ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

A Service Based Data Segregation & Extraction Using Big Data In E-Health Insurance

[1]R.Jayaraj and [2]N.S.Badrinath

[1]Agni college of Technology, Anna University, Computer Science and Engineering, S.Yoganand, Chennai. India
[2]Agni College of Technology, Anna University, Computer Science and Engineering, S.Yoganand, Chennai. India

ARTICLE INFO

Article history:
Received 10 March 2015
Received 20 in revised form
March 2015
Accepted 25 March 2015
Available online 10 April 2015

Keywords:
Big data ,e-health, hadoop hdfs, map reduce

ABSTRACT

The fast development of internet application is boosting the development of cloud computing. In this agile world, the usage and storage of data increases day by day and collectively called as Big Data. To handle those Big Data we move to a Distributed File System approach which we term as Hadoop Distributed File System provided by an open source vendor. Hadoop, also an open-source implementation of Map Reduce, which is a suitable tool to deal parallel with these kinds of applications. Big data technology cannot be applied to e-Health service directly and it require additional capabilities. Analysis of Discharge Summary, Drug & Pharma, Diagnostics Details, Doctors Report, Medical History, Allergies & Insurance policies are made and Useful Data is Extracted. What we going to do is that we are also adding e-insurance So that the person can see the diagnosis, availability of hospital and doctors and claim all his medical records as one..

To Cite This Article: R. Jayaraj, N.S.Badrinath,S.Yoganand, A Service Based Data Segregation & Extraction Using Big Data In E-Health Insurance ***Aust. J. Basic & Appl. Sci.,*** *9(15): 70-75, 2015*

INTRODUCTION

With the rapid development of Internet, we are experiencing an information explosion era. Large amounts of data are being stored and managed. Intel corporation has estimated that the world generates one petabyte (1,024 terabytes) of data every 11 seconds, the equivalent of thirteen years of high definition video. IDC estimated that in 2011 all the data created in the world amounted to 1.6 trillion gigabytes(1.56 billion terabytes).By 2020,50 billion devices will be connected to the networks and the internet. Bharati Airtel Ltd, handles around 8billion calls every day, generating petabytes of data to be analysed for identifying new revenue opportunities.

Big Data:

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. The definition of big data as the three V's: volume, velocity and variety. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible. When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Large volume and velocity of data leaves with only option.i.e. distributed computing. Hadoop's HDFS is used for solving Big Data problem . Hadoop is a kind of distributed computing platform to store and process petabytes of data.

Hadoop:

Hadoop, the open-source software for processing Big Data. Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day. Hadoop can handle all types of data from disparate systems:

Corresponding Author: R.Jayaraj, Agni College Of Technology Computer Science and Engineering, S.Yoganand, Chennai. India.
Phone (+919884082950, +918012377333,9994763389) jayaraj1805@gmail.com

structured, unstructured, log files, pictures, audio files, communications records, email– just about anything you can think of, regardless of its native format. Even when different types of data have been stored in unrelated systems, you can dump it all into your Hadoop cluster with no prior need for a schema. Hadoop's cost advantages over legacy systems redefine the economics of data. One of the cost advantages of Hadoop is that because it relies in an internally redundant data structure and is deployed on industry standard servers rather than expensive specialized data storage systems, you can afford to store data not previously viable.

Hdfs Architecture:

Hadoop Distributed File System is a simple file system inspired by Google's File System(GFS). Designed for very large size data sets on a cluster. Developed on a idea Write Once Read Many(WORM).Data set is copied from the source directory , then parallel analysis is performed on the local data set. Using HDFS we can perform operations like copy files, create directory, delete directory, copy files from local, delete files.It is a DFS that provides high performance access to data across Hadoop clusters. HDFS takes in data, breaks the information into separate pieces and distributes them to different nodes in a cluster allowing for parallel processing. This HDFS copies the piece of data multiple times and distributes the copies to individual nodes and placing atleast one copy on a different server rack than the others.

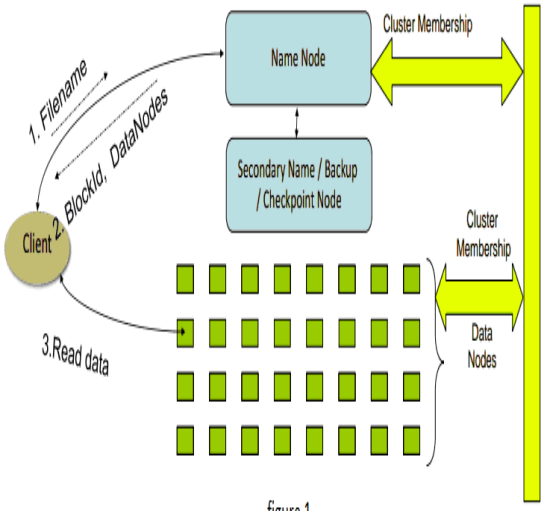


figure 1

From the figure it clearly explains, HDFS uses master/slave architecture with each cluster consisting of a single name node that manages file system operation and supporting data nodes that manages data storage on individual computer nodes Clusters are nothing but the group of servers and other resources that act like a single system and enables high availability and in some cases load balancing and parallel processing. size of the clusters can be varied. Maximum number of clusters on hardisk depends on the size of the FAT (File Allocation Table)entry.

