Enhancement of Healthcare Outcomes using Big Data Analytics

G.JayaLakshmi
Assistant Professor,VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

A.Srisaila
Assistant Professor,VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

P.Madhavi Latha
Assistant Professor, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

Abstract - Big have increased colossal consideration lately. Dividing enormous information is exceptionally regular prerequisite today and such necessities get to be bad dream when examining of mass information source, it is truly a major test to investigate the mass measure to get significance and distinctive examples of data on positive way. In this paper we discover knowledge from health information used in health care organizations using a new information management approach called as big data analytics to advance personalized care, improve patient outcomes and avoid unnecessary costs. Associations can store substantial datasets in Hadoop Distributed File Systems (HDFS) and utilize continuous investigation programming based on top of design such as spark to access information straightforwardly from HDFS, bypassing any information relocation logical efforts.

Keywords – Big data , Healthcare, Analytics, SPARK

I. INTRODUCTION

Essentially, big data is helping organizations become more creative, efficient and reduce costs. Like many other industries, healthcare has improved to data analytics not only for its financial returns but also for refining patient's quality of life. Big data is creating a considerable measure of hype in each industry including Healthcare. The healthcare industry historically has generated huge data, driven by record keeping, compliance & regulatory requirements, and patient care [1]. Data alone is insufficient, it's about how data is broke down to drive more quick choices about intervention and treatment alternatives.

Driven by mandatory requirements and the potential to improve the quality of healthcare delivery and reduce costs, this supports a range of medical and healthcare decisions, like clinical decision support, disease surveillance, and population health management, among others [2]. One of the guarantees of the developing minimum amount of clinical data collecting in electronic HEALTH record (EHR) frameworks is auxiliary use (or re-use) of the data for different purposes, for example, quality change and clinical examination.

Locating access and applying clinical and progressed analytics to this profitable new data empowers associations to enhance knowledge into risk, effects, resources, referrals, execution and readmissions, and to make narrow move. Uniting organized and unstructured data strengthens more perceptive investigation that empowers customized and evidence-based medicine, more proficient procedures and motivating forces that can improve patient behavior.

By definition, huge data in medical services does not allude to electronic healthcare data sets so extensive and complex that they are troublesome to make do with conventional programming and/or equipment; nor would they be able to be effortlessly made do with customary or basic data administration devices and systems [3]. The big data that healthcare organizations need to collect and analyze may come from hospitals, ambulatory care facilities, wellness centers, referral networks, labs and imaging centers, research and other nontraditional data sources. Collecting, integrating and analyzing data can be a complex task because the data resides in many internal and external locations and its level of quality may be unknown. In addition, about 80 percent of medical data is unstructured, which further increases the challenge

The wording encompassing the utilization of extensive and changed sorts of data in human services is developing, yet the term examination is accomplishing wide utilize both all through healthcare. The huge data that social insurance associations need to gather also, investigate might originate from doctor's facilities, wandering consideration offices, wellbeing focuses, referral systems, labs and imaging focuses, research and other nontraditional data sources[4]. Gathering, incorporating and investigating data can be a complex assignment in light of the fact that the data lives in numerous interior and outside areas and its level of value might be obscure. What's more, around 80 percent of medicinal data is unstructured, which advance builds the test.

There are various terms identified with data analysis. A center procedure in data analysis is machine learning, which is the territory of software engineering that intends to manufacture frameworks and calculations that gain from data. One of the significant methods of machine learning is data mining, which is characterized as the handling and demonstrating of a lot of data to find already obscure examples or relationships. A sub area of data mining is content mining, which applies data mining strategies to for the most part unstructured literary data [5].

A. Big Data Analytics in Healthcare:

Big Data offers great opportunity to better insights like patients with different risks can be identified and vaccinated within hours of the outbreak, patients' blood pressure and sugar levels can be spiked, can reduce readmission rates by assigning a care to heart patients with no emergency. Get past straightforward relationships in light of finding codes and SNPs. Examination stage to break down extensive data sets of ideas versus ideas, for example, lab results, genotypes, drugs, finding codes, phenotypes.

Huge data suggests huge volumes of data. It used to be representatives made data. Since data is produced by machines, systems and human association on frameworks such as online networking the volume of data to be broke down is outrageous.

