WORKS

One of the extensively investigated concept in recent days is data privacy preservation [4]. The scalability problems of anonymization are addressed by introducing sampling techniques, decision trees, etc. Iwuchukwu and Naughton [5] proposed R-tree index-based approach by creating a spatial index over data sets by achieving high efficiency. Fung et al. [6], [7], [8] proposed a TDS approach, which aimed at producing anonymous data sets by not taking the data exploration problem [6]. In literature, we could find various distributed algorithms, which are proposed for preserving privacy of multiple data sets retained by multiple parties. Mohammed et al. [9] and Jiang [10] proposed distributed algorithms for anonymizing vertically partitioned data from different sources. This work aimed at preserving the privacy information without disclosing privacy information from one party to another. Mohammed et al. [11] and Jurczyk [12] proposed a distributed algorithmic approach to anonymize horizontally partitioned data set.

3. PROPOSED SYSTEM

Figure 3.1 shows the architecture diagram of the proposed system. The health records of various patients that have been registered are stored in the database. Generally, big data can be processed incrementally and data are divided into equally sized blocks (data chunks). Data chunks created using k-anonymity and generate various anonymization levels. Figure 3.2 shows the data flow diagram of the proposed system. The data is collected from the healthcares and it is partitioned into data blocks. The partitioned data enter into the TPTDS phase with the help of Map-Reduce Top down Specialization (MRTDS) approach. The results of the two phases are the anonymized datasets. Then the datasets enters to the reduce phase to providecounts to the datasets. After that, the datasets are sent to the IGPL analysis and IGPL update phase to calculate the anonymization level. At last the data is shared and accessed in the cloud.

