International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 206 – 208

Big Data Harmonization – Challenges and Applications

Prof. Jigna Ashish Patel
Assistant Professor, CE Dept,
Institute of Technology, Nirma University
Ahmedabad, India
Jignas.patel@nirmauni.ac.in

Dr. Priyanka Sharma
Professor,Rakshashakti University
Meghaninagar
Ahmedabad,India
pspriyanka@yahoo.co.in

***Abstract*—**As data grow, need for big data solution gets increased day by day. Concept of data harmonization exist since two decades. As data is to be collected from various heterogeneous sources and techniques of data harmonization allow them to be in a single format at same place it is also called data warehouse. Lot of advancement occurred to analyses historical data by using data warehousing. Innovations uncover the challenges and problems faced by data warehousing every now and then. When the volume and variety of data gets increased exponentially, existing tools might not support the OLAP operations by traditional warehouse approach. In this paper we tried to focus on the research being done in the field of big data warehouse category wise. Research issues and proposed approaches on various kind of dataset is shown. Challenges and advantages of using data warehouse before data mining task are also explained in detail.

Keywords-*Data warehouse, big data*

*****

I. INTRODUCTION

Big data is usually pronounced as data with a large volume, variety of data and great velocity. In this digital world everyone generate data, collectively it becomes huge. It is required that we deal with these data in order to provide the factual data analytics. Real time Data warehouse has to be updated to deal with this three V's, volume, variety and velocity[3] In new paradigm shift we cannot ignore big data for business intelligence. Plenty of innovations has been done on BI and data analytics, only few of them can be feasible if updated or modified for big data. It is not only sufficient if any solution can support only three V's. Challenges like scalability, work distribution and integration is also that important. Big data technology become very popular amongst every field nowadays. It may use to optimize businesses. Today, almost all type of companies/users who apply the business intelligence utilities are using the raw data generated from company and develop intelligence to take decisions. The huge amount of profile users data are processed and analyzed for advertising.

OLAP and data warehouse are typical fields of data science which have been talked since numerous decades by the Data Warehousing and database research groups. Modern data warehouse deals with every type of data which is controversy with traditional data warehouses. In the context of Big Data research, computing OLAP data cubes over Big Data becomes the most motivating challenges in the research community [5]

Traditional data warehouses process more like SQL type of dataset. It follows the steps like data preprocessing, data modelling, data normalization/denormalization, OLAP operations, and data visualization. Relational OLAP (ROLAP) faces lot of problems in big data era as data volumes get increased exponentially. Scalability, table join operations, unstructured data are major flaws in ROLAP [6]

ROLAP commonly uses star model and snowflake model, which stores dimensions and measures into relational tables, through foreign keys we refer those tables. In this big data era the performance of ROLAP is not accepted only because of costly join operations. Lot of research has been found to increase the performance of ROLAP by various techniques like indexing and hashing. On the other side Multidimensional OLAP (MOLAP) is best suitable solution for big data as it provides fast response time in join operations specifically when the data volume is high. MOLAP system offers robust performance, but it need additional storage to maintain the mappings between dimensions and measures.

Similarly the basic problem of computing OLAP data cubes are varied in the wide variety of data types like classical relational data sets, graph data sets, XML data sets, and social network data sets. Unluckily straight forward solutions won't helpful for computing OLAP data cube over big data. The reason behind this complexity is increasing data rate and dimensionality of model. When OLAP is merged with data mining, the technique is known as OLAM (Online Analytical Mining) which is very popular for its high performance but not suitable for semi structured or complex data. Lot of solutions (theoretical and practical) are proposed in order to get efficient computation and retrieval of the data from data warehouse and analyzing various heterogeneous data. Scalability and proficiency have been issues in social media from its emergent. Data warehouse is the key phase of the whole procedure from user query to faster response. Data warehouse is always in the

research for achieving Business Intelligence (BI). Though massive raw data is generated in every social media website, by using unique and intelligent BI tools and algorithms the target is achieved. Twitter provides tremendous amount of research dataset to work with your proposed algorithm. Since twitter is the most popular microblogging website, here we discussed the OLAP about the tools for twitter data set

