A Study of Medical Big data mining using Lambda Architecture

R. Meena, H. Vidhya,
Assistant Professor, Prathyusha Institute of technology and Management,
Tiruvallur, Tamilnadu, India

Abstract--Business intelligence and analytics has become more important ever. The data evolves into exabytes day by day. Big data gives the solution to two main challenges of data such as storage and analysis. Lambda architecture suggested for storing the information which is been processed and they can be an ideal way to minimize the processing. The paper suggests that medical big data can be stored and analysed using hadoop can use lambda architecture to store the results so that it can be fetched. The sample medical dataset has been used and an example analysis was shown to store the results of data.

I .INTRODUCTION

Day by day we are generating exabytes of data through various forms. The usage of computers has been exponentially increased through a decade. Internet usage through laptops and other devices gives lot of digital records. The data generated by all our digital interactions to online shopping, text messages to tweets, Facebook updates creates unstructured data which traditional systems cannot handle. We generate 12TB tweets every day and more than 2 billion people on the net makes data. Using data, intelligence can be collect from it, and that entails effective data analysis. Big data gives the solution to two main challenges of data such as storage and analysis. The data can be distributed and stored in a massively parallel processing system and analysis will be done. This can be applied in many fields such as research, agriculture, logistics, urban design, energy, retailing, medical data mining, etc.., complex volumes of unstructured data are generated by healthcare activities, such as patient prescription, test reports, CT and MRI images, nerve and muscle biopsy reports, skeletal radiographs, ECG, etc.., Mining and intensive analysis on medical data can bring potential results which will assist in genetic disorder prediction, and drug discovery. Map Reduce by itself is capable for analysing large distributed data sets; but due to the heterogeneity, velocity and volume of Big Data, it is a challenge for traditional data analysis and management tools.[2] The lambda architecture introduced by Marz [4] is an interesting proposal to the latency challenges in real-time stream processing. The remainder of the paper is structured as follows: In Section 2, we discuss on lambda architecture to model on medical big data. The representation of frequent instances on medical dataset has been suggested using the sample dataset with the recommendation of lambda architecture in section 3, the conclusion and future enhancement of the paper in Section 4.

II. LAMBDA ARCHITECTURE ON MEDICAL BIGDATA

The main framework or tool which has been highly recommended for big data is hadoop. Hadoop has its storage on hadoop distributed file system [HDFS]. Hadoop has the component such as HDFS and Map reduce where HDFS access for data storage and map reduce for data analysis. Components of HDFS are name node, data node and secondary name node. The processing in hadoop follows the principle write once read many, where we can store petabytes of data and analyse it. For processing data, lambda architecture can be implemented in hadoop where we can store the results in a serving layer and can retrieve when needed and this prevents terabytes of data to be processed repeatedly. Three layers such as batch layer, service layer and speed layer has been suggested for storing the information which is been processed and they can be an ideal way to minimize the processing. The layers have defined operations to use. Serving layer indexes batch views so that they can be queried in ad hoc with low latency. Batch layer can managing the master dataset, an immutable, append-only set of raw data – pre-computing arbitrary query functions, called batch views. Speed layer accommodates all requests that are subject to low latency requirements. This provides views of processed data.

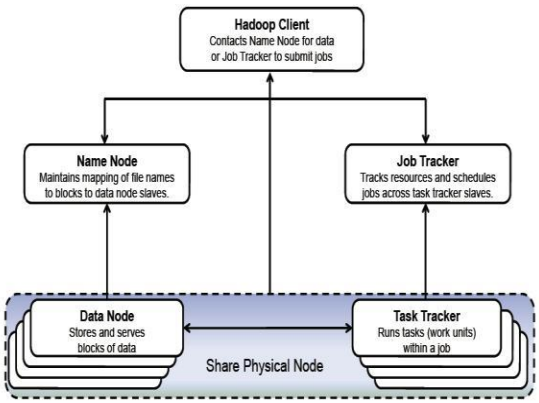


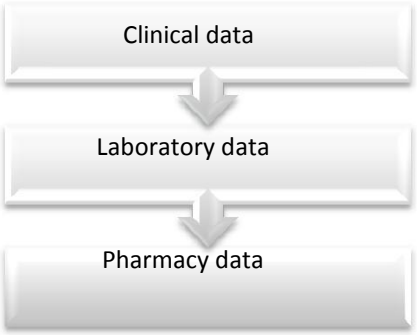
Fig 1.Hadoop architecture

Structured and unstructured e-health records, medical imaging genomic data has been the major sources. Data acquired from the medical data set can be studied deeply but the lack of seamlessly integrated, productized, and scalable infrastructure to support data management, analysis, and reporting is a limitation. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity. For a DNA or genomic-related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data.[1] Variety of electronic patient reports may have the valuable veiled facts which can be reported if proper inquiry is done. For instance, an integrated analysis of cross-modality glioblastoms (GBM) data, including DNA copy number, gene expression, and DNA methylation aberrations, helped dissect genome-wide regulatory mechanisms for further investigation into the identification of candidate biomarkers for GBM tumors and potential therapeutic targets.[8]

Recent studies revealed that every dollar we invested to map the human genome returned $140 to our economy. Hadoop integrated with lambda architecture can be a suitable solution to

- Have a regulated support and standardisation to analyse data
- Build a promising system to collect, regulate and analyse clinical data
- Perform storage on frequently mined patterns without repeated processing
- Diagnose variant genomic data such as gene and DNA sequencing for patient

Meaningful use of data standards in particular have already simplified the task of identifying clinical phenotyping patterns in electronic health records.



ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization. There can be similar data acquire from information extraction, feature selection and prediction may result in similar patient records and also seasonal diseases can have similar results. These can be stored in the serving layer which gives out potential, relevant attributes of result. Patient similarity learns a customized distance metric for a specific clinical context.

III. SAMPLE DATASET

ICD9

dd: 2768
ds: 4

ld: HYPOPOTASSEMIA
dd: 3051
ds: 6

ld: TOBACCO USE DISORDER
dd: 4010
ds: 1

ld: MALIGNANT HYPERTENSION

dd: 4254 ds: 3

ld: PRIM CARDIOMYOPATHY NEC

dd: 5819 ds: 2

ld: NEPHROTIC SYNDROME NOS

dd: 6826 ds: 5

ld: CELLULITIS OF LEG dd: 7823
ds: 4
ld: HYPOPOTASSEMIA

Solution

id: D5W el: 1

cu: CCU cg: -1

io: D5W 100.0ml IV Infusions vo: 100
du: ml
rt: Intravenous Push

id: Po Intake el: 1

cu: CCU cg: -1

io: Po Intake PO/Gastric du: ml

rt: Oral

Additive

id: Heparin el: 100

cu: CCU cg: RN

io: D5W 250.0ml + 25000Uhr Heparin IV Infusions am: 25000

du: Uhr rt: IV Drip

id: Nitroglycerine el: 100

cu: CCU cg: RN

io: D5W 250.0ml + 100mcgmin Nitroglycerine IV Infusions am: 100

du: mcgmin rt: IV Drip

The represented dataset indicates the log of data which is a clinical dataset. The problems with id numbers and the suggestions given have been collected and this grows into gigabytes when collected across years. The data has been semi structured since it has numbers and text formats. Using hadoop we can analyse the number of patients having a particular complaint and the results.

Example 1: – find the ds count for the particular day

Map Output: (4,1),(6,1),(5,1),(2,1),(4,1)

Reduce: (4,(1,1),(6,1),(5,1),(2,1))

(2,1),(4,2),(5,1),(6,1)

These simple analysis can be made and using which we can do deeper studies. The investigation can be made again and again if it has been queried and the results will be generated. We suggest that the way of placing the data somewhere and fetch it. The architecture will store this in the layer (service) and as requested the data will be fetched.

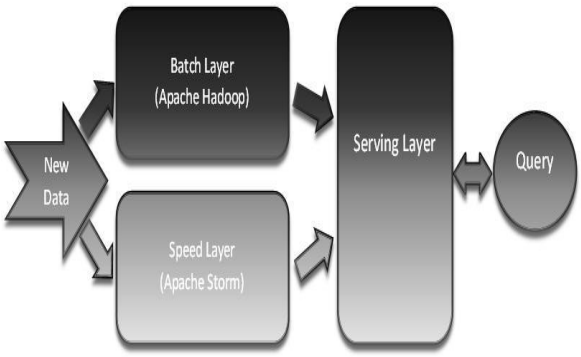


Fig 2.Illustration of lambda architecture

IV.FUTURE WORK AND CONCLUSION

Big data analytics is a promising right direction which is in its infancy for the healthcare domain. Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs. Developing a coordinated approach using big data and lambda architecture will be the right research way from where the healthcare can be improved further.

REFERENCES

[1] Data Mining with Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

[2] Jefry Dean and Sanjay Ghemwat,.MapReduce: Simplified data processing on large clusters, Communications of the ACM, Volume 51 pp. 107–113, 2008

[3] Investigating the Lambda Architecture Nicolas Bar of Zurich ZH, Switzerland University at Zurich

[4] Marz, N. (2013). Big Data: Principles and best practices of scalable realtime data systems. O'Reilly Media.

[5] Oracle Health Sciences Translational Research Center: A Translational Medicine Platform to Address the Big Data Challenge

[6] physiobank/database/mimic2cdb/ s25222/

[7] Big data: A review, Collaboration Technologies and Systems, IEEE, 978-1-4673-6403-4

[8] Comprehensive genomic Characterization defines humanglioblastoma genes and core pathways. Nature, 2008. 455(7216): p. 1061–1068.