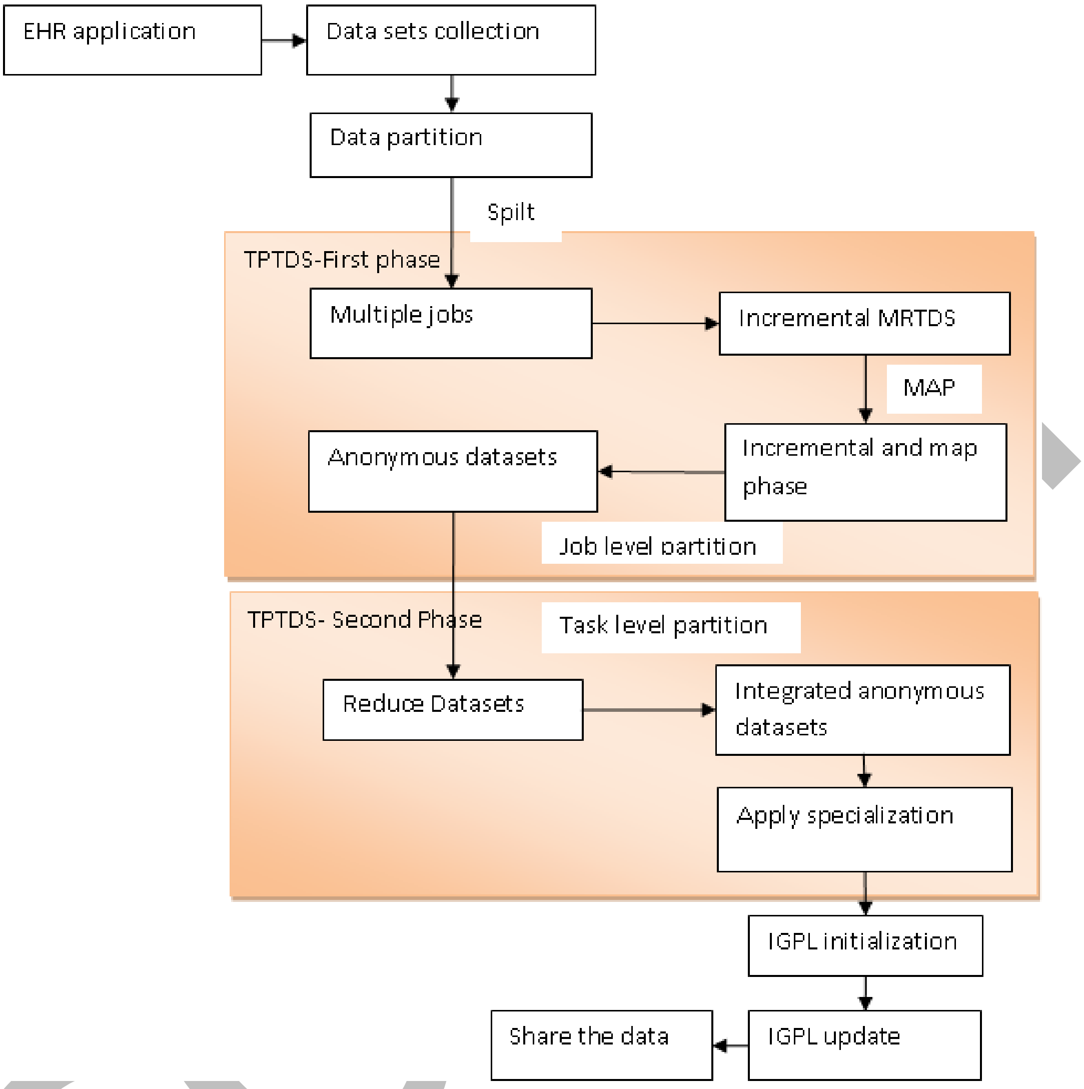


Figure 3.1 Architecture of the proposed system

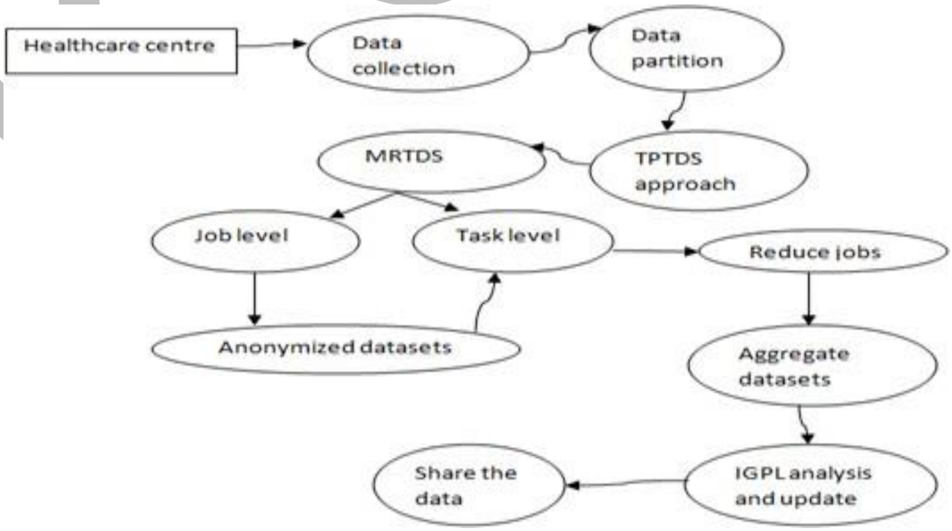


Figure 3.2 Data flow diagram of the proposed system

3.1. Module Description

There are four modules in the proposed system and they are as follows:

a. EHR system development

b. Data chunk creation

c. MRTDS implementation

d. IGPL analysis

a) EHR System Development

EHRs are real-time, patient-centric management system that helps in maintaining recordsof patients that make information available instantly and securely to authorized users. Figure 3.3 shows the homepage design of EHR system. It contains various tabs named home, hospital, patient, research centre, laboratory, specialist, pharmacy. By clicking on those tabs, the details of the particular option can be obtained as well as entered. The design of the homepage contains the Administrator, who has the central control access of all the information in the system.

EHRs may include a range of data, including medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. EHR system is created to gather big data from the users ofthe system such as patients, doctors and health care's. The system is designed in such a way for representing data and that accurately captures the state of the patient at all times. It allows for an entire patient history to be viewed without the need to track down the patient's previous medical record volume and assists in ensuring data is accurate, appropriate and legible. It cut down the timeinvolved in data replication as there is only one modifiable file, which means the file is constantly up to date, when viewed at a later date reduces the issue of paperwork. Due to all the information being maintained in a single file, it makes it much more effective when extracting medical data for the examination of possible trends and long term changes in the patient condition.

While an EHR does contain the medical and treatment histories of patients, it is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broaderview of a patient's care. One of the prime features of an EHR is that, the authorized providers can create and manage the information in a digital format that is capable of being shared with other providers across more than one health care organization. It can share information with other health care providers and organizations – such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics – so they contain information from all clinicians involved in a patient's care.

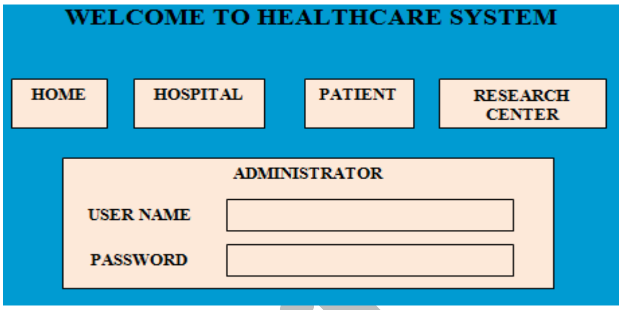


Figure 3.3 Home page of EHR system

b) Data Chunk Creation

The data records available in the file or database are spitted into data chunks. A chunk is a fragment of information. Each chunk contains a header which indicates some parameters e.g. the type of chunk, size. In the middle, there is a variable area, containing data which are decoded by the program from the parameters in the header. Chunks may also be fragments of information which are downloaded or managed by distributed programs. In distributed computing, a chunk is a set of data which are sent to a processor or one of the parts of a computer for processing. Figure 3.4 shows the creation of data chunks.

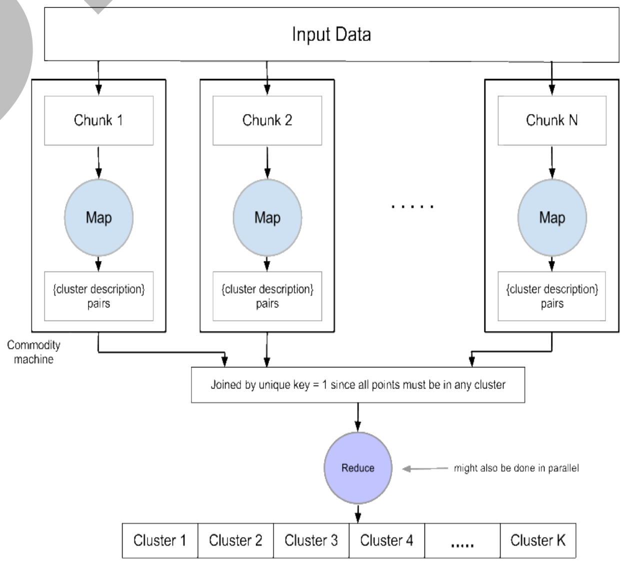

Input Data
Chunk 1
Chunk 2
Chunk N
Map
Map
Map
.....
{cluster description} pairs
{cluster description} pairs
{cluster description} pairs
Commodity machine
Joined by unique key = 1 since all points must be in any cluster
Reduce
might also be done in parallel
Cluster 1
Cluster 2
Cluster 3
Cluster 4
.....
Cluster K

Figure 3.4 Data chunk creation

The algorithm for data partition map and reduce is as follows:

Input : Data record (IDr, r), r□D, partition parameter p.
Output: D_i, $1 \le i \le p$.
Map : Generate a random number rand, Where 1 ≤rand ≤p; emit (rand, r).Reduce: For each rand, emit (null, list(r)).

