INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 2 (2016) PAGES 510-516
RECEIVED : 26/02/2016. PUBLISHED : 7/03/2016
March 7, 2016

# APPLING ONTOLOGY KNOWLEDGE FOR MAP REDUCE HADOOP IN BIG DATA FRAME WORK

[1]B. Kalaiselvi*, Assistant Professor, Mahendra Engineering College for Women, Tamilnadu, India
[2]V. Ravindra Krishna Chandar, [1]Assistant Professor, Paavai Engineering College, Namakkal, India
[3]Dr. M. Thangamani, Assistant Professor, Kongu Engineering College, Tamilnadu, India
[2]manithangamani2@gmail.com

***ABSTRACT-*** In this problem of aligning ontologies and database schemas across different knowledge bases and databases is fundamental to knowledge management problems, including the problem of integrating the disparate knowledge sources that form the hadoop. Domain based research articles published is a laborious task which is usually done manually. Because of the enormous amount of journal articles published in various domains, mapping technique is the need of the hour. In this article, a fast ontology based approach map reduces Frame work system research articles to their domains using cloud on Big data. In this research to present a number of order to offer research on more informed choices when they deciding upon a Map Reduce hadoop in Big data framework to Improve their performance evaluation.

## I.INTRODUCTION

In ontology, Map Reduce has been widely applied in various fields of data and compute intensive applications and also it is an important programming model for cloud computing. Hadoop is an open-source implementation of Map Reduce which operates on terabytes of data. The core of Hadoop includes file System, RPC, and serialization libraries and also it provides basic services for building a cloud computing environment. The Hadoop Distributed File System (HDFS) is a distributed file system which stores terabytes or petabytes of data also it provides high speed access to the application data. It is highly designed to run on clusters of commodity machines. Map Reduce framework is processing large datasets on compute clusters in distributed way. Map Reduce framework handles all complexities and distribution of the data as well as of map and reduce task.

The main point in using Hadoop is to handle large datasets efficiently. In this system, we have applied Hadoop Map Reduce model to analyze. Cloud computing as a distributed computing paradigm aims at large datasets to be processed on available computer nodes by using a Map Reduce framework. MapReduce is a software framework introduce Big data connotes performing database operations and computations for substantial amounts of data remotely from the data owner enterprise. The Information is being created and gathered at a rate that is rapidly nearing the Exabyte or year range. It creates and collection is speeding will come close to the Zetta byte or year range in a few years. The challenge is not only to store and manage this enormous volume .but also to analyze and draw significant value from it .It focuses on analysis of problems and challenges in conforming and going for Big data technology and its optimal solution using HDFS and Map Reduce framework. A large number of organizations are facing the problem of explosion of data and the size of the databases used in today enterprises has been growing at exponential rates. Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured process Today business applications are having enterprise features like large scale, data-intensive, web-oriented and accessed from diverse devices including map reduce. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task.

In term “Big data” is used for large data sets. In size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data found in current scenario includes web logs, RFID generated data, sensor networks, satellite and geo-spatial data, social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research, military. Surveillance, medical records, photography archives, video archives, and large-scale ecommerce. Due to

the latest development of efficient data transfer, the capability of handling huge data and the retrieval of data efficiently needs to be considered. Since the data that is stored increases voluminously, methods to retrieve relative information and security related concerns are to be addressed efficiently to secure in bulk data. Also with emerging concepts of big data, these security issues are a challenging task. It secures data transfer using the concepts of data mining in cloud environment using hadoop Mapreduce. In the proposed research describes Hadoop Map Reduce programming model with applying ontology for analyzing Big data. So that it could get hit count of specific web application. This system uses Hadoop file system to store log file and results are evaluated using Map and Reduce function as based on ontology.

### 1.1 ONTOLOGY KNOWLEDGE

Ontology Knowledge really facilitates the construction of ontologies by the ontology engineer. The vision of ontology includes a number of complementary disciplines that feed on different types of unstructured, semi-structured and fully structured data in order to support a semi-automatic, cooperative ontology engineering process. Ontology learning framework proceeds through ontology import, extraction, pruning, refinement, and evaluation giving the ontology engineer a wealth of coordinated tools for ontology modeling. Besides of the general framework and architecture, ontology Knowledge cycle that, research have implemented in our environment, Text-To Onto, such as ontology knowledge from free text, from dictionaries, or from legacy ontologies, and refer to some others that need to complement the complete architecture. In automated construction of knowledge discovery workflows, given the types of inputs and the required outputs of the knowledge discovery process.

