Secure & Efficient Large Scale Healthcare Data Storage System in Hadoop

Priyanka Achar
M.Tech Final Year, Department of Information Science & Engineering,
Dayananda Sagar College of Engineering,
Bangalore, Karnataka

Rekha Jayaram
Assistant Professor, Department of Information Science & Engineering,
Dayananda Sagar College of Engineering,
Bangalore, Karnataka

Abstract: The measure of computerized information created worldwide is exponentially developing. While the wellspring of this information, all in all known as Big Data, fluctuates from among versatile administrations to digital physical frameworks and past, the invariant is their undeniably fast development for a long time to come. Tremendous information exist, in healthcare space and to inquire about in human sociologies, that rouse preparing progressively greater information to separate data and learning in order to enhance procedures and advantages. Thusly, the requirement for more proficient figuring frameworks custom-made to such huge information applications is progressively increased. Such custom structures would expectedly grasp heterogeneity to better match each period of the calculation. We propose a review state of the art as well as envisioned future large-scale data processing customized for batch processing of big data applications in the MapReduce technique. We likewise give our perspective of current imperative patterns significant to such frameworks, and their effects on future models and compositional highlights anticipated that would address the requirements of tomorrow huge information handling in this worldview.

We additionally explore secure incorporation of MapReduce into our plan, which makes our plot greatly reasonable for distributed computing condition. Careful security examination and numerical investigation do the execution of our plan as far as security and effectiveness.

Keywords: MapReduce, distributed computing, Hadoop, Resilient Distributed Datasets (RDD), Pregel.

I. INTRODUCTION

Ideal models for enormous information investigation can be extensively ordered into clump handling and stream preparing. As the names infer, the previous is utilized when the information is as of now gathered, for example, the instance of record age for web wide hunt by Google, while the last is commonly utilized when the information is delivered on the web and is intended to be prepared on the fly, for example, the instance of examining the twits posted on Twitter. Here we mention the previous class and the structures to enhance its execution. MapReduce presented by Google is among the most generally utilized programming standards in this class, and Hadoop is its open-source usage that made it accessible to numerous different clients outside that organization. Various different overviews exist on MapReduce, however their objective is chiefly giving a more profound comprehension of the MapReduce worldview and its product executions or a particular utilization of it; to the best of our insight this is the main study concentrating on different models, equipment programming, and equipment just ones, proposed to enhance execution and proficiency of MapReduce calculation.

It is imperative that various expansions to the fundamental MapReduce worldview or general enormous information bunch handling likewise exist that are picking up notoriety; this incorporates Spark and furthermore Pregel. Start is celebrated for its in-memory registering ability and also different highlights, for example, bolster for iterative calculations, execution of a stream diagram of activities, and not withstanding gushing highlights. Pregel is particularly intended to parallelize chart preparing calculations planned to be connected on huge diagrams. Start presents Resilient Distributed Datasets (RDD) as its essential information structure, and Pregel characterizes Super steps as its unit of calculation advance where tasks are performed in and on vertices of the enormous diagram dispersed on specialist hubs. Therefore, the execution display, including calculation and also correspondence models and examples, in such standards is unique in relation to the first MapReduce, and subsequently, they require their own

particular design examination and changes. In this paper, we crack down on the first MapReduce worldview, its execution display, and proposed custom models for it

II. Related Works

MapReduce Framework

MapReduce is the programming system for preparing expansive scale datasets in a disseminated way. As appeared in Fig.1, to process an undertaking with enormous measures of information, MapReduce isolates the assignment into two stages: outline lessen. These two stages are communicated with outline lessen capacities, which take <key,value> matches as info and yield information arrange. In a bunch, hubs that are in charge of guide and lessen capacities are called mappers and reducers separately. In a MapReduce assignment, the system parts input datasets into information pieces, which are prepared by autonomous mappers in parallel. Each guide work forms information and creates middle yield as <key,value> sets. These moderate yields are sent to reducers after rearrange. As indicated by the key space of <key,value> matches in middle of the road yields, every reducer will be relegated with a parcel of sets. In MapReduce, middle of the road <key,value> yields with the same key are sent to a similar reducer. From that point forward, reducers sort and gather every single middle of the road yield in parallel to produce the last result. Nevertheless, all these designs only focus on improving the computational performance over large-scale datasets, and none of them take privacy protection into consideration.