Assortment alludes to the numerous sources and sorts of data both organized and unstructured. We used to store data from sources such as spreadsheets and databases. Enormous Data Velocity manages the pace at which data streams in from sources such as business procedures, machines, systems and human collaboration with things such as online networking destinations, cell phones, and so on.

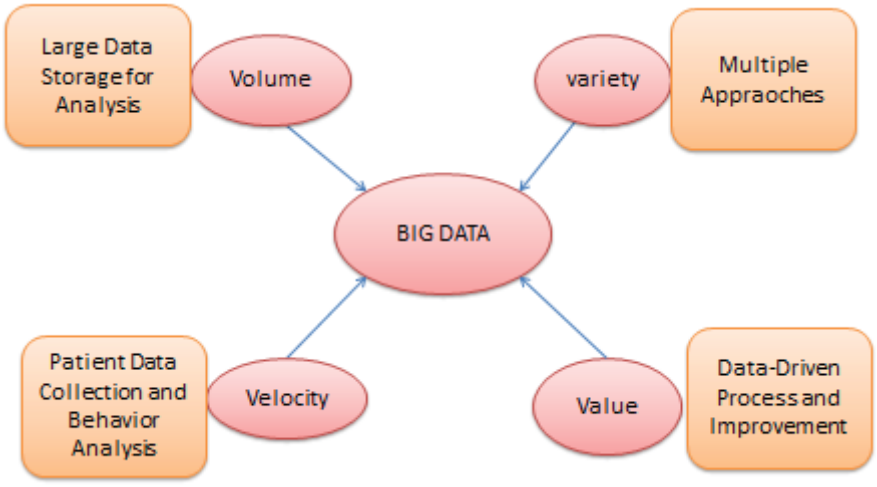

Figure 1. Big data Analytics for Health Care Data

Huge information in human services can originate from inside (e.g., electronic healthcare records, clinical choice emotionally supportive networks, CPOE, and so forth.) and outside sources (government sources, research facilities, drug stores, insurance agencies and HMOs, and so forth.), frequently in various organizations (level records, .csv, social tables, ASCII/content, and so forth.) and living at different areas (geographic and in addition in various medicinal services suppliers' destinations) in various legacy and different applications (exchange handling applications, databases, and so forth.).

II. RELATED WORK

Medicinal services IT frameworks can enhance the productivity and nature of consideration, by imparting and coordinating patient information crosswise over various offices and institutions, while holding security controls. The expanding digitization of human services data is opening new conceivable outcomes for suppliers and payers to enhance the nature of consideration, medicinal services comes about, and minimize the expenses. The most recent devices and innovations are utilized on computerized data of medicinal services associations can create profitable bits of knowledge. Associations should likewise divide interior and outside patient data to all the more precisely measure hazard and results. In the meantime, numerous suppliers and payers are attempting to build information straightforwardness to create new understanding learning.

KiyanaZolfaghar et al. [6] has presented prediction model in which training data is given to the Hadoop record framework, basic information is preprocessed and changed over to classifiable information which will be in encoded group called as a vector that will be given as data to the Mahout structure, utilizing arbitrary forest algorithm. Joseph M. Woodside [7] has presented, inefficient vendors can be identified, and who is poor in the member's routine life decisions and amenableness with preventative care programs.

Medical images are a significant source of data often used for diagnosis, therapy assessment and planning [8]. Such data requires large storage volumes if stored for long term. It also demands fast and accurate algorithms if any result assisting automation were to be performed using the data. From a data point of view, medical images might have 2, 3, and four dimensions. Positron emission tomography (PET), CT, 3D ultrasound, and functional MRI (fMRI) are considered as multidimensional medical data. Modern medical image technologies can produce high-resolution images such as respiration-correlated or “four-dimensional” computed tomography (4D CT) [9].

Thommandram et al.[10] used data streams with a different goal attempting to detect and classify neonatal cardio respiratory spells and system is called Artemis used a steady stream of physiological data from the new born patient and both detect and ascertain which type of cardio respiratory spell the patient is experiencing all in real-time.

Ashish thutoo[13] et al. created a platform called Smart Health Informatics Program (SHIP) with the goal of helping patients connect to the medical experiences of other patient sported throughout the internet via message boards (i.e. forums, blogs etc.). This study used a pool of 50,000 discussions including over 400,000 posts from four different message board websites (inspire.com, medhelp.com, and 2 others).