Once partitioned data sets D_i , $1 \le i \le p$ are obtained , we run MRTDS (D_i,K_I,AL^0) on these datasets in parallel to derive intermediate levels AL_i^*, $1 \le i \le p$.
AL- Anonymization level

To employ anonymisation level and to manage the process of anonymisation, the followingsteps are to be followed:

Step 1 : Initializes the values of information gain and privacy loss for all specializations, withthe job IGPL initialization.

Step 2 : i) It is an iterative process, which is the main body of the algorithm.

ii) The best specialization is selected from valid specializations in current anonymisationlevel as described in step i.

iii) A specialization is said to be valid, if it satisfies two conditions. One is that its parent value is not a leaf, and the other is that the anonymity A_c(spec)>k, i.e., the data set is still k- anonymous if is performed. Then, the current anonymisation level is modified via performing the best specialization in Step ii, i.e., removing the old specialization and inserting new ones that are derived from the old one. In Step iii, information gain of the newly added specializations and privacy loss of all specializations need to be recomputed, which are accomplished by the job IGPL Update. The iteration continues until all specializations become invalid. Ultimately, D is concretely anonymised according to the resultant AL in Step 3 by the job Data Anonymisation.

MRTDS produces the same anonymous data as the serial TDS, because they follow the same steps. MTRDS mainly differs from the serial TDS in calculating IGPL values. However, calculating IGPL values dominates the scalability of TDS, as it requires TDS algorithms to count the statistical information of data sets iteratively. MRTDS exploits Map Reduce to carry out the computation of IGPL in a parallel and scalable way.

i- Denotes the count of the data to be portioned

ii- D_i - Partitioned Data
D - Data set

c) MRTDS Implementation

MRTDS implements map reduce system which contains two steps, such as map andreduce. Incremental map provides the necessary control over the alignment and granularity of the input parts. Incremental reduce processes the output of the map function grouped by the keys of the generated key value pairs. Map reduce version of MRTDS concretely conducts the computation required in TPTDS. MRTDS only leverages the task level parallelization's of map reduce. The map function emits anonymous records. The reduce function simply aggregates these anonymous records and counts their number. The figure 3.5 shows that the execution framework overview of MRTDS.
The algorithm for MRTDS driver is as follows:

Input : Data set D, the topmost AL AL^0 , and k-anonymity parameter k.
Output : Anonymous data set $D^{*.}$

Step 1 : Initialize the IGPL values for each specialization spec □ $U^m{}_{j=1}$ .Cut_j with respect toAL^0, using the job IGPL initialization;

Step 2 : While spec □ $U^m{}_{j=1}$ Cut j is valid.Find the best specialization from AL_i, $spec_{Best}$;

i. Update AL_i to AL_{i+1};

ii. Update information gain of the new specializations in AL_{i+1}, and privacy loss foreach specialization via job IGPL Update;

Step 3: Anonymise data set D in terms of resultant AL via job Data Anonymisation.

d) IGPL Analysis

The main task of IGPL Initialization is to initialize information gain and privacy loss of all specializations in the initial AL. The statistical information is required for each specialization to calculate information gain. The number of records in each current QI-group needs to be computed. The algorithm for IGPL Initialization Map is as follows:

Input : Data record (ID_r,r), r□D; anonymisation level AL.
Output: Intermediate key-value pair (Key,Count)

Step 1: For each attribute value v_i in r, find its specialization in current AL:spec; Let q be theparent in spec, and c be the q's child value that is also ancestor of v_i in TT_i;

Step 2: For each v_i, emit ((q, c, sv), count);

Step 3: Construct quasi-identifier qid*=(q1,q2...qm) where q_i,1≤ i≤ m, is the parent ofspecialization in current AL. Emit ((qid*,$,#),count);

Step 4: For each i□[1,m] ,replace q_i in qid* with its child c_i, where c_i is also the ancestor of v_i.

Let the resultant quasi identifier be qid. Emit ((qid*,p_i,#),count).

The inputs are data sets that consist of a number of records. It is the sequence number ofthe record. Step 1 gets the potential specialization for the attribute values. Then Step 2 emits key- value pairs containing the information of specialization, sensitive value, and the count information of this record. According to the above information, we compute information gain for a potential specialization in the corresponding Reduce function. Step 3 aims at computing the current anonymity, while Step 4 is to compute anonymity after potential specializations. The symbol ‘#’ is used to identify whether a key is emitted to compute information gain or anonymity loss, while the symbol ‘$’ is employed to differentiate the cases whether a key is for computing A_a(spec) or A_c(spec).

