Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 6, June-2018

De-Identification of Health Data under Efficient Recommendation using Map Reduce

Nancy Rupani[1], Koyana Wavhal [2], Upender Yadav [3], Rajat Shrivastava[4], Abha Jain [5]

[1,2,3,4] Student, Department of Computer Engineering

[5]Assistant Professor, Department of Computer Engineering

[1,2,3,4,5] DYPIEMR, Pune, India

Abstract— Many knowledge homeowners square measure needed to unleash the information during a sort of world application, since it's of great importance to discover valuable data keep behind the information. However, prevailing re-identification attacks on the AOL and ADULTS knowledge sets have shown that publishing such data directly might pose huge threats to the individual privacy. Thus, it's imperative to resolve every kind of re-identification risks by recommending effective de-identification policies to ensure each privacy and utility of the information. De-identification policies is one amongst the models which we will need to succeed such needs, however, the quantity of de-identification policies is exponentially massive thanks to the broad domain of quasi-identifier attributes. To manage the trade off between knowledge utility and knowledge privacy, skyline computation will be needed to choose such policies, however it's nevertheless difficult for economical skyline process over sizable amount of policies. During this paper, we tend to propose one parallel algorithmic rule known as SKY-FILTER-MR, that relies on Map scale back to beat this challenge by computing skylines over massive scale de-identification policies that's drawn by bit-strings. For improving the performance, a completely unique approximate skyline computation theme was projected to prune unqualified policies exploitation through the domination relationship. With approximate skyline, the facility of filtering within the policy area generation stage was greatly strong to effectively decrease the value of skyline computation over various policies. In depth experiments over each real world and artificial datasets demonstrate that our projected SKY-FILTER-MR algorithmic rule well outperforms the baseline approach by up to fourfold faster within the best case, that indicates sensible quantifiability over massive policy sets.

***Keywords*-Big Data; Access Control;Privacy-preserving Policy; De-identification policies.**

I. Introduction

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data.

II.Existing system

Existing re-identification attacks on the AOL and ADULTS knowledge sets have shown that publishing such data directly might cause tremendous threats to the individual privacy. Thus, it's important to resolve all types of re-identification risks by recommending effective de-identification policies to ensure each privacy and utility of the information. Their work has limitations in many ways. First, their framework needs a lattice that contains all the choice policies to arrange with time value. Second, their

algorithms are approximate approaches that haven't any guarantee of best resolution.We give the formal definition of recommendation over de-identification policies and denote the problem as RIDP. We propose algorithms using MapReduce to speedup the parallel computation efficiency and obtain high scalability.Through analyzing the characteristics of data distribution,we give the formal definition of independent property, which can be used to generate new policies effectively.To reduce the sort overhead of skyline and decrease the number of alternative policy set in Map phase of the first round, a new scheme was introduced for recommendation of de-identification policies. We demonstrate the superiority of our methods through extensive experiments, and the results show that our approach can preserve the privacy substantially with high data utility and query efficiency.

III.Proposed system

In this paper, we propose a system wherein the patients record would be safeguarded and also it will be diagnosed properly. This system helps to maintain a centralized data system whose admin will have the entire data but not with everyone. The patient would login and feed the system with their symptoms. The symptoms would be identified from the training datasets we have and then the related disease is reduced. The specialized doctor gets the notification of the patient and thus the patient is treated and provided with the cure. If the patient is not satisfied the diagnosis then they may go to the chief doctor as well and the same procedure is followed.

All this identification is done through the sky-filter MR algorithm and KNN clustering algorithm where all the diseases are provided with key-value pair and closest to this key-value pair the disease is identified. Thus, the system protects users sensitive data and helps them to get the cure without the fear of endangering their privacy.