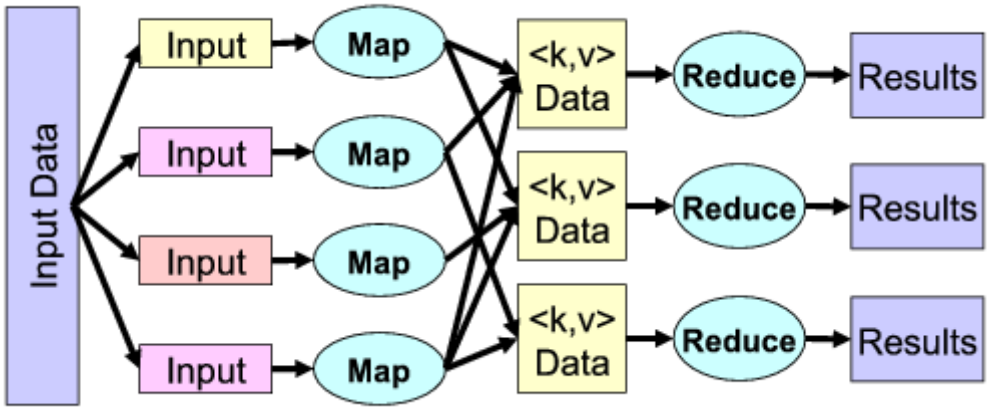

Fig. 1.MapReduce Framework

Privacy-preserving Clustering

The issue of protection safeguarding clustering has likewise been contemplated in the circulated setting. These plans principally depend on multi-party secure calculation procedures; for example, secure circuit assessment, homomorphic encryption and unaware exchange. Protection safeguarding circulated grouping has an alternate reason with security safeguarding outsourcing of grouping. These plans include various elements, which perform grouping over their common information without unveiling their information to each other. In an unexpected way, the dataset in clustering outsourcing is possessed by a solitary substance, which needs to limit neighborhood computational cost for substantial scale clustering. A different line of research that is identified with this work is protection saving KNN seeks, since both K-means and KNN utilize Euclidean separation to gauge the similitude of information vectors. An effective grid based security safeguarding KNN look plot is first proposed by Wong et al., in which they change over the Euclidean separation correlation with scalar item calculation. By and by, as exhibited by Yao et al., is powerless against the straight investigation assault when the cloud server gets an arrangement of information objects from the dataset. To overcome such a security vulnerability, Yao et al. present a secure solution by adopting a novel partition-based secure Voronoi diagram design. Unfortunately, their scheme only supports data with no more than two dimensions, and thus becomes impractical for most types of data in the domain of clustering. In an unexpected way, our proposed plan can bolster information of any number of measurements, is impervious to straight investigation assaults as appeared and does not present any precision misfortune. Moreover, thinking about privacy preserving Euclidean separation correlation just, our plan fundamentally lessens computational cost and capacity overhead contrasted. Furthermore, stretching out protection saving KNN to help the outsourcing of K-means grouping isn't a paltry undertaking. Not at all like the KNN look through that is a solitary round assignment, K-means clustering is an iterative procedure and requires the refresh of grouping focuses in view of all articles in the dataset after each round of clustering. To ensure the productivity and security of the whole clustering process, our plan particularly makes these updates good with MapReduce and enables them to be principally taken care of by the cloud server over cipher texts. Especially, the dataset proprietor just needs to play out a consistent number of activities for the refresh of grouping focuses as which is autonomous to the span of the expansive scale dataset.

III. PROBLEM STATEMENT

The problem of privacy-preserving clustering has also been studied in the distributed setting. These schemes mainly rely on multi-party secure computation techniques, such as secure data evaluation, homomorphic encryption and oblivious transfer. Nevertheless, privacy-preserving distributed clustering has a different purpose with privacy-preserving outsourcing of clustering. These designs involve multiple entities, which perform clustering over their shared data without disclosing their data to each other. Differently, the dataset in clustering outsourcing is owned by a single entity, who wants to minimize local computational cost and storage space for large-scale clustering. Another line of research that is related to this work is privacy-preserving KNN clustering, since both K-means and KNN use Euclidean distance to measure the similarity of data vectors. To overcome security vulnerability, Yao et al. present a secure solution by adopting a novel partition-based secure data storage system which has lot of limitations. To overcome these problem, the proposed system uses Map Reduce technique for multiple user environment with KNN clustering and DNA Light Weight Encryption.