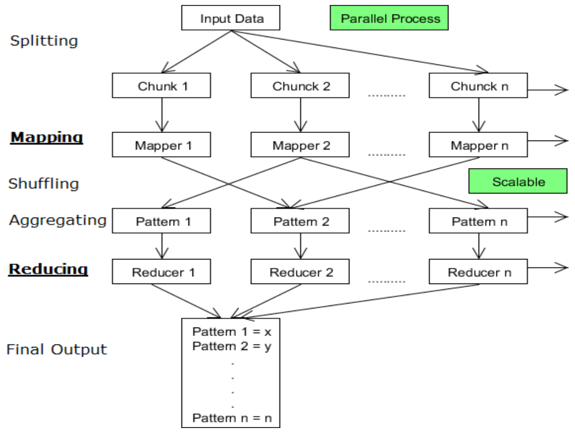


Figure 3.5 Execution framework overview of MRTDS

The algorithm for IGPL Initialization Reduce is as follows:

Input : Intermediate key-value pair (key, list (count)).

Output: Information gain (spec,IG(spec)) and anonymity (spec,A_c(spec)), (spec,A_q(spec)) for all specializations.

Step 1 : For each key, sum□□count;

Step 2 : For each key, if key.sv≠#, update statistical counts:

i. |(R_c,sv)|□sum, |R_c|□sum+|R_c|,|(R_q,sv)|□sum+|(R_q,sv)|,|R_q|□sum+|R_q|

ii. If all sensitive values for child have arrived, compute I(R_c) according to (4-3);

iii. If all children c for parent p have arrived, compute I(R_q)and IG(spec), Emit (spec

,IG(spec));

Step 3: For each key, if key.sv=#, update anonymity.

i. If key.c=$ and sum<A_q(spec), update current anonymity:A_q(spec)□sum;

ii. If key.c≠$ and sum<A_c(spec), update potential anonymity ofSpec: A_c(spec) □sum;

Step 4: Emit (spec, A_q(spec)) and emit (spec, A_c(spec)).

The first step is to accumulate the values for each input key. If a key is for computing information gain, then the corresponding statistical information is updated in Step 2(i) I(R_q), I(R_c) and IG(spec) are calculated if all the count information they need has been computed in Step 2 (ii and iii) in terms of (3) and (4). A salient Map Reduce feature that intermediate key-value pairs are sorted in the shuffle phase, makes the computation of IG(spec) sequential with respect to the order of the specializations arriving at the same reducer. Hence, the reducer just needs to keep statistical information for one specialization at a time, which makes the reduce algorithm highly scalable.

To compute the anonymity of data sets before and after a specialization, Step 3 (i) finds the smallest number of records out of all current QI-groups, and Step 3 (ii) finds all the smallest number of records out of all potential QI-groups for each specialization. Step 4 emits the results of anonymity. Note that there may be more than one key-value pair (spec, A(spec)) for one specialization in output files if more than one reducer is set. But we can find the smallestanonymity value in the driver program. Then the privacy loss PL(spec) is computed. Finally, IGPL(spec) for each specialization is obtained.

The algorithm for IGPL update map as follows:

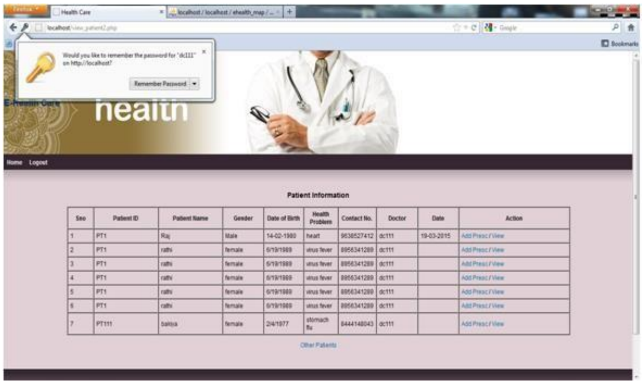
Input : Data record (ID_r, r), r□D; Anonymisation level AL.

Output: Intermediate key-value pair (key,count).

Step 1: Let attr be the attribute of the last best specialization. The value ofthis attribute in r is v. Find its specialization in current AL: spec. Let q be the parent in spec, and c be q’s child that is also an ancestor of v; Emit((q,c,sv),count);

Step 2: Construct quasi-identifier qid*=(q1,q2. .... qm), q_i ,1≤i≤m, is the parent of a specialization in current AL and is also an ancestor of v_i in r.

Step 3: For each i □[1,m], replace q_i in qid* with its child c_i if the specialization related to q_i is valid, where c_i is also the ancestor of v_i.Let the



Patient Information

Sno	Patient ID	Patient Name	Gender	Date of Birth	Health Problem	Contact No.	Doctor	Date	Action
1	PT1	Raj	Male	14-02-1990	heart	9638527412	dc111	19-03-2015	Add Presc / View
2	PT1	rathi	female	6/19/1989	virus fever	8956341289	dc111		Add Presc / View
3	PT1	rathi	female	6/19/1989	virus fever	8956341289	dc111		Add Presc / View
4	PT1	rathi	female	6/19/1989	virus fever	8956341289	dc111		Add Presc / View
5	PT1	rathi	female	6/19/1989	virus fever	8956341289	dc111		Add Presc / View
6	PT1	rathi	female	6/19/1989	virus fever	8956341289	dc111		Add Presc / View
7	PT111	baliya	female	2/4/1977	stomach flu	8444148043	dc111		Add Presc / View

Other Patients

resultant quasi-identifier be qid.
Emit ((qid, qi,
#),count).

The IGPL Update job dominates the scalability and efficiency of MRTDS, since it is executed iteratively as described in MRTDS driver. The IGPL Update job is quite similar to IGPL Initialization, except that it requires less computation and consumes less network bandwidth. Thus,the former is more efficient than the latter. IGPL update map describes the Map function of IGPL Update. The Reduce function is the same as IGPL Initialization, already described in IGPLinitialization reduce.