It consists of two main ingredients. The first one is defining a formal conceptualization of knowledge types and data mining algorithms by means of knowledge discovery ontology. The second one is workflow composition formalized as a planning task using the ontology of domain and task descriptions. The baseline version demonstrates suitability of the knowledge discovery ontology for planning and uses Planning Domain Definition Language (PDDL) descriptions of algorithms to this end, a procedure for converting data mining algorithm descriptions into PDDL was developed. Directly queries the ontology using a reasoned. In the ontology extraction phase of the ontology process, major parts it complete ontology or large chunks reflecting a new subdomain of the ontology, are modeled with learning support exploiting various types of (Web) sources. Thereby, ontology learning techniques partially rely on given ontology parts. It encounters an iterative model where previous revisions through the ontology learning cycle may propel subsequent ones. In this information retrieval stand on ontologies, dynamic semantic network, and lexical chains, significant an approach for achieve and indexing consequences by means of a narrative metric to compute semantic relatedness between words. It has numerous novelties in meticulous concerning utilize of a collective knowledge base from which we mine precise domain ontologies.

This domain-specific ontology is to get rid of conceptual and terminological confusion. It accomplishes this by specifying a set of generic concepts that characterizes the domain as well as their definitions and interrelationships some algorithms for identifying relations and constructing an information technology in Ontology, while extracting the concepts and objects from different sources. The Ontology is constructed based on three main resources ACM, Wikipedia and unstructured files from ACM Digital Library. Ontology's make the task of searching similar pattern of text that to be more effective, efficient and interactive. In domain ontology mining algorithm is used for supporting real-world ontology engineering. In particular, contextual information of the knowledge sources is exploited for the extraction of high quality domain ontologies and the uncertainty embedded in the knowledge sources is modeled based on notion. Fig. 1 shows the ontology knowledge Architecture.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 2 (2016) PAGES 510-516
RECEIVED : 26/02/2016. PUBLISHED : 7/03/2016
March 7, 2016

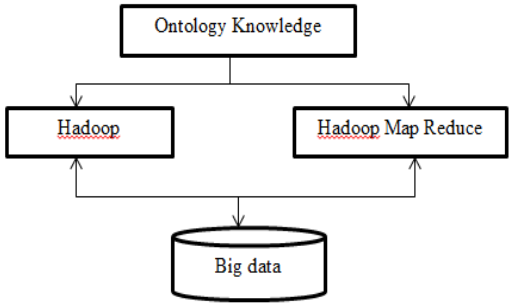


Fig.1 Ontology Knowledge Architecture

Empirical studies have confirmed that the proposed method can discover high quality domain ontology knowledge. Ontology knowledge is used from their domain-specific text documents. By using a full text parsing technique and in corporate both statistical and lexico-syntactic methods the knowledge extracted by our system is more concise and contains a richer to compared with alternative systems.

### 1.2 HADOOP

There is requirement of Big Data analytic frameworks for the organization that deal with different types of Big Data. There are several methods available for big data analysis like Apache Hadoop, Apache Drill etc., still very less work done in this approach. Hadoop is the technological answer to big data which provide Hadoop Distributed File System and MapReduce programming model is used for storage and retrieval of big data. But sometime execution may fail during processing. So more improvement is required at hadoop architecture to get better performance and overcome reliability.

Hadoop is an open source project hosted by Apache Software Foundation. It consists of many small sub projects which belong to the category of infrastructure for distributed computing. Hadoop mainly consists of two systems

- File System (The Hadoop File System)
- Programming Paradigm (Map Reduce)

The other subprojects provide complementary services or they are building on the core to add higher-level abstractions. There exist many problems in dealing with storage of large amount of data. Though the storage capacities of the drives have increased massively but the rate of reading data from them hasn't shown that considerable improvement. The reading process takes large amount of time and writing process is also slower. This time can be reduced by reading from multiple disks at once. Only using 100 of a disk may seem wasteful. But if there are 100 datasets, each of which is one terabyte and providing shared access to them is also a solution. There occur many problems also with using many pieces of hardware as it increases the chances of failure.

The main problem of combining data is being read from different devices. Many methods are available in distributed computing to handle this problem but still it is quite challenging. All the problems discussed are easily handled by Hadoop. The problem of failure is handled by the Hadoop Distributed File Systems problem of combining data is handled by Map reduce programming Paradigm. Map Reduce basically reduces the problem of disk reads and writes by providing a programming model dealing in computation with keys and values. Hadoop thus provides a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by MapReduce.