IV. PROPOSED SYSTEM

In the first place we are introducing prepared informational collection for each unique group which is identified with restorative Information. After, the clustering calculation partition document into number of lumps and for each block hash code is created for the security reason. Before putting away into HDFS System, arrangement calculation orders that record have a place with which bunch class.

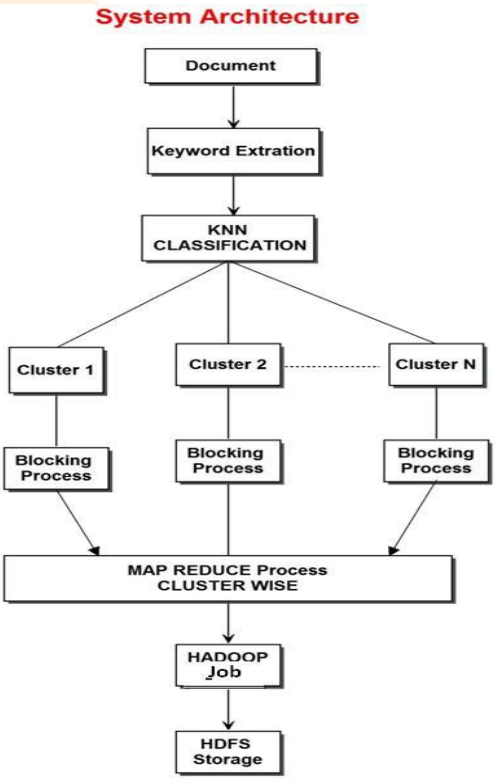

Fig 2: System Architecture of Real-Time Healthcare Data into HDFC

Unfortunately, these security saving KNN seek plans are restricted by the powerlessness to direct investigation assaults, the help up to two measurement information, or precision misfortune. In expansion, KNN is a solitary round inquiry errand, however K-mean grouping is an iterative procedure that requires the refresh of grouping fixates in view of the whole dataset after each round of grouping. Thinking about the productive help over vast scale datasets, these refresh forms additionally should be outsourced to the cloud server in a protection saving way.