System Architecture

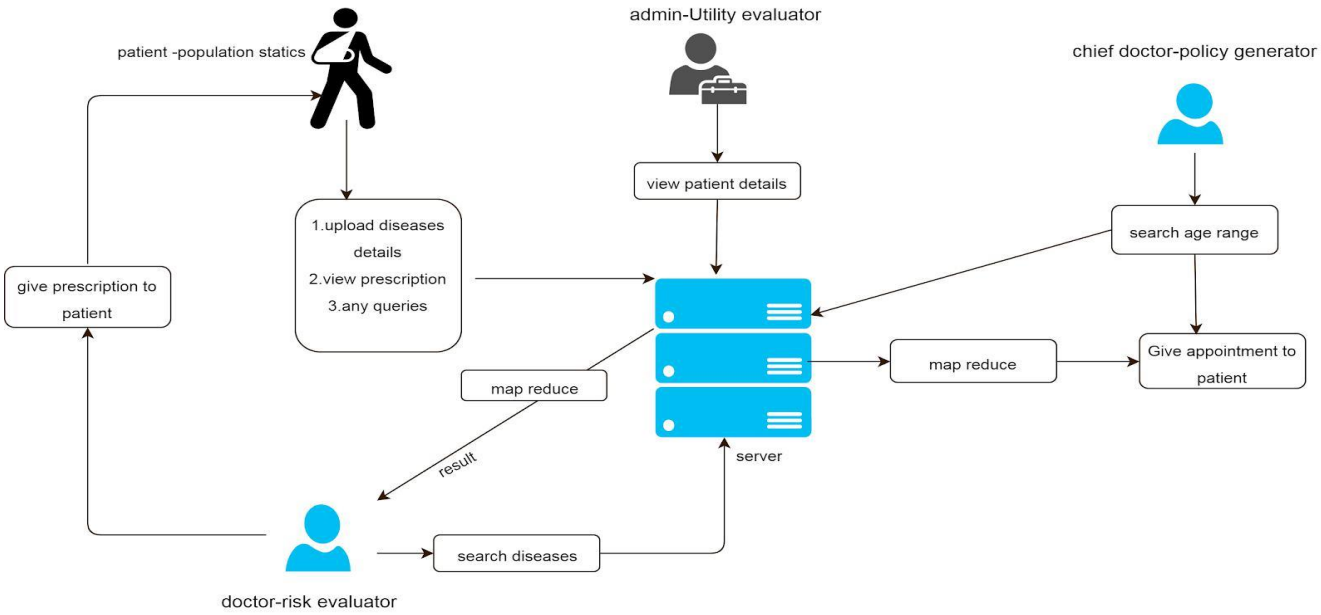

Fig:- Proposed System Architecture.

V.Methodology

Map Reduce:-MapReduce is a framework for processing computation on Bigdata like hadoop using large number of systems, commonly known as a cluster or a grid . Processing can occur on structured as well as unstructured database like hadoop,cloud,relational etc. MapReduce can take advantage of the locality of data, processing it near the place it is stored in order to minimize communication overhead.

A MapReduce framework is mainly composed of three operational steps:

1. **Map:** each active system applies the map function to the local dataset, and writes the output to a dynamic storage. A parent system or node ensures that only one copy of redundant input data is processed.
2. **Shuffle:** worker nodes reorient the data based on the output keys produced by the map function, such that all data belonging to one key is located on the same worker node.
3. **Reduce:** worker nodes now process each group of output data, per key, in parallel.

Algorithm :
1) **SKY-FILTER-MR**({*S*,(*R*,*U*)},(β,α,*h*)
Input: *{S;* (*R;U*): the policy space, _: the sampling parameter,
": the precision, *h*: the depth of filter.
Output: *F*: the policy skyline frontier.
1: *{G;* (*R;U)}*= FilterPolicy-M(*{S;* (*R;U*)}β,α, *h*);
2: *F* = SKY-MIN-MR(*{;G* (*R;U*),β}_)
3: **Function** *{G;* (*R;U)}* = FilterPolicy-M(*{S;* (*R;U*),α, *h*);
4: **map**(key = R,value = string);//string = R + U + policy
5: risk = key, utility = string.U;
6.flag=false

7.**for** each policy in *G* by inverse sequence **do**// reverse traversal
8.a = risk*α - policy.R, b = utility*α - policy.U;
9.depth = 0;// the depth of reverse traversal
10.**if** (a*b)_ 0 and (a+b)>0 **then**
// a >=0, b>=0, a+b≠0
11.depth = depth + 1;
12.flag = true;//policy is in the approximately dominated
area
13.break;
14.**if** depth > *h* **then**
15.break;
16.**if** flag == false **then**
17.policy -> *G*;
18.output(key,value);
19.**return** *{G;* (*R;U)}*;

2) **KNN(K-Nearest Neighbor)-**K nearest neighbor is a simple predictive algorithm that stores all available cases and classifies new cases based on different parameters like distance , functions,class etc.

pseudo code-

1. Load the data from the database
2. Initialise the value of k=1
3. For getting the predicted class, iterate from 1 to total number of data in the training data set
    a. Calculate the distance between sample data and each row of training data. Here we will use Euclidean distance as our distance metric as it is the most used method.
    b. Sort the calculated distances in either ascending or descending order based on users choice.
    c. Get top k rows from the modified array
    d. Get the most frequent class of these rows
    e. Return the identified class.