II. CHALLENGES IN BIG DATA WAREHOUSE:

By considering data warehouse and big data, we found the following challenges:

1. Data Quality:

Biggest and very common challenge to deal with data is to ensure the data quality. Building a data warehouse require 75% of the efforts, such as readying the data and transporting it into the data warehouse. Get data from all heterogeneous sources and of different formats,it is real challenge to provide a single platform for all different kinds of data.To ready the data various data quality tools can be used to maintain the data quality. Lot of research in order to face the data problems and compared the available tools [5]

2. Scalability

Any type of data warehouse should deal with increasing data rate. Storing capacity of data warehouse should be flexible with real data size. It should support dynamic scaling. In the era of cloud computing perfect solution for big data warehouse is dynamic scaling. We may choose horizontal scaling or vertical scaling for our purpose. Variousplatforms are available for horizontal scaling like Hadoop and for vertical scaling like GPU (Graphics Processing Unit).

3. Efficiency

As far as efficiency of data warehouse is concern it is related to construction of data warehouse as well as its operating efficiency. Big Data mining techniques either applied via data warehouse or directly on data warehouse depending on convenience. If data warehouse is able to respond faster for the millions of queries then it is big efficiency concern [6]

4. Heterogeneity

Data coming from various heterogeneous sources results into variety of data, like structured, semi structured and unstructured data set. Some sources follow RDBMS type and some follows NoSQL databases. Every type of dataset must be provided a unique layer of data integration. Data warehouse should be flexible enough to deal with heterogeneous dataset so that data warehouse won't suffer from the cost of reconstruction [6].

III. APPLICATION AREA FOR OLAP AND BIG DATA

Social media

Today synonyms for social media is Facebook, Twitter, Quora, Instagram, you tube, WhatsApp, blogs and what not! Every social media sites and blogs are popular in their environment. If we talk about Twitter it has more than 500 million users and more than 340 million tweets per day. By efficient use of data streaming algorithms and efficient tools & technology tweets can be favorited, embedded, unlike, replied to, shared. With this millions of data and social networks Twitter performs analytics too. By taking this meta data and applying OLAP operations over it in the combination of data mining tasks lot of information and knowledge extraction like behavior of users, emerging trends issues can be analyzed. [3]

To extend the functionality and to overcome the limitations of OLAP, research is continue in the area of social media. Techniques and functionalities are modified in order to get good accuracy. Opinion mining and recommender systems are using semi structured data warehouses to extract knowledge to deal with unstructured as well as semi structured datasets. OLAP in social media should be extend to discover underlying measure for unstructured dataset [4].

Text data

For making the business wiser people are taking the help of reviews of users, advertisement, recommendation systems and lot more. Each of this methods are using textual data. To provide the platform for OLAP processing for textual data document warehousing is popularly used. Document warehousing is the solution for storing multidimensional documents and to do analysis over it for proficient text mining. By using document warehousing approach various heterogeneous document data can be integrated in well-formed infrastructure. Challenges like scaling, performance and security are also introduced in big data concern for document warehousing.[17]

Research point of view, document warehousing is the thirst area to contribute in OLAPing. The paper emphases on giving an improved solution of data warehousing in the big data era. Methodology mainly consists of three stages documentation, aggregation and data loading stage. Documentation stage remove the data from data sources by including that data to simple text files. Aggregation phase uses MapReduce process to finish ETL from various data files received from the first stage. In this phase all the results generated will be transformed into JSON objects. By using this approach parallelism can achieved better and big data problem can be solved [15]

Spatial Data

For the analytics of remote sensing data, spatial on line analytical processing (SOLAP) is used. SOLAP is a perfect solution for decision support system for exploring multidimensional perspective of spatial data. It can be used in spatio-temporal analytics for whether and environment monitoring systems. As the data generated from earth observation, it is very challenging to manage because of large scale and aggregation point of view. SOLAP cube uses the concept of map reduce in order to get higher parallelism. Newer approach is implemented on Hadoop framework using the traditional operations like roll-up/drill down and slice/dice on optimized ROLAP/MOLAP/HOLAP cube [1].