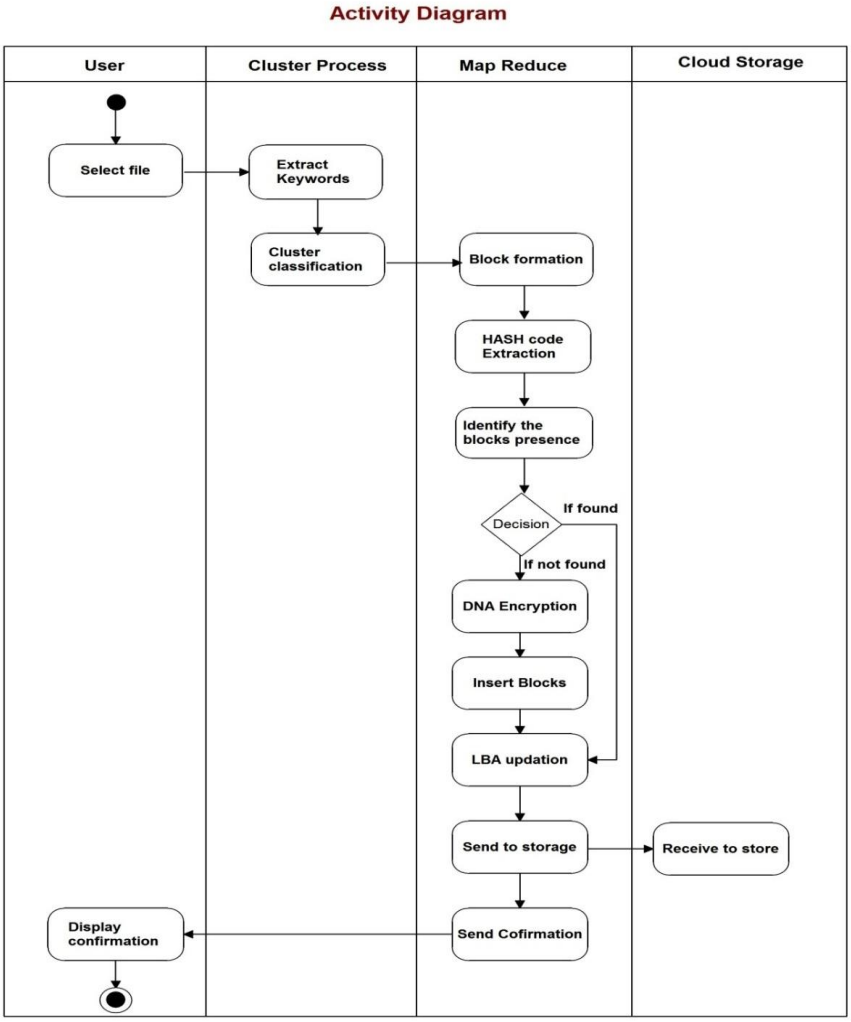

Fig 3: Activity Diagram of Healthcare Data Uploading process

PRACTICAL KNN SEARCH

Step 1: Get the File (F)

Step 2: Extract the keywords with weight age and store it in array K[]

Step 3: Let N be the Number of Classification

Step 4: Initialize an Array Class_Weight[N]

Step 5: Let M be the number of extracted Keywords

Step 6: For I = 1 to M

Step 7 : Let KWord = k[I]

Step 7: For J = 1 to N

Step 8: Check the presence of KWord in Jth Classification Keywords

Step 9: If it present Class_Weight[J] = Class_Weight[J] + KWord Weight

Step 10: Next J

Step 11: Next I

Step 12: Fetch the (Next) highest Class_Weight Value and Index

Step 13: Add Index in Classification Array

Step 13: W = W + Fetched Class_Weight

Step 14: if W >= Threshold then Goto Step

Step 15: Goto Step 12

Step 16: Print all the categories in Classification Array

Step 17: Stop

DNA IMPLEMENTATION

PHASE1: ENCRYPTION OF SECRET DATA

Algorithm for Encryption:

Step one: Convert binary data to DNA sequences.

A=00,

T=01,

C=10, and

G=11.

Step two: Complementary pair rule.

Complementary pair rule is a unique equivalent pair which is assigned to every nucleotides base pair.

Example:

Complementary rule: ((AC) (CG) (GT) (TA))

DNA strand: AATGCT

Applying complementary rule on DNA strand: CCATGA.

Step three: Representing DNA sequences as numeric data.

We extract the index of each couple nucleotides in DNA reference sequence, numerically.

Example:

Assume the reference sequence to be CT1GA2TC3CC4GC5AT6TT7.

Then the numerical representation will be 040602.

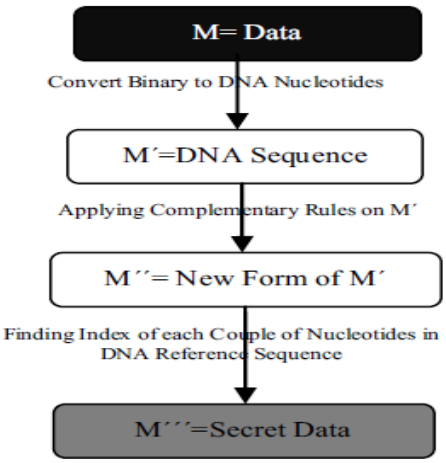

Figure 4: Encryption Process

There is an original data M which the client decides to upload via a network to cloud computing environments. So, there are three sub-phases to provide the final form of M which is M´´´ and upload it to Hadoop. The data M is read as integer and converted to binary form.

In order to convert binary data into amino acids as a DNA sequence, the base pairing rules must be used. Synthesizing nucleotides in real environment (biology) is done in constant rules.

PHASE2: EXTRACTING ORIGINAL DATA

Client2 takes the secret data in form of some numbers. For the purpose of extracting the original data from DNA reference sequence, phase two with its sub phases will extract the original data, correctly depicted in figure 5.

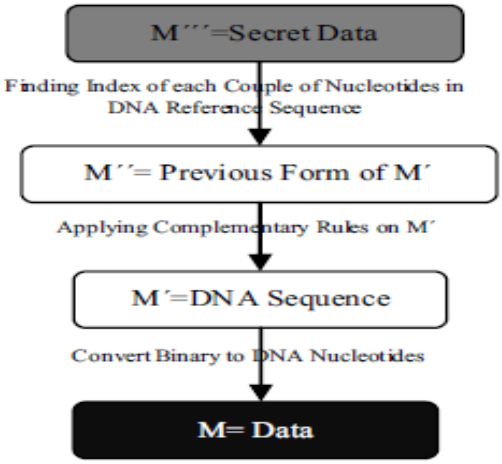

Figure 5: Decryption Process

Algorithm for Decryption:

Step One: Convert numeric data to DNA sequences.