Rolia et al.[14] devise a new system to use social health forums to help patients learn about their condition from posts by other patients with similar conditions. Their method consists of three steps:

1. Determining the patient’s current medical condition from the personal health record (PHR).
2. The system will ascertain which other users have a similar condition.
3. A metric will be implemented evaluating and ranking the forum topics to determine the most relevant to present to the user. They describe their system implementation for a test case of type II diabetes mellitus (DM II).

A Healthcare through Big Data Technologies:

Changing Health Care through Big Data includes the commitments of people from health system, health information technology, academic, and health policy domains. Healthcare Providers historically used data warehouses and business intelligence tools to report on and analyze financial results, optimize operation of their facilities, and measure outcomes and quality of care.

- Improving patient consequence and illness prevention
- Improving Efficiency and Managing Costs
- Proactive and better diagnoses faster with less errors

- Enable better use of facilities and equipment
- Effective processing and leveraging of IT resources

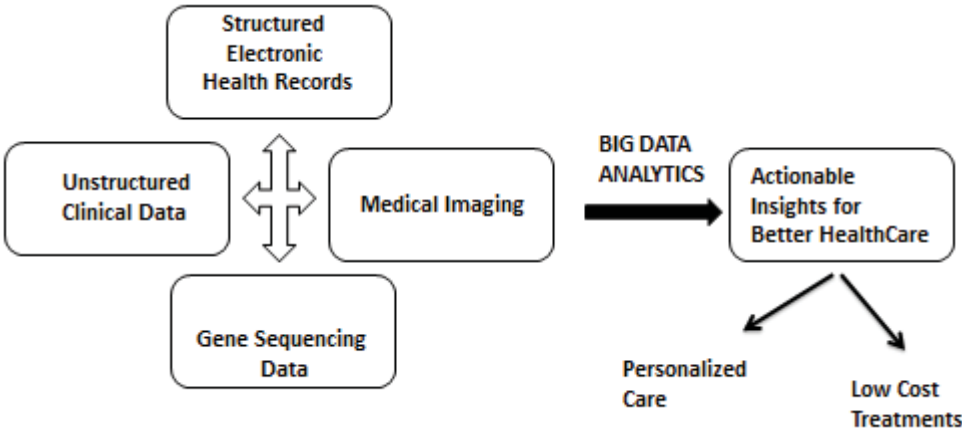

Figure 2. Processing of data through Big Data Analytics for Healthcare Data

B Data Collection:

- Electronic Medical Records (EMRs)
- Transcriptions
- PACS
- Medication Administration
- Financial
- Laboratory

With the end goal of huge information investigation, this information needs to be shared. In the second segment the information is in a "crude" state and should be handled or changed, at which point a few choices are accessible. The information stays crude and administrations are utilized to call, recover and prepare the information. Another methodology is information warehousing wherein information from different sources is collected and made prepared for preparing, despite the fact that the information is not accessible in real-time. Depending on whether the information is organized or unstructured, a few information configurations can be info to the huge information examination stage. In this next part in the reasonable structure, a few choices are made with respect to the information data approach, appropriated outline, apparatus determination and investigation models.

III. PROPOSED ALGORITHM

Spark can be used for processing larger open data from organizations that publish millions of records yearly and update these monthly or weekly. In order to utilize this open information and inference data, they must be handled through different steps like downloading, extraction, cleaning, and transforming.
Here Spark optimizes the steps in processing workflows and support lazy evaluation of healthcare data queries. As Spark holds intermediate results in memory, it is easy to work on same dataset multiple times.

The conventional scripts won't work with such a large amount of information, it will as a rule takes all the more time to complete the process of handling and redesign information back to mysql. We require an appropriated way to deal with procedure information, so taking after the enormous information idea; we utilize Spark for handling these records.

