ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Data Mining with Big Data e-Health Service Using Map Reduce

Abinaya.K

PG Student, Department Of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamilnadu, India

Abstract*:* Introducing the new knowledge of Big Data for belief apprehension of large-volume, complex, growing data sets with several autonomous sources. HACE theorem that characterizes the features of big data revolution and perform the operation in data mining perspective. Big Data e-Health Service application has promised to transform the whole healthcare heart disease process to become more efficient, less expensive and higher quality. This application involves data-driven model demand-driven aggregation of information sources. Big Data is transforming healthcare, business, as e-Health heart disease becomes one of key driving factors during the innovation process. Look into BDeHS (Big Data e-Health Service) to fulfil the Big Data applications in the e-Health service domain. Existing Data Mining technologies such cannot be simply applied to e-Health services directly. Our design of the BDeHS for heart disease that supplies data operation management capabilities and e-Health meaningful usages.

Keywords: Big data, Data mining, HACE theorem, Cloud storage, Hadoop map reduce.

I. INTRODUCTION

Data mining is the process of discovering useful, possibly unexpected, patterns in data and it is an non-trivial extraction of implicit existing unknown data and it is potentially useful for information from data. It will do exploration and analysis by automatically or semi-automatic of large data in order to discover meaningful patterns.

Big data that concerns a large volume complex and growing data sets with several autonomous sources. Big data is rapidly growing in all science and engineering .Big data is characterized as (vvvvc) (ie) volume, velocity, variable, variety and complexity.In the recent years big data is growing in all science and engineering here we going to implement BDeHS for heart disease.

Today everybody talks about e-health, but some people have come up with a clear definition by comparing with new term. In the previous year of 1999 it used rarely but now it seems to serve as a general "buzzword", which is used to characterize not only internet medicine, but also everything which is virtually related to computers and medicine. This term was evidently first used by the industry leaders and marketing people rather than academics and others. Such people's created and used this term in line with other e-words such as e-commerce and e-business it also includes like e-solutions and e-filing in an attempt to transfer the principle and excitement around e-commerce to the health arena not only this but also to give an account of the new possibilities to internet for opening up the area of health care.

Big data health care becomes one of the key driving factors during the innovation process, the most fundamental challenge of big data is to explore large volume of data's and extract the useful information or knowledge. This paper focus to transform the health care service to big data and especially for heart disease. Today heart disease is growing higher and higher to know about the disease causes and to know about the heart disease from all the sources this health care service will be useful. This paper introduce HACE theorem that characterizes the features of big data revolution and perform the operation in data mining perspective. As sources and volume of information increase, the expectations in utilizing those large volume of e-Health data for heart disease is to reduce costs, boost outcomes and improve treatment. We investigate BDeHS heart disease to fulfil the Big Data applications in the e-Health service domain.

This paper is organized as follows. In section we describe the characteristic and functionality of HACE theorem. To further extend the Big Data service approach into a national and global framework for e-Health care heart disease. In section II we described our related work. In section III we describe our propose design of the BDeHS heart disease architecture that provides the operational management capabilities, and promising e-Health heart disease meaningful usages. In Section IV we describe the algorithm used. The final section concludes with a summary of our contributions

HACE THEOREM:

HACE theorem stands for Heterogeneous Autonomous Sources with Distributed and Decentralized Control and evolving the relationship among various data. This theorem categorize the extreme challenge for discovering the useful information or knowledge from big data .Autonomous data sources with distributed and decentralized is the main characteristic for the application of big data. Being autonomous each data source is able to produce or generate information without involving the centralized. Thus the data heterogeneity and diverse dimensionality issues become one of the major challenges if we are enabling data aggregation by the combination of data from all sources. The main fundamental characteristics of the Big Data are the large volume of data represented by heterogeneous and diverse dimensionalities. This is happened because different

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

information collectors used their own design or protocols for data recording, so the nature of different applications also described in diverse data. Let us take one example as each individual human being in a biomedical world can be represented by a simple demographic information such as name, age, sex, family disease, etc.

To verifying X-ray and CT scan of every human being the digital image or videos are used to show the results because they will supply a information in a visual manner to the particular doctor to give the detailed description about it. Thus the volume of the Big Data increases, the complexity and the relationships also increases. The previous stage of centralized data information systems focus on finding the good feature values to describe all observations.