We extract the couple nucleotides in DNA reference sequence according to the index read from the file.

Step two: Complementary pair rule.

Complementary pair rule is a unique equivalent pair which is assigned to every nucleotides base pair.

Step three: Convert DNA sequences to binary data.

DNA PSEUDO CODE

Step 1: Get the Message

Step 2: Covert the String into the Streams

Step 3: Let Consider n be the length of String S1 (e.g. n = 7)

Step 4: Pad the beginning of each with a blank to simplify things (e.g. S1 = "_WRITERS")

Step 5: Fill an initially empty by 0

Step 6: Let M be a Original data Convert binary data to DNA sequences.

Step 7: Let M' = DNA Sequences

A=00,

T=01,

C=10, and G=11.

Step 8: Apply the Base Pairing rule on M'

(A= 00, T= 01, C= 10, G= 11): M´= TAAT

Step 9: Applying complimentary rule M''

((AC) (CG) (GT) (TA)): M´´= ACCA

Step 10: Indexes: M´´´= 0706(Encrypted data)

Decryption Process:

Step 11: Convert numeric data to DNA sequences.

Step 12: M´´´=0706 (Input)

By referring the DNA sequence:

Sub-phase1 (Indexes): M´´= ACCA.

By using Complementary rule:

Sub-phase2 ((AC) (CG) (GT) (TA)):M´= TAAT

By using Base Pair Rule:

Sub-phase3 (A= 00, T= 01, C= 10, G= 11):

M=01000001 (A)(Output)

V. RESULT AND ANALYSIS

Classification	Actual Storage (KB)	Map Reduce Storage (KB)	Map Reduce Storage %	Saved %
Cardiology	520	340	65.38	34.62
Neurology	450	250	55.56	44.44
Dental	636	323	50.79	49.21
Digestive	341	212	62.17	37.83

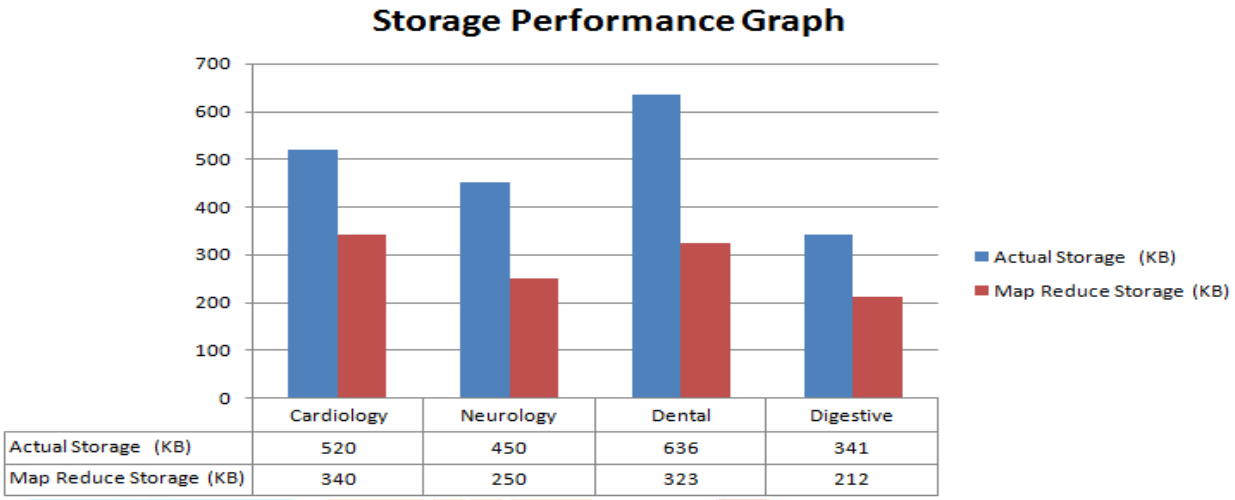

Figure 6: Storage Performance Graph

By considering above set of classifications such as cardiology, neurology, dental and digestive data storage spaces. We are conducting an analysis on storage space occupied in both MapReduce and Actual Storage, where we can observe efficient and optimized storage in the above figure 6.