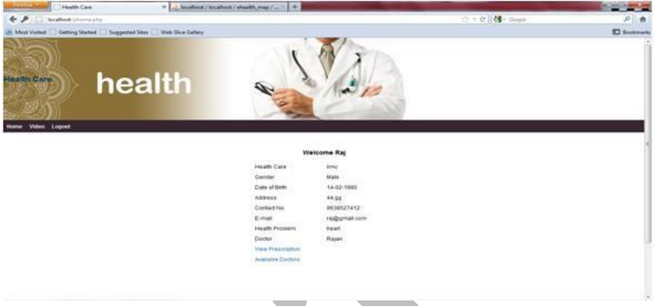
After a specialization spec is selected as the best candidate, it is required to compute the information gain for the new specializations derived from spec. So, Step 1 in IGPL update map only emits the key-value pairs for the new specializations, rather than all in IGPL initialization map. Note that it is unnecessary to re-compute the information gain of other specializations because conducting the selected specialization never affects the information gain of others. Compared with IGPL initialization, only part of data is handled and less network bandwidth is consumed.

On the contrary, the anonymity values of other specializations will be influenced with high probability because splitting QI-groups according to spec changes the minimality of the smallest QI-group in the last round. Therefore, it is needed to compute A_c(spec) for all specializations in AL, as described in Step 2 and 3 of IGPL update map. Yet A_q (spec) can be directly obtained from the statistical information kept by the last best specialization. Note that if the specialization related to qi in Step 3 is not valid, no resultant quasi-identifier will be created.

Since the IGPL update job dominates the scalability and efficiency of MRTDS, hence analyze its complexity as follows. Let n denote all the records in a data set, m be the number of attributes, s be the number of mappers and t be the number of reducers. As a mapper emits (m+1) key-value pairs, it takes O(1) space and O(m*n/s) time a reducer takes O(1) space and O(m*n/t) time. Note that a reducer only needs O(1) space due to the Map Reduce feature that the key value pairs are sorted in the shuffle phase. Otherwise, the reducer needs more space to accumulate statistic information for a variety of specializations. The communication cost is O(m*n) according to the map function, but communication traffics can be reduced heavily by Map Reduce techniqueslike Combiner.

4. RESULTS

Login of Admin

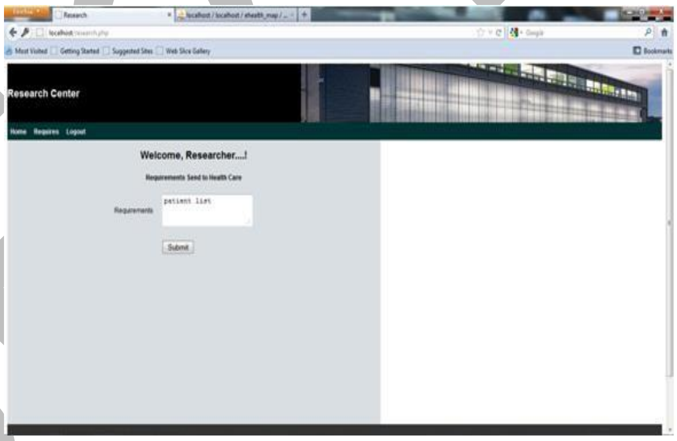


Welcome Raj

Health Care	kmc
Gender	Male
Date of Birth	14-02-1990
Address	44,gg
Contact No.	9638527412
E-mail	raj@gmail.com
Health Problem	heart
Doctor	Rajan

View Prescription
Available Doctors

Viewing patient details

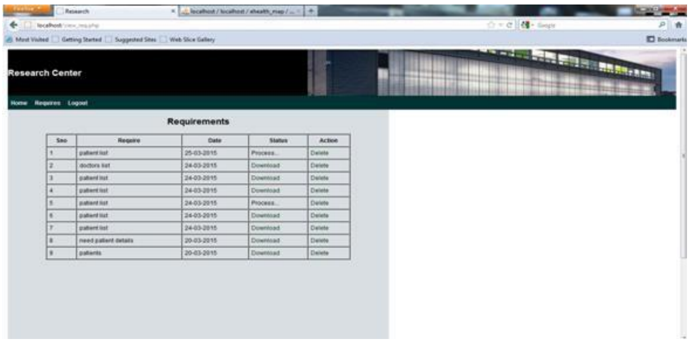


Welcome, Researcher....!

Requirements Send to Health Care

Requirements: patient list

Raising query to the admin by researcher


Research
Most Visited Getting Started Suggested Sites Web Slice Gallery
Research Center
Home Requires Logout

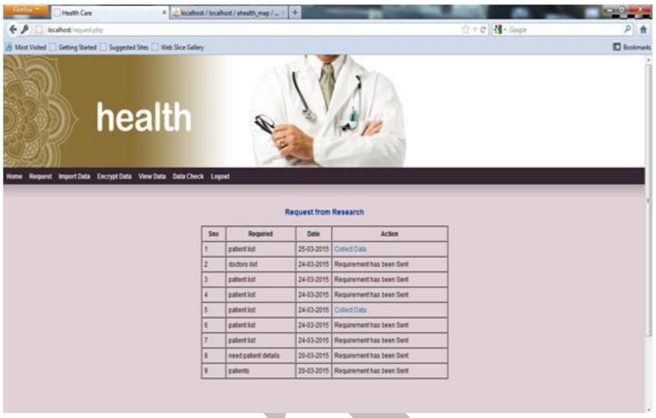
Requirements

Sno	Require	Date	Status	Action
1	patient list	25-03-2015	Process...	Delete
2	doctors list	24-03-2015	Download	Delete
3	patient list	24-03-2015	Download	Delete
4	patient list	24-03-2015	Download	Delete
5	patient list	24-03-2015	Process...	Delete
6	patient list	24-03-2015	Download	Delete
7	patient list	24-03-2015	Download	Delete
8	need patient details	20-03-2015	Download	Delete
9	patients	20-03-2015	Download	Delete