### 1.3.HADOOP MAPREDUCE

Hadoop MapReduce provides a mechanism for programmers to leverage in distributed systems for processing data sets. Map Reduce can be divided into two phases:

- Map Phase: Divides the workload into smaller sub-workloads and assigns tasks to Mapper, which processes each unit block of data. Output of Mapper is a sorted list of (key, value) pairs. This list is passed (also called shuffling) to the next phase.

- Reduce: Analyzes and merges the input to produce the final output. The final output is written to the HDFS in the cluster.

The following section provides an introduction to Map Reduce steps while processing a job:
Input step: Loads the data into HDFS by splitting the data into blocks and distributing to data nodes of the cluster. The blocks are replicated for availability in case of failures. The Name node keeps track of blocks and the data nodes in job

- Step-1: Submits the Map Reduce job and its details to the Job Tracker.
- Step-2: The Job Tracker interacts with Task Tracker on each data node to schedule a Reduce tasks.
- Step-3: Mapper process the data blocks and generates a list of key value pairs. • Sort step: Mapper sorts the list of key value pairs.
- Step-4: In shuffle step Transfers the mapped output to the reducers in a sorted fashion. • Reduce step: Reducers merge the list of key value pairs to generate the final result.

Finally, the results are stored in HDFS and replicated as per the configuration. The results are finally read from the HDFS by the clients. Here it is fault tolerant which is achieved by its daemons using the concept of replication. The daemons associated with the Map Reduce phase are job-tracker and task-trackers. Map-Reduce jobs are submitted on job-tracker. The Job Tracker pushes work out to available Task Tracker nodes in the cluster, striving to keep the work as close to the data as possible. A heartbeat is sent from the Task Tracker to the Job Tracker every few minutes to check its status whether the node is dead or alive. In this Programming model is designed to process large volumes of data in parallel by dividing the Job into a set of independent Tasks. The Job referred to full Map Reduce program, which is the execution of a Mapper or Reducer across a set of data. A Task is an execution of a Mapper or Reducer on a slice of data. So the Map Reduce Job usually splits the input data set into independent chunks, which are processed by the map tasks in a completely parallel manner. The Hadoop consists of a single Master node that runs a Job tracker instance which accepts Job requests from a client node and Slave nodes each running a Task Tracker instance. The Job tracker assumes the responsibility of distributing the software configuration to the Slave nodes, scheduling the job's component tasks on the Task Trackers, monitoring them and reassigning tasks to the Task Trackers when they failed. It is also responsible for providing the status and diagnostic information to the client. The Task Trackers execute the tasks as directed by the Job Tracker. The Task Tracker executes tasks in separate java processes so that several task instances can be performed in parallel at the same time. It depicts the different components of the Map Reduce framework. The high-level pipeline of the Hadoop Map Reduce. The Map Reduce input data typically come from the input files loaded into the HDFS. These files are evenly distributed across all the nodes in the cluster. In Hadoop, computer nodes and data nodes are all the same, meaning that the Map Reduce and HDFS run on the same set of nodes. At the mapping phase, the input file is divided into independent input splits and each split of these splits describes a unit of work that comprises a single map task in the Map Reduce job. The map tasks are then assigned to the nodes in the system based on the physically residence of the input file splits. Several map tasks can be assigned to an individual node, which attempts to perform as many tasks in parallel as it can. When the mapping phase has completed, the intermediate outputs of the map tasks are exchanged between all nodes and they are also the input of the reduction tasks. This process of exchanging the map intermediate outputs is known as the shuffling.

### 1.4. BIG DATA

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term "Big Data" is believed to be originated from the Web search companies the query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties

- Volume: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,
- Variety: Today data comes in all types of formats emails, video, audio, transactions etc.,
- Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

The other two dimensions that need to consider with respect to Big Data are Variability and Complexity

- Variability: Along with the Velocity, data flows can be highly inconsistent with periodic.

- Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Technologies today not only support the collection of large amounts of utilizing such data effectively. Transactions made all over the world with respect to a Bank, Walmart customer transactions, and Facebook users generating social interaction data. When making an attempt to understand the concept of Big Data, the words and "Hadoop" cannot be avoided believed to be originated from the web search volume - storing transaction. Fig.2 illustrates the Bigdata storage device.