Mapreduce:

We will focus on Hadoop Map Reduce, which is the most popular open source implementation of the Map Reduce framework proposed by Google (Lu, X., et al., 2012). Generally speaking, a Hadoop Map Reduce job mainly consists of two user-defined functions: map and reduce. The input of a Hadoop Map Reduce job is a set of key-value pairs(k; v) and the map function is called for each of these pairs. The map function produces zero or more intermediate key-value pairs(k0; v0). Then, the Hadoop Map Reduce framework groups these intermediate key-value pairs by intermediate key k0 and calls the reduce function for each group. Finally, the reduce function produces

zero or more aggregated results. The beauty of Hadoop Map Reduce is that users usually only have to define the map and reduce functions. The framework takes care of everything else such as parallelization and failover. The Hadoop Map Reduce framework utilizes a distributed file system to read and write its data. Typically Hadoop Map Reduce uses the Hadoop Distributed File System (HDFS), which is the open source counterpart of the Google File System (https://nppes.cms.hhs.gov.). Therefore, the I/O performance of a Hadoop Map Reduce job strongly depends on HDFS. we will introduce Hadoop Map Reduce and HDFS in detail. We will contrast both with parallel databases. In particular, we will show and explain the static physical execution plan of Hadoop Map Reduce and how it affects job performance. In this part, we will also survey high level languages that allow users to run jobs even more easily.

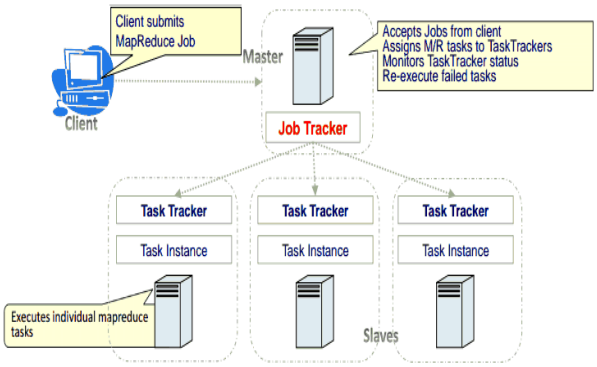


figure 2

If a node remains silent for longer than the expected interval, the master node makes node and reassigns the work to other nodes. The key to how Map Reduce works is to take input as, conceptually, a list of records. The records are split among the different computers in the cluster by Map. The result of the Map computation is a list of key/value pairs. Reduce then takes each set of values that has the same key and combines them into a single value. So Map takes a set of data chunks and produces key/value pairs and Reduce merges things, so that instead of a set of key/value pair sets, you get one result. You can't tell whether the job was split into 100 pieces or 2 pieces. Map Reduce isn't intended to replace relational databases: it's intended to provide a lightweight way of programming things so that they can run fast by running in parallel on a lot of machines.

3.Additional E-Health Capabilities:

3.1 Data Federation and Aggregation:

The various types have been described in the previous section, and those sources reflect the fragmentation of e-Health data among the various stakeholders, including payers, providers, labs, ancillary vendors, data vendors, standards organizations, insurance institutions and regulatory agencies. Solutions for big data will break the traditional model, in which all data is loaded into a warehouse. Data federation will emerge as a solution in which the big data architecture is based on a collection of nodes within and outside the enterprise and accessed through a layer that integrates the data and analytics. Additional central data collection points are emerging as HIE, RHIO and NHIN work to facilitate the exchange of healthcare data.

3.2Security and Regulatory Concerns:

This is the most fundamental requirements that distinguish the Big Data services for e-Health. They deal

with additional challenges, such as privacy, security and legal concerns, as well as questions about authenticity, accuracy and consistency. The entire healthcare system can realize benefits from democratizing big data access and the cloud (Park, E.K. and W. Liu, 2013) makes exposing and sharing big data easy and relatively inexpensive. However, significant security and privacy concerns exist, including the Health Insurance Portability and Accountability Act (HIPAA). A credentialing process could facilitate and automate this access, but there are complexities and challenges. Since providers, patients and other interested parties such as researchers need various secure accesses, data security policies have to control by group, role and function. Finally, the security of the data once it leaves the cloud also needs to be assured

3.3 Predictive Analytics:

Heart patients weigh themselves at home with scales that transmit data wirelessly to their health center. Algorithms analyze the data and flag patterns that indicate a high risk of readmission, alerting a physician. people.