ISBN: 978-81-910765-1-6

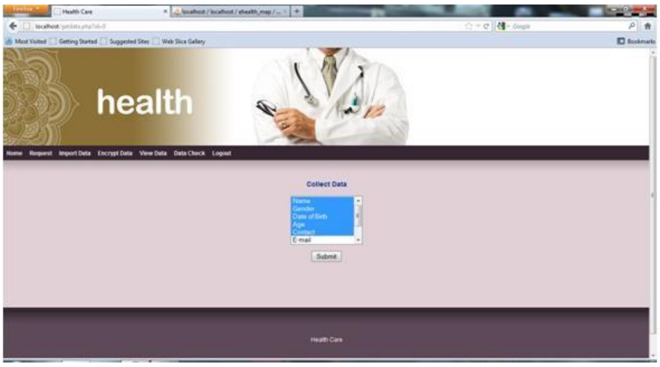
Processing of the query

Responding to the query by the admin



Request from Research

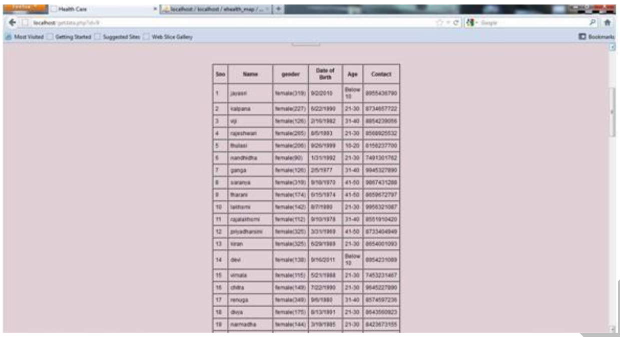
Sno	Required	Date	Action
1	patient list	25-03-2015	Collect Data
2	doctors list	24-03-2015	Requirement has been Sent
3	patient list	24-03-2015	Requirement has been Sent
4	patient list	24-03-2015	Requirement has been Sent
5	patient list	24-03-2015	Collect Data
6	patient list	24-03-2015	Requirement has been Sent
7	patient list	24-03-2015	Requirement has been Sent
8	need patient details	20-03-2015	Requirement has been Sent
9	patients	20-03-2015	Requirement has been Sent


Health Care
localhost/getdata.php?id=9
health
Home Request Import Data Encrypt Data View Data Data Check Logout
Submit
Health Care

Collect Data

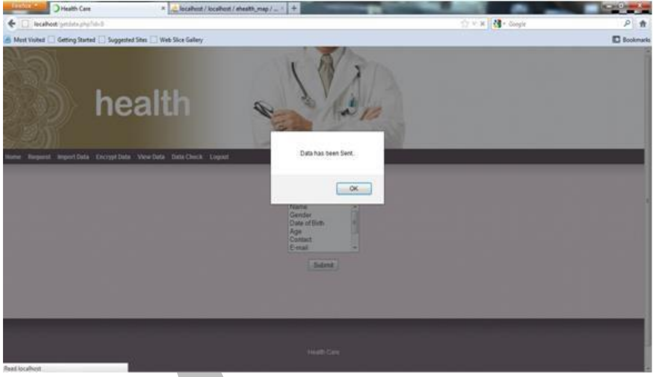
Name
Gender
Date of Birth
Age
Contact
E-mail

Selecting the fields required by the researcher



Sno	Name	gender	Date of Birth	Age	Contact
1	jayasri	female(318)	9/2/2010	Below 10	8955436790
2	kalpana	female(227)	6/22/1990	21-30	8734657722
3	viji	female(126)	2/16/1982	31-40	8854239056
4	rajeshwari	female(265)	8/5/1993	21-30	8568925532
5	thulasi	female(206)	9/26/1999	10-20	8156237700
6	nandhidha	female(90)	1/31/1992	21-30	7491301762
7	ganga	female(126)	2/5/1977	31-40	9945327890
8	saranya	female(318)	9/18/1970	41-50	9867431288
9	tharani	female(174)	6/15/1974	41-50	8659672797
10	lakhsmi	female(142)	8/7/1990	21-30	9956321087
11	rajalakhsmi	female(112)	9/10/1978	31-40	8551910420
12	priyadharsini	female(325)	3/31/1969	41-50	8733404949
13	kiran	female(325)	6/29/1989	21-30	8654001093
14	devi	female(138)	9/16/2011	Below 10	8954231089
15	vimala	female(115)	5/21/1988	21-30	7453231467
16	chitra	female(149)	7/22/1990	21-30	9645227890
17	renuga	female(349)	9/6/1980	31-40	8574597236
18	divya	female(175)	8/13/1991	21-30	8643560923
19	narmadha	female(144)	3/19/1985	21-30	8423673155

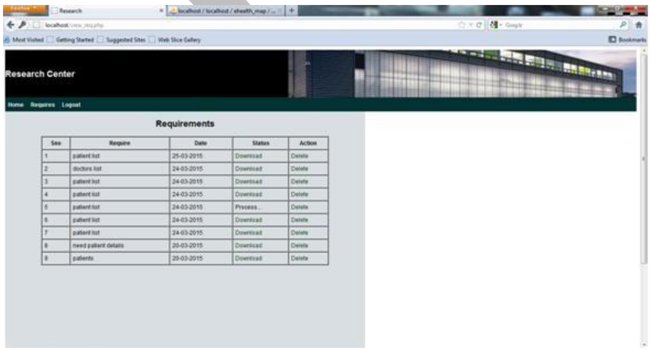
Displaying the details to be sent


Health Care
localhost/getdata.php?id=9
health
Home Request Import Data Encrypt Data View Data Data Check Logout
OK
Submit
Health Care

Data has been Sent.