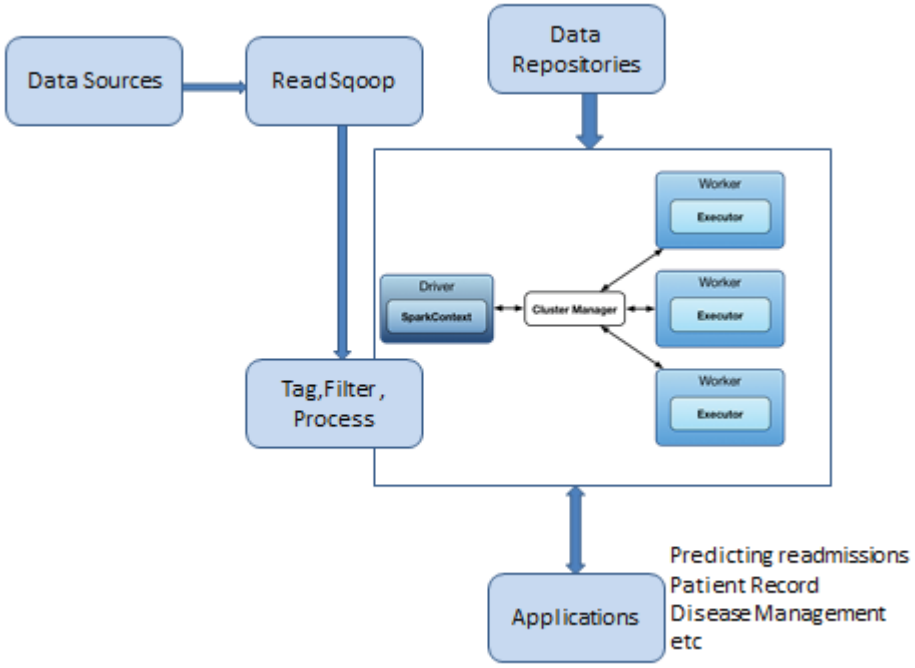

Fig 3. Predicting readmission Patient Record Disease Management

Algorithm: Enhancement of Healthcare outcomes Using Big Data Analytics
Step 1: Structure spark standalone cluster.

Use spark-ec2 script to create a 10 node spark cluster.

Step 2: Load data in spark cluster

Use Sqoop to load data from MySql into HDFS of spark standalone cluster.

Step 3: Create a pattern config file (using json)

{

"affiliation_tag": {

"NYU": {

"AFFILIATION_PATTERNS": "Hospital1",

"DEPARTMENT_LIST": [

"Division of Cardiac Surgery",

"Division of Pediatric and Adult Congenital Cardiac Surgery",

"Department of Surgery",

"Acute Care Surgery",

"Bariatric Surgery",

]

},

"UPEN": {

"AFFILIATION_PATTERNS": " Hospital2",

"DEPARTMENT_LIST": [

"Genetics",

"Medical Ethics and Health Policy",

"Dermatology",

"Emergency Medicine",

]

},

"RIC": {

"AFFILIATION_PATTERNS": " Hospital3",

"DEPARTMENT_LIST": [

"Cancer Rehabilitation",

"Back Injuries and Back Pain",

]

},..}

Step 4: Spark application which runs on spark cluster.
Subsequent to stacking the information to HDFS through step 2, we will prepare the information through flash application written in python. We have to setup the record way by giving HDFS url of document. The aftereffects of handling are upgraded back to mysql through the same flash script. For string matching fuzzy string matching functions were used, which requires to be submitted to cluster and also set cluster url in application using set Master option in spark context, then submit the application to cluster.
$ spark-submit aff_tagger.py