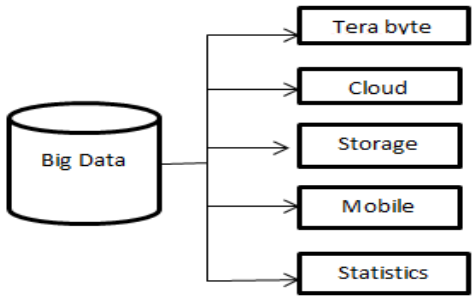


Fig.2 Big data storage device

**1.5 HADOOP BASED BIG DATA ANALYTICS**

Big data analytics is the process of examining large amounts of data in an effort to uncover hidden patterns or unknown correlations. Big Data Analytics Applications (BDA Apps) are new type of software application which analyzes big data using massive parallel processing frameworks. In Developers of such applications typically develop them using a small sample of data in a pseudo-cloud environment. They deploy applications in a large-scale cloud environment with considerably more processing power and larger input data. Big Data Analytics Applications (BDA Apps) are a new category of software applications that leverage large-scale data. This is typically too large to fit in memory or even on one hard drive to uncover actionable knowledge using large scale parallel-processing infrastructures. The big data can come from sources such as runtime information about traffic tweets during the Olympic Games, stock market updates, usage information of an online game or the data from any other rapid growing data-intensive software system.

**2. RELATED WORKS**

Ontology based approach has automatically map to research articles in their domains. It will help to quickly ascertain the domain of each research article and also to unravel their trending topics and the number of articles published in each domain in a particular time span. . Last 3 years of DBLP bibliographic dataset is used in this study. Hadoop map reduce technique is used to speed up the keyword extraction and subsequent ontology mapping Process. In Experiment studies revealed that the proposed ontology based research article topic in mapping technique framework is accurate, efficient and scalable process (Swaraj K P et al., 2015). The most common method for representing documents is the vector space model, which represents their document features as a bag of words and does not represent semantic relations between words. To solve the problem of clustering intensive data documents, they have addition to integrating the Word Net ontology with bisecting k-means in order to utilize the semantic relations between words to enhance document clustering results. In lexical categories for nouns only enhances their internal evaluation measures of document clustering and decreases the documents features from thousands to tens features in ( Abdelrahman Elsayed et al., 2015). An upper ontology describes very general concepts that are the same in all knowledge domains. One of the important functions of an upper ontology is to support broad semantic interoperability among large number of Ontologies. This enables the well-formed information to be semantically searchable on the web environment. It uses Protégé software to develop the ontology for testing purpose ontology using in SPARQL querying language. In attempt to represent their knowledge of published articles on Semantic Web and try to accomplish the vision of semantic web ( Ms swaminarayan priya et al., 2012). Ontology's make to search their similar pattern to text that to become more effective, efficient and

interactive. The current method for grouping in research project selection is using an ontology based text mining. In semantic knowledge is used to form ontology domain-specifically in text documents. By using a full text parsing technique and incorporating both statistical and lexico-syntactic methods, the knowledge extracted is more concise and contains a richer semantics compared with alternative systems. Quantitative evaluation are comparing with a state-of-the-art ontology learning system known as Text-To-Onto, has shown that CRCTOL produces much better precision and recall for both concept and relation extraction (Xing Jiang and Ah-Hwee Tan 2005). Hadoop is an efficient management of large scale Web services, where Hadoop can overcome the drawback which occurs in the traditional Web service management infrastructure. The two components of Hadoop, HBase and MapReduce, are integrated. The HBase table is non-functional property index mechanisms are designed to strengthen and retrieving their performance in functional and nonfunctional properties of Web services (Shangguang Wang et al., 2012). Hadoop Distributed File System, Master Namenode Failure affects the performance of the Hadoop Cluster. In current Scenario to overcome the name node failure, it replicates the Name node on the other Data node (Karwande V.S et al., 2015). In clusters mode using Hadoop Map Reduce model is used for maintaining the workload of the jobs. It would bring optimization not only in terms of privacy preservation but also with enhanced resource utilization in BigData based applications (Chhaya S Dule et al., 2014). It focuses on analysis of problems and challenges in conforming and going for Big data technology and its optimal solution using Hadoop Distributed File System (HDFS) and Map Reduce framework (Swapnil A. Kale1 et al., 2014). Hirdesh Shivhare and Nishchol Mishra, 2013, provides advantages and disadvantages of cloud in different examples the cloud services are different enterprises in the field of cloud computing cloud related with big data In use of visualization techniques to facilitate user understanding of the ontology alignment results. AlViz is implemented as a tab plug-in for Proteg´e. Ontology bootstrapping process integrates the results of both methods and apply in third method to validate the concepts using the service free text descriptor, thereby offering a more accurate definition of ontologies(Aviv Segev et sl., 2012).