Name
Gender
Date of Birth
Age
Contact
E-mail

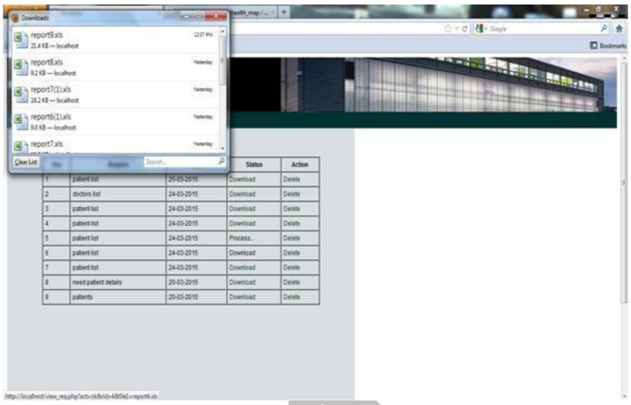
Data has been sent to the researcher



Requirements

Sno	Require	Date	Status	Action
1	patient list	25-03-2015	Download	Delete
2	doctors list	24-03-2015	Download	Delete
3	patient list	24-03-2015	Download	Delete
4	patient list	24-03-2015	Download	Delete
5	patient list	24-03-2015	Process...	Delete
6	patient list	24-03-2015	Download	Delete
7	patient list	24-03-2015	Download	Delete
8	need patient details	20-03-2015	Download	Delete
9	patients	20-03-2015	Download	Delete

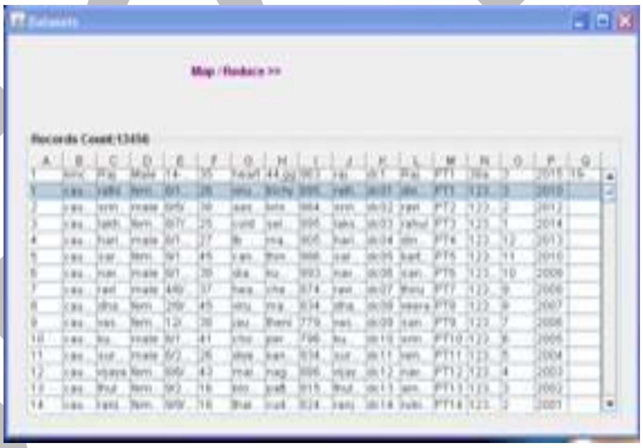
Data received by the researcher


Downloads
Clear List
Search...
http://localhost/view_req.php?act=ok&rid=4&file1=report4.xls

report9.xls 21.4 KB — localhost 12:37 PM
report8.xls 9.2 KB — localhost Yesterday
report7(1).xls 18.2 KB — localhost Yesterday
report6(1).xls 9.0 KB — localhost Yesterday
report7.xls Yesterday

Sno	Require	Date	Status	Action
1	patient list	25-03-2015	Download	Delete
2	doctors list	24-03-2015	Download	Delete
3	patient list	24-03-2015	Download	Delete
4	patient list	24-03-2015	Download	Delete
5	patient list	24-03-2015	Process...	Delete
6	patient list	24-03-2015	Download	Delete
7	patient list	24-03-2015	Download	Delete
8	need patient details	20-03-2015	Download	Delete
9	patients	20-03-2015	Download	Delete