Historical EMR Analysis:

Hadoop reduces the cost to store data on clinical operations, allowing longer retention of data on staffing decisions and clinical outcomes. Analysis of this data allows administrators to promote individuals and practices that achieve the best results.

Processing Healthcare Data:

Depending on the use case, healthcare organizations process data in batch (using Apache Hadoop Map Reduce and Apache Pig); interactively (with Apache Hive); online (with Apache HBase) or streaming (with Apache Storm).

3.4Analyzing Healthcare Data:

Once data is stored and processed in Hadoop it can either be analyzed in the cluster or exported to relational data stores for analysis there. These data stores might include:

- Enterprise data warehouse
- Quality data mart
- Surgical data mart
- Clinical info data mart
- improve data quality.
- Diagnosis data

3.5 Data Interoperability Management:

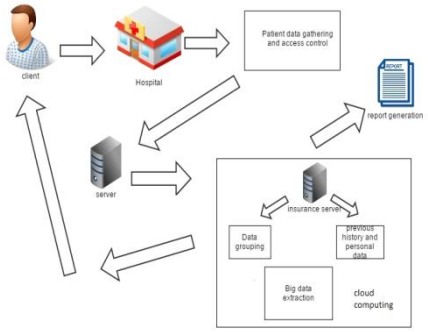
Our solution architecture is flexible enough to cope with not only the additional sources but also the evolution of schemas and structures used for transporting and storing data. To ensure analytics are meaningful, accurate and suitable, metadata and semantic layers are supported that accurately define the data and provide business context and guidance, including appropriate and inappropriate uses of the data. This evolution of standards will eventually improve data quality

The Proposed System:

In our proposed system, we proceed by collecting large amount of data sets, here in our project for example we consider medical records from various hospitals. These medical records are thus considered as Big Data and thereby creating an application using struts and hibernate in which frontend was designed using html5 and css3.The end users of this application are the patient's, Insurance company, doctors of the various hospitals. The main aim of this application is that the end users can access their record from anywhere and view their past medical history, present details and other details of the patient's using their unique patient ID. In this application we use mysql as our database from which data are transferred to HDFS using Hive which is an open source product using which the data can be transferred to Hadoop database. Map Reduce coding is done to process massive amounts of unstructured data parallely. We are analyzing more number of Factors like Disease Types with its Corresponding Reasons, Insurance policy Details with Sanctioned Amount, Family Grade wise Segregation.

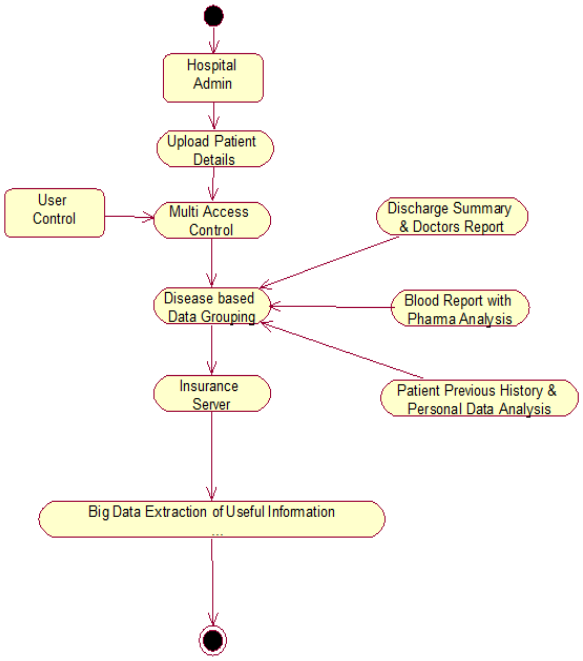
System Architecture:

Data are collected during the first visit at the outpatient clinics and includes information of personal information, date of last-ve and first +ve test, route of infection and drug usage. Data are collected on the basis when every time the patient comes to treatment. It includes treatment, symptoms and lab results. Separate accounts for patient and insurance are created for user and maintained. The user request are respond by the service provider accordingly. we create a server in the data warehouse in which that data from the hospital database is moved to the insurance server. This server is consist of medical claim information of the patient. At any point of time we can query the sever for details. Information present on the server are managed and extracted using hadoop. Previous history of patient consist of past records if he is hospitalized. Patient history and personal details are maintained consistently and stored as single modifiable file. There is no need of changing the replicated data because it is updated continuously. There are various report like discharge summary, diagnosis, laboratory report and pharmacy details. Based on the kind of report, the information are provide from the insurance server as report. Map reduce is a technique of big data which allows to extract information from the server by the analyst. The analyst has a permission to access the insurance server and he will use the above technique to extract the useful information from the vast amount of data



Program Flow:

The entire system is according to official Hadoop manual.HDFS in charge of store the large data sets we need to process, offer the high throughput of data access rather than low latency. Map Reduce component in charge of the processes distributed.