**3. CONCLUSION**

## References

1. Abdelrahman Elsayed , Hoda M. O. Mokhtar and Osama Ismail. (2015). Ontology Based Document Clustering Using Mapreduce, 7, 1-12.
2. Aviv Segev, Quan Z. Sheng. (2012). Bootstrapping Ontologies for Web Services. IEEE Transactions On Services Computing, 5, 33-44.
3. Chhaya S Dule,Dr. Girijamma H.A,Rajasekharaiah K M. (2014). Privacy Preservation Enriched MapReduce for Hadoop Based Big Data Applications. American International Journal of Research in Science, Technology, Engineering & Mathematics, 6, 293-299.
4. Hirdesh Shivhare, Nishchol Mishra AlViz. (2006). AlViz - A Tool for Visual Ontology Alignment. Tenth international conference on information Visualization, 430 - 440 .
5. Karwande V.S, Dr. S. S.Lomte, Prof. R. A. Auti. (2015) . The Data Recovery File System for Hadoop Cluster -Review Paper. International Journal of Computer Science and Information Technologies, 6,365-367.
6. Swapnil A. Kale1, Prof. Sangram S.Dandge. (2006)., Understanding The Big Data Problems And Their Solutions Using Hadoop And Map-Reduce , International Journal of Application or Innovation in Engineering & Management, 3,439-445.
7. Shangguang Wang, Wei Su and Xilu Zhu . (2012). A Hadoop-based Approach for Efficient Web Service Management. International Journal of Web and Grid Services, 1-13.
8. Swaraj K P1 , Dr. Manjula D, Adhithyan. (2015). A novel ontology Based Framework for mapping. International Journal of Science, Technology and Management , 4, 1342-1347.
9. Swaminarayan priya. R, Ms. Nehal Daulatjada, Dr P.V Virpania, Dr V.R Rathod., (2012). Knowledge Representation of "Published Articles" in Semantic Web using Upper Ontology, International Journal of Advanced Research in Computer Science and Software Engineering, 2, 294-299.
10. Xing Jiang and Ah-Hwee Tan. (2005). Mining Ontological Knowledge from Domain-Specific Text Documents. Proceedings of the Fifth IEEE International Conference on Data Mining, 1-4.

**Ms. B. Kalaiselvi** has completed Master of Engineering in Computer Science and Engineering in Anna University Application. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing and ontology. She has presented any papers in national and international conferences in the above fields. She is currently working as Assistant Professor in Mahendra Engineering College for Women.

**Mr. V. Ravindra Krishna Chandar** is currently working as Assistant Professor in the Department of Computer Science and Engineering at Paavai Institutions, Namakkal District. He has a Master's degree in Software Engineering (2010) at Periyar University, Salem. He is proficient in Ontology, Medical Data Mining, Big Data Analytics, Cloud Computing and Database Management System. He has published the papers in International and National Journals and presented the papers in National conferences. He has reviewer in Science Publication.

M. Thangamani completed her B.E., from Government College of Technology, Coimbatore, India. She completed her M.E in Computer Science and Engineering from Anna University and PhD in Information and Communication Engineering from the renowned Anna University, Chennai, India in the year 2013. **Dr. M. Thangamani** possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 70 articles in refereed and indexed journals, books and book chapters and presented over 67 papers in National and International conferences in above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has organized many self-supporting and government sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She has received the International Award for the "Women of Distinction" from Venus International Foundation on 5th March, 2016 and "Senior Women Educator and Scholar Award" from the National Foundation for Entrepreneurship Development on 8th March 2016. She continues to actively serve the academic and research communities and presently guiding 6 PhD Scholars under Anna University. She is on the editorial board and reviewing committee of leading research journals, which includes her nomination as the Associate Editor to International Journal of Entrepreneurship and Small & Medium Enterprises at Nepal, Editor, International Scientific Journal of Contemporary Research in Engineering, Science and Management (ISJCRESM) and on the program committee of top international data mining and soft computing conferences in various countries. She is also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and countries like California, Dubai, Malaysia, Singapore, Thailand and China. She has Associate Editor in Canadian Arena of Applied Scientific Research, Canada. She has Life Membership in ISTE, Member in CSI, International Association of Engineers and Computer Scientists in China, IAENG, IRES, Athens Institute for Education and Research and Life member in Analytical Society of India. She is currently working as Assistant Professor at Kongu Engineering College at Perundurai, Erode District.