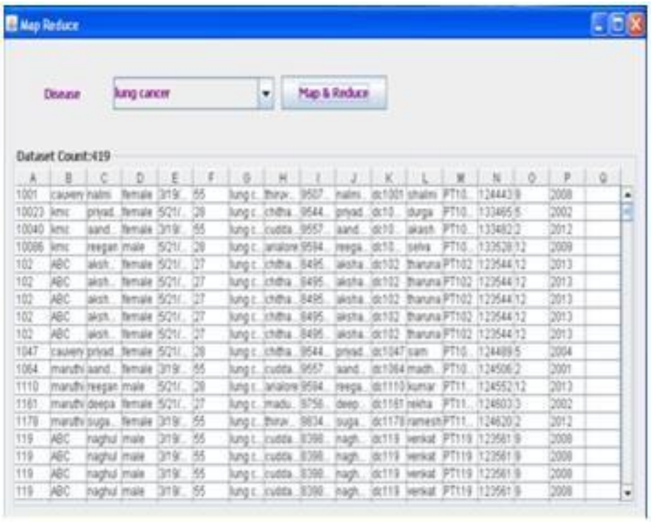
Downloading the data received



Records Count:13456

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	kmc	Raj	Male	14-	35	heart	44 gy	961	raj	dr1	Raj	PT1	394	3	2015	15
1	cau	rathi	fem	6/1	26	renu	trichy	995	rath	dr01	ab	PT1	123	3	2010	
2	cau	srin	male	6/6/	30	sas	kela	984	srin	dr02	ravi	PT2	123	2	2012	
3	cau	lakh	fem	8/7/	25	cold	sel	995	laks	dr03	rahul	PT3	123	1	2014	
4	cau	hari	male	6/1	27	flu	ma	905	hari	dr04	sim	PT4	123	12	2013	
5	cau	sar	fem	9/1	45	can	thon	998	sar	dr05	kart	PT5	123	11	2010	
6	cau	nan	male	9/1	30	dia	ku	993	nan	dr06	san	PT6	123	10	2009	
7	cau	ravi	male	4/6/	37	hea	che	874	ravi	dr07	thiru	PT7	123	9	2008	
8	cau	dha	fem	2/9/	45	viru	ma	834	dha	dr08	veera	PT8	123	8	2007	
9	cau	nas	fem	12/	30	jau	theni	779	nas	dr09	san	PT9	123	7	2006	
10	cau	ku	male	9/1	41	cho	pan	796	ku	dr10	sim	PT10	123	6	2005	
11	cau	sur	male	6/2	26	dise	kan	934	sur	dr11	ken	PT11	123	5	2004	
12	cau	vijaya	fem	6/6/	43	mal	nag	996	vijay	dr12	man	PT12	123	4	2003	
13	cau	thul	fem	9/2	16	sto	pati	915	thul	dr13	am	PT13	123	3	2002	
14	cau	rani	fem	9/9/	16	thai	cud	824	rani	dr14	subi	PT14	123	2	2001	

Huge amount of data records



Disease: lung cancer

Dataset Count:419

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1001	cauvery	nalini	female	3/19/...	55	lung c...	thirav...	9507...	nalini...	dc1001	shalini	PT10...	124443	9	2008	
10023	kmc	priyad...	female	5/21/...	28	lung c...	chitha...	9544...	priyad...	dc10...	durga	PT10...	133465	5	2002	
10040	kmc	aand...	female	3/19/...	55	lung c...	cudda...	9557...	aand...	dc10...	akash	PT10...	133482	2	2012	
10086	kmc	reegan	male	5/21/...	28	lung c...	arialore	9594...	reega...	dc10...	selva	PT10...	133528	12	2009	
102	ABC	aksh...	female	5/21/...	27	lung c...	chitha...	8495...	aksha...	dc102	tharuna	PT102	123544	12	2013	
102	ABC	aksh...	female	5/21/...	27	lung c...	chitha...	8495...	aksha...	dc102	tharuna	PT102	123544	12	2013	
102	ABC	aksh...	female	5/21/...	27	lung c...	chitha...	8495...	aksha...	dc102	tharuna	PT102	123544	12	2013	
102	ABC	aksh...	female	5/21/...	27	lung c...	chitha...	8495...	aksha...	dc102	tharuna	PT102	123544	12	2013	
102	ABC	aksh...	female	5/21/...	27	lung c...	chitha...	8495...	aksha...	dc102	tharuna	PT102	123544	12	2013	
1047	cauvery	priyad...	female	5/21/...	28	lung c...	chitha...	9544...	priyad...	dc1047	sam	PT10...	124489	5	2004	
1064	maruthi	aand...	female	3/19/...	55	lung c...	cudda...	9557...	aand...	dc1064	madh...	PT10...	124506	2	2001	
1110	maruthi	reegan	male	5/21/...	28	lung c...	arialore	9594...	reega...	dc1110	kumar	PT11...	124552	12	2013	
1161	maruthi	deepa	female	5/21/...	27	lung c...	madu...	9756...	deep...	dc1161	rekha	PT11...	124603	3	2002	
1178	maruthi	suga...	female	3/19/...	55	lung c...	thirav...	9834...	suga...	dc1178	ramesh	PT11...	124620	2	2012	
119	ABC	raghul	male	3/19/...	55	lung c...	cudda...	8398...	ragh...	dc119	venkat	PT119	123561	9	2008	
119	ABC	raghul	male	3/19/...	55	lung c...	cudda...	8398...	ragh...	dc119	venkat	PT119	123561	9	2008	
119	ABC	raghul	male	3/19/...	55	lung c...	cudda...	8398...	ragh...	dc119	venkat	PT119	123561	9	2008	
119	ABC	raghul	male	3/19/...	55	lung c...	cudda...	8398...	ragh...	dc119	venkat	PT119	123561	9	2008	

Reduced amount of data records

5. CONCLUSION

In this paper, scalability problem of large-scale data anonymization by TDS has been implemented for an electronic health record maintenance application. A highly scalable two-phase TDS approach using MapReduce on cloud has been proposed. In the first phase, data sets were partitioned and anonymized. In the second phase, the intermediate results, generated by the first phase were merged and further anonymized for producing consistent k-anonymous data sets. In cloud environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. Based on the contributions herein, we plan to further explore the next step onscalable privacy preservation aware analysis and scheduling on large-scale data set.

REFERENCES

1. S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud," Proc. 31st Symp. Principles of Database Systems (PODS '12), pp. 1-4, 2012.
2. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53,no. 4, pp. 50-58, 2010.
3. L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 2,pp.296-303, Feb. 2012.
4. B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Devel- opments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.
5. T. Iwuchukwu and J.F. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), pp. 746-757, 2007.
6. B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.
7. N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4, Article 18, 2010.
8. B. Fung, K. Wang, L. Wang, and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data and Knowledge Eng., vol. 68, no. 6, pp. 552-575, 2009.
9. N. Mohammed, B.C. Fung, and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants," VLDB J., vol. 20, no. 4, pp. 567-588, 2011.
10. W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," VLDB J., vol. 15, no. 4, pp. 316-333, 2006.
11. N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4, Article 18, 2010.
12. P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers," Proc. 23rd Ann. IFIP WG 11.3 Working Conf. Data and Applications Security XXIII (DBSec '09), pp. 191-207, 2009.