Benefits Of Deploying Hadoop:

Hadoop is scalable storage platform it can store very large data sets. It also enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. In addition, hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection. A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy is available for user. Despite the fact that the open source version of Hadoop is free, the claim that Hadoop is cheap is a myth. While Hadoop is a much more cost effective solution for storing large volumes of data, it can still require major investments in hardware, development, maintenance and expertise. This level of investment may be worth it for companies that plan on using Hadoop regularly, but for those who only need to run occasional analytics or simply don't have the budget for the upfront investment, the cost of Hadoop can be a real issue. Big data analytics has a lot to offer. From a more complete view of the consumer to more efficient manufacturing and improved product innovation, valuable insights lie within the mass stores of multi-structured data being created.

Conclusion:

Hadoop achieves more and more attention by the academia industry. And its application is more and more widespread. This paper describes the architecture of Hadoop in the cloud. Thus hadoop in the cloud will be the future technology since markets are growing the data's are also growing the best solution to handle the Big Data is hadoop. Thus moving Hadoop to Cloud will result in many advantages like flexibility, cost efficient and in power consumption. Big Data Stream setup and provision to support application flow associations, data federation and aggregation along the streaming paths, partitioning of E-Health messages along the processing. Security to meet stringent e-Health environments and regulations is an integral part of the infrastructure. Multiple flows can be managed by a common MPMD model. End-to-end regulatory oversights can be guaranteed. Quality of Service guarantees includes real-time processing, reconfiguration and enhancement of processing cluster and network capacity, data interoperability management, as well as reporting capabilities. So, the proposed method will play an important role in future.

REFERENCES

Liu, W. and E.K. Park, 2013."e-Health AON (Application Oriented Network)", IEEE International Conference on Computer Communication Networks, BMAN Workshop, Nausa, Bahamas.

Park, E.K. and W. Liu, 2013. "e-Healthcare Cloud Computing Application Solutions", IEEE ICNC2013, International Conference on Computing, Networking and Communications, San Diego, CA.

Liu, W., E.K. Park and Udo R. Krieger, 2012."e-Health Interconnection Infrastructure Challenges and Solutions Overview", IEEE HealthCom-2012, the 14th IEEE International Conference on e- 9877 Health Networking, Application & Services, Beijing, China.

Chute, C.G., 2012."Obstacles and options for big-data applications in biomedicine: The role of standards and normalizations", 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Li, D., C. Tao, H. Liu and C. Chute, 2012. "Ontology-Based Temporal Relation Modeling with MapReduce Latent Dirichlet Allocations for Big EHR Data", Second International Conference on Cloud and Green Computing (CGC).

Lu, X., H. Tang, W. Cheng and T. Zhang, 2012. "Heterogeneous Data Source Middleware for Android E-Health Application", Eighth International Conference on Mobile Ad-hoc and Sensor Networks (MSN).

Diaz, M., G. Juan, O. Lucas and A. Ryuga, 2012."Big Data on the Internet of Things: An Example for the E-health, "Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS).

Liu, W. and E.K. Park, 2012. "e-Healthcare Security Solution Framework", IEEE International Conference on Computer Communication Networks, MobiPST-2012, Munich, Germany, August 2012.

Liu, W. and E.K. Park, 2011."e-Health Service Characteristics and QoS Guarantee", IEEE International Conference on Computer Communication Networks, Workshop on Context-aware QoS Provisioning and Management for Emerging Networks, Applications and Services, Maui, HI.

Health Level Seven International, http://www.hl7.org/implement/ standards.

National Provider Identification (NPI), https://nppes.cms.hhs.gov.

Health Industry Number System (HIN), http://www.hibcc.org/ hinsystem.htm.

NCPDP, National Council for Prescription Drug Program,http://www.ncpdp.org.

Digital Imaging and Communications in Medicine (DICOM), http://medical.nema.org.

US Congress, "Health Insurance Portability And Accountability Act", 1996.

ISO/IEEE11073, "Medical/Health Device Communication Standards", 2004 (base standards) ~ 2012 (additional parts and revisions).

U.S. Department of Health & Human Services, "National Health Information Network", http://healthIT.hhs.gov.

http://storm-project.net/

http://hadoop.apache.org/

Kaladevi, S and B. Gayathri, 2014. " Efficient resource sharing on cloud computing using hadoop based system"/published.