II. RELATED WORK

Twitter mood predicts the stock market by J. Bollen, H. Mao, and X. Zeng this paper describes the Basic behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making[1]. Evaluate the text content of daily Twitter feeds by two mood tracking tools, namely Opinion Finder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy).

Efficient influence maximization in social networks by Y.-C. Chen, W.-C. Peng, and S.-Y. Lee. In this paper Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence [2]. Match the efficient influence maximization from two complementary directions. One is to improve the original greedy algorithm and its improvement to further reduce its running time, and the second is to propose new degree discount heuristics that improves influence spread.

Algorithms for mining the evolution of conserved relational states in dynamic network by R. Ahmed and G. Karypis**.**This paper help to identify the transitions from one conserved state to the next and may provide evidence to the existence of external factors that are responsible for changing the stable relational patterns in these networks [3]. Presents a new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states.

Collective mining of bayesian networks from distributed heterogeneous data by R. Chen, K. Sivakumar, and H. Kargupta.This paper uses Local Bayesian network at each site using the local data[4]. Then each site identies the observations that are most likely to be evidence of coupling between local and non-local variables and transmits a subset of these observations to a central site. Another Bayesian network is learnt at the central site using the data transmitted from the local site. The local and central Bayesian networks are combined to obtain a collective Bayesian network and that models the entire data.

Mining high speed data streams by P.Domingos and G. Hulten.In this paper databases that grow without limit at a rate of several million records per day[5]. Mining these continuous data streams brings unique opportunities, but also new challenges. Describes and evaluates VFDT, an anytime system that builds decision trees using constant memory and constant time per example. Describes and evaluates VFDT, an anytime system that builds decision trees using constant memory and constant time per example. VFDT can in-corporate tens of thousands of examples per second using on the self hardware. It uses Hoeffding bounds to guarantee that its output is asymptotically nearly identical to that of a conventional learner. VFDT's properties and demonstrate its utility through an extensive set of experiments on synthetic data.

III. MINING BIG DATA USING MAPREDUCE

Big Data e-Health Service Application Knowledge includes data accessing and computing, information sharing, data privacy and domain and application knowledge. It is the term for a collection of data Clusters so large and complex that it becomes difficult to process. Implement the HACE theorem. This application managed both structured and unstructured users data and applies the cluster filter by using map reduce technique.

Mapreduce:

Mapreduce is a parallel programming model that is used to retrieve the data from the Hadoop cluster. In this model, the library handles lot of messy details that programmers doesn't need to worry about. For example, the library takes care of parallelization, fault tolerance, data distribution, load balancing, etc.This splits the tasks and executes on the various nodes parallel, thus speeding up the computation and retrieving required data from a huge dataset in a fast manner. This provides a clear abstraction for programmers. They have to just implement (or use) two functions: map and reduce. The data are fed into the map function as key value pairs to produce intermediate key value pairs.

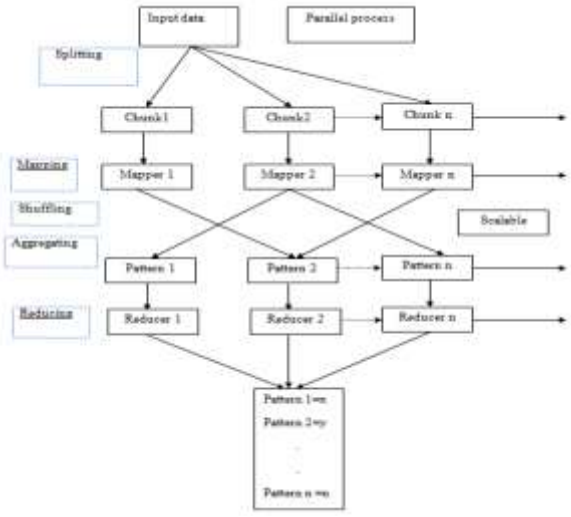

Fig. 1 Map reduce architecture

Once the mapping is done, all the intermediate results from various nodes are reduced to create the final output.JobTracker keeps track of all the Mapreduce jobs that are running on various nodes. This schedules the jobs,

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

keeps track of all the map and reduce jobs running across the nodes. If any one of those jobs fails, it reallocates the job to another node, etc. In simple terms, Job Tracker is responsible for making sure that the query on a huge dataset runs successfully and the data is returned to the client in a reliable manner.TaskTracker performs the map and reduce tasks that are assigned by the Job Tracker. Task Tracker also constantly sends a heartbeat message to Job Tracker, which helps Job Tracker to decide whether to delegate a new task to this particular node or not. We've only scratched the surface of the Hadoop. This is just the first article in our ongoing series on Hadoop. In the future articles of this series, we'll explain how to install and configure Hadoop environment, and how to write Mapreduce programs to retrieve the data from the cluster, and how to effectively maintain a Hadoop infrastructure.