<pre>#aff_tagger.py

```
classAffiliationFilter():
   """docstring for AffiliationFilter"""
def __init__(self, **kwargs):
self.source = kwargs.get("source", None)
with open("affiliation_tagging_map.py") as json_data:
self.rmap       = simplejson.load(json_data)
self.enrich_map  =self.rmap['affiliation_tag']#.values()#[self.source.upper()]

deffind_match(self, jsonl):
       """ find the match using pattern in config file"""
match = True
foraff_canon in self.enrich_map.keys():
aff_search_pattern=  self.enrich_map[aff_canon]['AFFILIATION_PATTERNS']
             p = re.search(aff_search_pattern,jsonl['AFFILIATION'],re.IGNORECASE)
```

```
if match and p:
jsonl.update({"AFFILIATION_CANON":aff_canon})
aff_dept_list = self.enrich_map[aff_canon]['DEPARTMENT_LIST']
                t = {}
fordept in aff_dept_list:
ifdept in jsonl["AFFILIATION"]:
t["AFFILIATION_DEPT"] = dept
elifTokenMatch().site_name_match(dept,jsonl["AFFILIATION"])[0] &amp;gt; 58:
t["AFFILIATION_DEPT"] = dept
if not t:
t["AFFILIATION_DEPT"] = ""
jsonl.update(t)
match = False
if match == False:
returnjsonl
if match:
jsonl.update({"AFFILIATION_CANON": "UNTAGGED","AFFILIATION_DEPT": ""})
returnjsonl

defconvert_time(self,string,format = "%Y-%m-%d"):
    """converts to Y-m-d format"""
returnstrftime("%Y-%m-%d %H:%M:00",strptime(string, format))

defget_month_from_date(self,date_obj):
returndate_obj.strftime('%B')

defget_quarter(self, month):
    """ gets quarter from month"""
returnmath.ceil(month/3)

defjsonize(self, data):
data =  json.loads(data)
return data

defget_affiliation(self, jsonl):
    # if "new york university" in jsonl["AFFILIATION"].lower() or "nyu" in jsonl["AFFILIATION"].lower():
       # printjsonl["YEAR"],jsonl["AFFILIATION"]
returnjsonl

defupdate_mysql(self, iterator):
    """ updates the mysql tuple back with the tag of institute"""
conn     =     MySQLdb.connect(host=SELECT_STAR_DB_SERVER_IP,     port=SELECT_STAR_DB_PORT,
user=SELECT_STAR_DB_USER, passwd=SELECT_STAR_DB_PASSWORD, db="select_star",
charset='utf8', use_unicode=True)
conn.autocommit(True)
cur = conn.cursor(cursorclass=MySQLdb.cursors.DictCursor)
print "partitions...",
    f = open("/root/mysqlf.txt","w+")
    k = 0
for i, jsonl in enumerate(iterator):
sql = "update select_star_core_hcp set  aff_tag = '%s' where record_id = '%s';"
sql = sql %(jsonl["AFFILIATION_CANON"],jsonl["RECORD_ID"])
f.write(sql+ "\n")
printcur.execute(sql)
        sql1 = """SELECT pre_enriched from raw_records WHERE record_id = '%s';""" %jsonl["RECORD_ID"]
f.write(sql1+ "\n")
```

```
      r = cur.execute(sql1)
      r = cur.fetchall()
iflen(r) &amp;gt; 0:
if r[0]["pre_enriched"]:
ujsonl = simplejson.loads(r[0]["pre_enriched"])
ujsonl["AFF_TAG"] = jsonl["AFFILIATION_CANON"]
ujsonl["AFFILIATION_DEPT"] = jsonl["AFFILIATION_DEPT"]
query = """UPDATE  raw_records SET pre_enriched = %s WHERE  record_id = %s;"""
cur.execute(query,[simplejson.dumps(ujsonl),jsonl["RECORD_ID"]])
else:
        k+=1
print "no of row in this partition %s %s" %(i,k)
conn.close()

def f(self, iterator):
print "partition"
for x in iterator:
break

deftuple_json(self,tupl):
tupl = tupl.split(',')
return {"RECORD_ID":tupl[0],"AFFILIATION":','.join(tupl[1:])}

if __name__ == '__main__':
conf = SparkConf().setAppName(
  "select_star_filter_affilition_v1").setMaster(
  "spark://ec2-54-167-xx-xx6.compute-1.indianhealth.com:7077")
  # .set("spark.speculation","true")
  # sc = SparkContext(conf=conf)
sc =
SparkContext(conf=conf,pyFiles=["/root/metrics/token_matcher.py","/root/metrics/normalize.py","/root/metrics/sett
ings.py","/root/metrics/token_constants.py"])
path = "hdfs://ec2-54-167-xx-xx.compute-1.indianhealth.com:9000/selectaff"
  A = AffiliationFilter()
select_star_RecordRDD = sc.textFile(path).repartition(40)
select_star_RecordRDD = select_star_RecordRDD.filter(lambda l: l).map(A.tuple_json).filter(lambda l:
l.get("AFFILIATION",None))
filtered_rdd = select_star_RecordRDD.map(A.find_match)
col = filtered_rdd.filter(lambda l: l["AFFILIATION_CANON"] not in ["NYU","UPEN"]) # in
["NYU","UPEN","RIC","MFMER"]
  # col.saveAsTextFile("hdfs://ec2-54-147-63-178.compute-1.indianhealth.com:9000/ric.txt")
print "%s " %("*"*500)
  #todo use foreach on each rdd...
  #filtered_rdd.foreach(a.update_mysql)
printcol.foreachPartition(A.update_mysql)
```

Healthcare data can preprocessed using Apache spark and assign affiliation to the organization records, based on string matching. The results indicate that provide the mechanism needed to generate more meaningful knowledge that can positively impact patient outcomes. The results illustrates and analysis the data collected for the experiment purpose by Apache Spark using healthcare database .If there is any change in pattern, then the hospital wanted an alert to be generated to a team of doctors and assistants. This provides the best clinical support, reduce the cost of care measurement and manage the population of at-risk patients.