3)**ITERATIVE ALGORITHM-**The conceptual idea for iterative algorithms in MapReduce is to link multiple jobs together, using the output of the last one as the input of the next one. An important consideration is that, given the usual size of the data, the termination condition must be computed within the MapReduce program.

A particular functionality has to be carried out iteratively till grouping criteria meets. At a time a Mapping is done and then te

reduce function is called for doing work in one iteration. After that the need to call Map-reduce then again map-reduce then again map-reduce till certain condition is satisfied. The most important of all the process is the termination condition. Unless the proper termination condition is met the process will go on.

IV.OUTPUT

In this section we evaluate the execution of the SKY-FILTER and KNN algorithm.

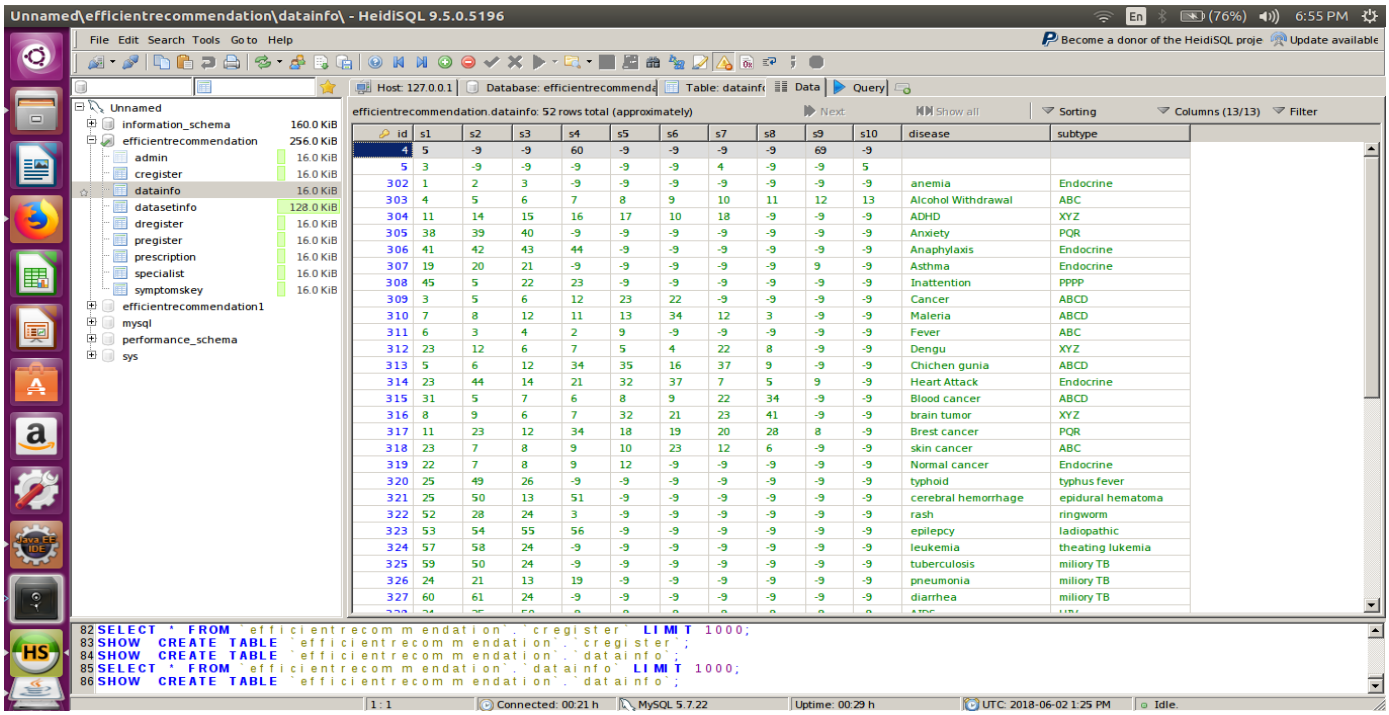

Fig:-Symptom's keys for identification of diseases.