Overall, cloud computing provides better agility and scalability, together with lower costs and faster time to market. However, it does require that applications be engineered to take advantage of this new infrastructure; applications built for the cloud need to be able to scale by adding more servers, for example, instead of adding capacity to existing servers.

On the storage layer, traditional relational databases were not designed to take advantage of horizontal scaling. A class of new database architectures, dubbed NoSQL databases, are designed to take advantage of the cloud computing environment. NoSQL databases are natively able to handle load by spreading data among many servers, making them a natural fit for the cloud computing environment. Part of the reason NoSQL databases can do this is that related data is always stored together, instead of in separate tables. This document data model, used in MongoDB and other NoSQL databases, makes them a natural fit for the cloud computing environment.

In fact, MongoDB is built for the cloud. Its native scale-out architecture, enabled by "sharding," aligns well with the horizontal scaling and agility afforded by cloud computing. Sharding automatically distributes data evenly across multi-node clusters and balances queries across them. In addition, MongoDB automatically manages sets of redundant servers, called "replica sets," to maintain availability and data integrity even if individual cloud instances are taken offline. To ensure high availability, for instance, users can spin up multiple members of a replica set as individual cloud instances across different availability zones and/or data centers. MongoDB has also partnered with a number of leading cloud computing providers, including Amazon Web Services, Microsoft and SoftLayer.

Combining Operational and Analytical Technologies; Using Hadoop

New technologies like NoSQL, MPP databases, and Hadoop have emerged to address Big Data challenges and to enable new types of products and services to be delivered by the business. One of the most common ways companies are leveraging the capabilities of both systems is by integrating a NoSQL database such as MongoDB with Hadoop. The connection is easily made by existing APIs and allows analysts and data scientists to perform complex, retroactive queries for Big Data analysis and insights while maintaining the efficiency and ease-of-use of a NoSQL database.

NoSQL,MPP databases and Hadoop are complementary: NoSQL systems should be used to capture Big Data and provide operational intelligence to users, and MPP databases and Hadoop should be used to provide analytical insight for analysts and data scientists. Together, NoSQL, MPP databases and Hadoop enable businesses to capitalize on Big Data.

Considerations for Decision Makers

While many Big Data technologies are mature enough to be used for mission-critical, production use cases, it is still nascent in some regards. Accordingly, the way forward is not always clear. As organizations develop Big Data strategies, there are a number of dimensions to consider when selecting technology partners the following techniques are used.

Online vs. Offline Big Data

Big Data can take both online and offline forms. Online Big Data refers to data that is created, ingested, trans- formed, managed and/or analysed in real-time to support operational applications and their users. Big Data is born online. Latency for these applications must be very low and availability must be high in order to meet SLAs and user expectations for modern application performance. This includes a vast array of applications, from social networking news feeds, to analytics to real-time ad servers to complex CRM applications. Examples of online Big Data databases include MongoDB and other NoSQL databases.

Offline Big Data encompasses applications that ingest, transform, manage and/or analyze Big Data in a batch context. They typically do not create new data. For these applications, response time can be slow (up to hours or days), which is often acceptable for this type of use case. Since they usually produce a static (vs. operational) output, such as a report or dashboard, they can even go offline temporarily without impacting the overall goal or end product. Examples of offline Big Data applications include Hadoop-based workloads; modern data warehouses; extract, transform, load (ETL) applications; and business intelligence tools.

Organizations evaluating which Big Data technologies to adopt should consider how they intend to use their data. For those looking to build applications that support real-time, operational use cases, they will need an operational data store like MongoDB. For those that need a place to conduct long-running analysis offline, perhaps to inform decision-making processes, offline solutions like Hadoop can be an effective tool. Organizations pursuing both use cases can do so in tandem, and they will sometimes find integrations between online and offline Big Data technologies. For instance, MongoDB provides integration with Hadoop.