IV. Analysis and Usability Study

The tool enables health researchers or other interested users to access health data and data mining tools through available interfaces on internet. The initial implementation on health problems exploration allows users to identify risk factors and high risk groups and allows a search for possible identifications without prior knowledge.The results of the analysis can then be visualized using various forms of knowledge representation methods. This tool and the embedded data mining tools can be used in broad areas of health data analysis.

V. CONCLUSION

Big Data is collection of data elements whose size, speed, complexity require to adopt software and hardware mechanisms to successfully store, analyze and visualize data. Healthcare is an important example to show how three Vs of data, velocity, veracity and volume are vital aspect of the data it produces. The data spread among multiple healthcare centers should provide a platform for global data transparency. Researchers are reviewing the complexity in healthcare data in terms of both characteristics of the data itself and the taxonomy of analytics that can be expressively performed on them. The goal of using Spark in Healthcare is to collect and analyze data from public health trends in a region of people to identify treatment options for one patient. And this helps most patient care systems working with unstructured data for analysis.

REFERENCES

[1] Raghupathi W, "Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity", Edited by Kudyba S. Taylor & Francis; 2010:211-223.

[2] Dembosky A, "Data Prescription for Better Healthcare", Financial Times, December12,2012,p.19.2012.

[3] Frost & Sullivan, "Drowning in Big Data, Reducing Information Technology Complexities and Costs for Healthcare Organizations".

[4] Big Data and Healthcare Payers.2013.

[5] Cohen AM and Hersh WR, "A survey of current work in biomedical text mining", Briefings in Bioinformatics, 2005. 6: 57-71.

[6] KiyanaZolfaghar, NarenMeadem, Ankurteredesai, SenjutiBasu Roy, Si-Chi Chin.Big, "Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients",2013 IEEE International Conference on Big Data, 978-1-4799-1293-3/13, http://dx.doi.org/10.1109/bigdata.2013.6691760

[7] Joseph M. Woodside, "Virtual Health Management", 2014, 11th International Conference on Information Technology: New Generations 978-1-4799-3187-3/14. http://dx.doi.org/10.1109/itng.2014.124

[8] F. Ritter, T. Boskamp, A. Homeyer et al., "Medical image analysis", IEEE Pulse, vol. 2, no. 6, pp. 60–70, 2011

[9] K. Bernatowicz, P. Keall, P. Mishra, A. Knopf, A. Lomax, and J. Kipritidis, "Quantifying the impact of respiratory-gated 4D CT acquisition on thoracic image quality: a digital phantom study",Medical Physics, vol. 42, no. 1, pp. 324–334, 2015.

[10] Thommandram A, Pugh JE, Eklund JM, McGregor C, James AG (2013), "Classifying neonatal spells using real-time temporal analysis of physiological data streams", Algorithm development In: IEEE Point-of-Care Healthcare Technologies(PHT 2013). IEEE, based in New York, USA, Bangalore, India, pp 240–243.

[11] Zhang Y, Fong S, Fiaidhi J, Mohammed S (2012), "Real-time clinical decision support system with data stream mining", J Biomed Biotechnol 2012: 8. [http://dx.doi.org/10.1155/2012/580186]

[12] Campbell AJ, Cook JA, Adey G, Cuthbertson BH (2008), "Predicting death and readmission after intensive caredischarge", British J Anaesth 100(5): 656–662.

[13] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy , "Hive - A Warehousing Solution Over a Map-Reduce Framework",VLDB ,09,August 24-28, Lyon, France, , 2009 .

[14] Rolia J, Yao W, Basu S, Lee WN, Singhal S, Kumar A,Sabella S (2013) Tell me what i don't know -making the most of social health forums. Tech. Rep: HPL-2013–43. Hewlett Packard Labs ,https://www.hpl.hp.com/techreports/2013/HPL-2013-43.pdf