Web data

Extreme use of internet and web generates massive web dataset. By the concept of Web warehousing the critical aspect related to decision support system can be built. Advantages like improved productivity and cost savings can be achieved by applying web warehousing. Web warehousing is the approach to build the OLAP cube and warehouse on web information in the form of semi structured data, graphics, text, sound, images, multimedia objects, videos and many more. In simple language we may say web warehousing is the combination of data warehouse and web technology. Research in this area is to show how efficient web warehouse than the data warehouse by applying the web data on traditional warehouse. For the big data concern again map reduce procedures are used to avail high parallelism. Using the Hadoop framework and HBase gives improved results.[18]

References

[1] J. Li, L. Meng, F. Z. Wang, W. Zhang, and Y. Cai, "A Map-Reduce-enabled SOLAP cube for large-scale remotely sensed data aggregation," Computers & Geosciences, vol. 70, pp. 110–119, Sep. 2014.

[2] C. Blanco, I. García-Rodríguez de Guzmán, E. Fernández-Medina, and J. Trujillo, "An architecture for automatically developing secure OLAP applications from models," Information and Software Technology, vol. 59, pp. 1–16, Mar. 2015.

[3] N. U. Rehman, S. Mansmann, A. Weiler, and M. H. Scholl, "Building a Data Warehouse for Twitter Stream Exploration," 2012, pp. 1341–1348.

[4] L. Petrazickis, M. Butuc, and B. Steinfeld, "Crunching big data with Hadoop and BigInsights in the cloud," in Proceedings of the 2012 Conference of the Center for Advanced Studies on Collaborative Research, 2012, pp. 241–242.

[5] A. Nandi, C. Yu, P. Bohannon, and R. Ramakrishnan, "Data Cube Materialization and Mining over MapReduce," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 10, pp. 1747–1759, Oct. 2012.

[6] A. Cuzzocrea, L. Bellatreche, and I.-Y. Song, "Data warehousing and OLAP over big data: current challenges and future research directions," in Proceedings of the sixteenth international workshop on Data warehousing and OLAP, 2013, pp. 67–70.

[7] S. Mansmann, N. Ur Rehman, A. Weiler, and M. H. Scholl, "Discovering OLAP dimensions in semi-structured data," Information Systems, vol. 44, pp. 120–133, Aug. 2014.

[8] J. Song, C. Guo, Z. Wang, Y. Zhang, G. Yu, and J.-M. Pierson, "HaoLap: A Hadoop based OLAP system for big data," Journal of Systems and Software, vol. 102, pp. 167–181, Apr. 2015.

[9] D.-H. Shin and M. J. Choi, "Ecological views of big data: Perspectives and issues," Telematics and Informatics, vol. 32, no. 2, pp. 311–320, May 2015.

[10] J. Dittrich and J.-A. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2014–2015, 2012.

[11] S. Lee, S. Jo, and J. Kim, "MRDataCube: Data cube computation using MapReduce," in Big Data and Smart Computing (BigComp), 2015 International Conference on, 2015, pp. 95–102.

[12] I. Triguero, D. Peralta, J. Bacardit, S. García, and F. Herrera, "MRPR: A MapReduce solution for prototype reduction in big data classification," Neurocomputing, vol. 150, pp. 331–345, Feb. 2015.

[13] T. Niemi, J. Nummenmaa, and P. Thanisch, "Normalising OLAP cubes for controlling sparsity," Data & Knowledge Engineering, vol. 46, no. 3, pp. 317–343, Sep. 2003.

[14] N. U. Rehman, A. Weiler, and M. H. Scholl, "OLAPing social media: the case of Twitter," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 1139–1146.

[15] M. Ben Kraiem, J. Feki, K. Khrouf, F. Ravat, and O. Teste, "OLAP of the tweets: From modeling toward exploitation," in Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on, 2014, pp. 1–10.

[16] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis," 2008, pp. 905–910.

[17] F. S. C. Tseng and A. Y. H. Chou, "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence," Decision Support Systems, vol. 42, no. 2, pp. 727–744, Nov. 2006.

[18] X. Tan, D. C. Yen, and X. Fang, "Web warehousing: Web technology meets data warehousing," Technology in Society, vol. 25, no. 1, pp. 131–148, Jan. 2003.