DOI 10.17148/IJARCCE.2015.4227

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

IV. MINING ALGORITHM

LL and MF for Multiple information sources:

Mostly Big Data e- Health Service applications are featured with autonomous sources and decentralized controls with aggregating distributed data sources to a centralized site for mining is systematically restricted due to the some potential transmission of cost and privacy concerns. This (local learning and model fusion for multiple information sources) is to carry out such mining activities at each distributed site the unfair view of the data collected at all site often leads to unfair decisions or models thus the big data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites can work together to achieve a global effective goal. Some model mining and correlations are the key steps to ensure that models or patterns discovered from several information sources can be combine to meet the aim of global mining . And more specifically, the global mining can be featured with a two-step one is local mining and the other is global correlation process at data model and at knowledge levels. In the data level each local site can calculate their data statistics according to local data sources and exchange the statistics between sites to obtain a global data distribution view. In the model or pattern level each site can carry out local mining activities to discover local patterns by exchanging patterns between multiple sources and new global patterns can be synthesized by aggregating patterns across each sites.

Mining from Sparse, Uncertain, and Incomplete Data:

In the spare uncertain and incomplete data are describing the features for Big Data applications. In every human being

sparse, number of data points is very less for drawing reliable conclusions and this is normally a complex of the data dimensionality issues where data in a high-dimensional space do not show clear trends or distributions. Such uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random error distributions. Data's like incomplete data's are refer to the missing of data field values for some samples the missing values can be caused by different realities such as the malfunction of a sensor node or some systematic policies to intentionally skip some values.

Mining Complex and Dynamic Data:

The growth of big data is driven by the rapid increasing of difficult data's and their changes in volumes and in nature documents like servers internet backbones, social networks communication networks are all featured with complex data. Where complex dependency structures concealed the data raise the complexity for our learning systems they also offer some exciting opportunities that simple data representations are being incapable of achieving.

Self-directed data sources with distributed and decentralized controls are main characteristic of Big Data applications. Each data source is able to generate and collect information without involving (or relying on) any centralized control. On this way the application of e-health care for disease datas are uploaded successfully and obtained the correct values of result for retrieval from various source of clusters in bigdata.

V. SYSTEM ARCHITECTURE

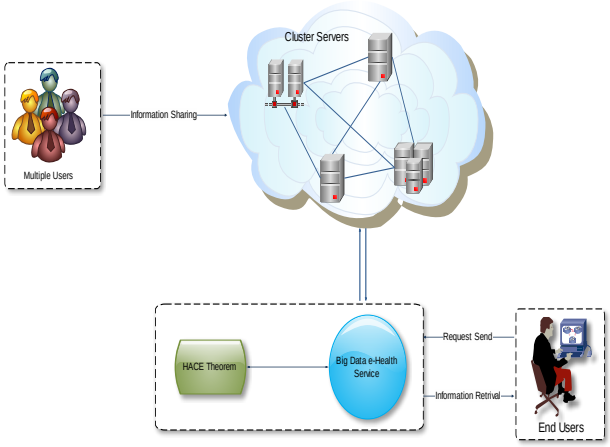

Fig. 2 Architecture for heart disease using map reduce technique

System design is the process of defining the architecture, components, modules, and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development blends the perspective of marketing, design, and manufacturing into a single approach to product development, then design is the act of taking the marketing information and creating the design of the product to be manufactured. System design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

VI. CONCLUSION

Big Data accurately concerns about data volumes from our HACE theorem suggests that the key characteristics of the Big Data e-Health Service application are huge with heterogeneous and diverse data sources, autonomous with distributed and decentralized control and complex and evolving in data and knowledge associations and combined characteristics suggest that Big Data require a “big mind e-Health Service” to consolidate data for maximum values.

ACKNOWLEDGMENT

I would like to thank my guide Prof. C.Chitra Asst. Prof., Computer Science and Engineering Department, Parisutham Institute of Technology and Science, Thanjavur for her help and guidance to enable me to propose this system.

REFERENCE

[1] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[2] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, “Efficient Algorithms for Influence Maximization in Social Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

[3] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[4] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.
[5] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

BIOGRAPHY

K.Abinaya received B.E (CSE) from Periyar Maniammai University in 2013. She is currently persuading M.E –Computer Science and Engineering in Parisutham Institute of Technology and Science.