CONCLUSION

In this work, we proposed a security safeguarding MapReduce based KNN clustering techniques in Hadoop Distributed File System (HDFS). Much appreciated to our light-weight DNA encryption configuration in view of the LWE hard issue, our plan accomplishes clustering velocity and precision that are tantamount to the KNN clustering with privacy protection. Considering the help of extensive scale dataset, we safely coordinated MapReduce structure into our outline; furthermore, make it greatly reasonable for parallelized preparing in distributed computing condition. Further to Map Reduce technique, the privacy preserving DNA encryption technique merged with KNN clustering improve the system performance to next level. We give exhaustive examination to demonstrate the security and effectiveness of our plan. Our model execution on many healthcare domain data and result shows that our plan is productive, versatile, and precise for extensive scale dataset.

REFERENCES

[1] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Siberschatz, and A. Rasin, "HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads," Proc. VLDB Endowment, vol. 2, no. 1, pp. 922–933, Apr. 2009.
[2] R. Choquet, M. Maaroufi, A. de Carrara, C. Messiaen, E. Luigi, and P. Landais, "A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research," J. Amer. Med. Informat. Assoc., vol. 22, no. 1, pp. 76–85, Jul. 2014.
[3] C. G. Chute, S. A. Beck, T. B. Fisk, and D. N. Mohr, "The enterprise data trust at Mayo Clinic: A semantically integrated warehouse of biomedical data," J. Amer. Medical Informat. Assoc., vol. 17, no. 2, pp. 131–135, Mar.-Apr. 2010.
[4] R. H. Dolin, B. Rogers, and C. Jaffe, "Health level seven interoperability strategy: Big data, incrementally structured," Methods Infor. Med., vol. 54, no. 1, pp. 75–82, Dec. 2015.
[5] K.N. Eggleston et al., "The net value of health care for patients with type 2 diabetes, 1997 to 2005," Ann. Internal Med., vol. 151, no. 6, pp. 386–393, Sep. 2009.
[6] M. Kimura et al., "High speed clinical data retrieval system with event time sequence feature: With 10 Years of clinical data of Hamamatsu University Hospital CPOE," Methods Inf. Med., vol. 47, no. 6, pp. 560–568, Nov. 2008.
[7] C. N. Mead, "Data interchange standards in healthcare IT–computable semantic interoperability: Now possible but still difficult, do we really need a better mousetrap?" J. Healthcare Inf. Manage., vol. 20, no. 1, pp. 71–78, Jan. 2006.
[8] M. J. Minn, A. R. Zandieh, and R. W. Filice, "Improving radiology report quality by rapidly notifying radiologist of report errors," J. Digit. Imag., vol. 24, pp. 492–498, 2015. [Online]. Available: http://www. ncbi.nlm.nih.gov/pubmed/25694167.
[9] T. Namli, G. Aluc, and A. Dogac, "An interoperability test framework for HL7-based systems," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 3, pp. 389–399, May 2009.

[10] A. Nguyen, J. Moore, G. Zuccon, M. Lawley, and S. Colquist, "Classification of pathology reports for cancer registry notifications," Studies Health Technol. Informat., vol. 178, pp. 150–156, Jul. 2012.
[11] F. Oemig and B. Blobel, "Semantic interoperability adheres to proper models and code systems. A detailed examination of different approaches for score systems," Methods Inf. Med., vol. 49, no. 2, pp. 148–155, Feb. 2010.
[12] D. Rajeev et al., "Development of an electronic public health case report using HL7 v2.5 to meet public health needs," J. Amer. Med. Informat. Assoc., vol. 17, no. 1, pp. 34–41, Jan. /Feb. 2010.
[13] P. Roberts, "Total teamwork–the Mayo Clinic," Radiol. Manage., vol. 21, no. 4, pp. 29–30, 32–36, Jul./Aug. 1999.
[14] J. Sayyad Shirabad, S.Wilk,W.Michalowski, and K. Farion, "Implementing an integrative multi-agent clinical decision support system with open source software," J. Med. Syst., vol. 36, no. 1, pp. 123–137, Dec. 2012.
[15] G. Schrijvers, A. vanHoorn, and N.Huiskes, "The care pathway: Concepts and theories: An introduction," Int. J. Integr. Care, vol. 12, Special Edition Integrated Care Pathways, pp. 1–7, Sep. 2012.
[16] W. Sujansky and T. Wilson, "DIRECT secure messaging as a common transport layer for reporting structured and unstructured lab results to outpatient providers," J. Biomed. Informat. vol. 54, no. 1, pp. 191–201, Apr. 2015.