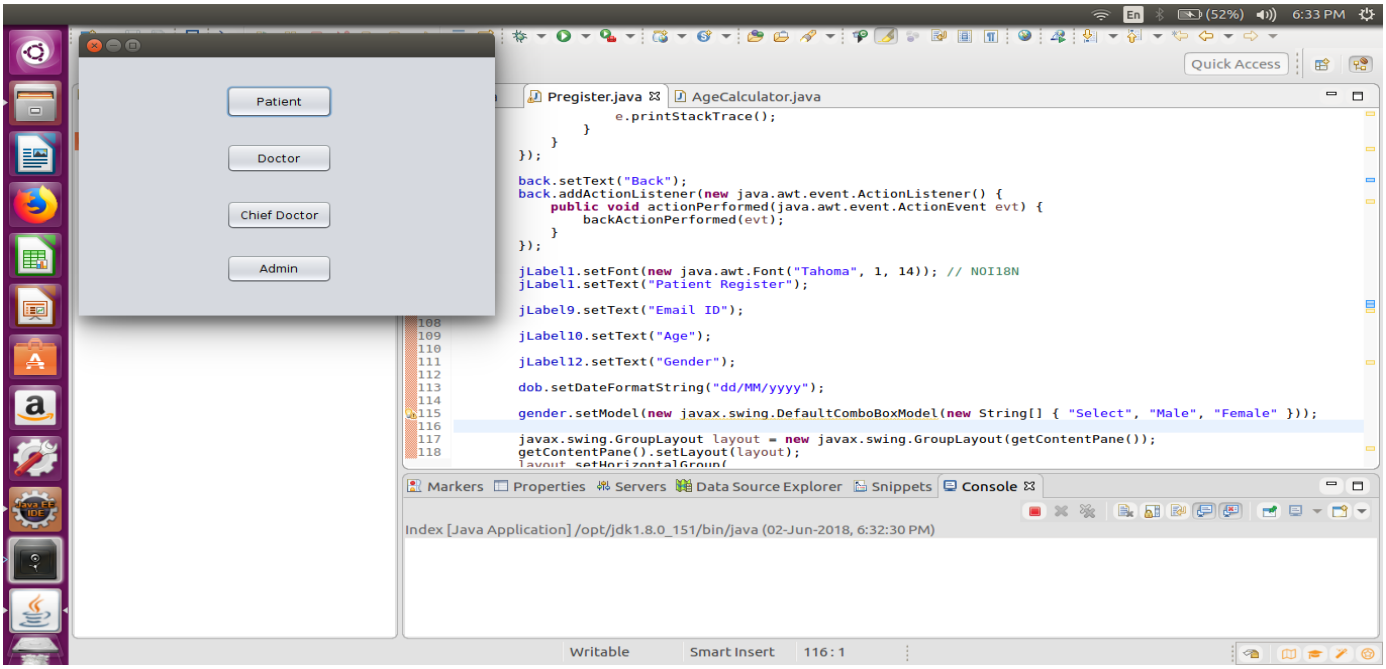

Fig:-Login Page.

VI.Conclusion

We study the recommendation on a great number of De-identification policies using Map Reduce. To put forward an effective way of policy generation on the basis of newly proposed definition, which can decreases the time of generating policies and the size of alternative policy set dramatically. We perform comprehensive experimental evaluation on both real-world and synthetic datasets, and the results indicate good performance and scalability.

References

[1] Xiaofeng Ding, Li Wang, Zhiyuan Shao, and Hai Jin, "Efficient Recommendation of De-identification Policies using Map-Reduce "DOI 10.1109/TBDATA.2017.2690660, IEEE

[2] X. MA, H. Li, J. Ma, Q. Jiang, S. Gao, N. Xi, and D. Lu, "Applet: A privacy-preserving framework for location-aware recommender system," *Sci China Inf Sci*, vol. 59, no. 2, pp. 1–15, 2016.W. Xia, R. Heatherly, X. Ding, J. Li, and B. Malin, "Efficient discovery of de-identification policies through a risk-utility frontier," in *CODASPY*, 2013, pp. 59–70.

[3] K. Benitez, G. Loukides, and B. Malin, "Beyond safe harbor: Automatic discovery of health information de-identification policy alternatives," in *IHI*, 2010, pp. 163–172.

[4] K. E. Emam, "Heuristics for de-identifying health data," *IEEE Security and Privacy*, vol. 6, no. 4, pp. 58–61, 2008.

[5] L. Sweeney, "*k*-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 555–570, 2002.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond *k*-anonymity," in *TKDD*, 2007, pp. 1–52.

[7] N. Li, T. Li, and S. Venkatasubramanian, "*t*-closeness: Privacybeyond *k*-anonymity and ℓ-diversity," in *ICDE*, 2007, pp. 106–115.

[8] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 70–78.

[9] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1388–1399, 2012.

[10] W. Xia, R. Heatherly, X. Ding, J. Li, and B. A. Malin, "Ru policy frontiers for health data de-identification," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1029–1041, 2015.

[11] B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no.