CHAPTER 21
Proposal and Improvement of Data Structure for Big Data Analytics

Dr.R.Senthilkumar
Shree Venkateshwara HiTech Engineering College, India
Dr.R.S.Kamalakannan
Shree Venkateshwara HiTech Engineering College, India
Ms.K.Indumathi
Shree Venkateshwara HiTech Engineering College, India
K.Mahalakshmi
Shree Venkateshwara HiTech Engineering College, India

ABSTRACT

The explosion in the amount of data, called —data deluge‖, is forcing to redefine many scientific and technological fields, with the affirmation in any environment of Big Data as a potential source of data. Official statistics institutions a few years ago started to open up to external data sources such as administrative data. The advent of Big Data is introducing important innovations: the availability of additional external data sources, dimensions previously unknown and questionable consistency, poses new challenges to the institutions of official statistics, imposing a general rethinking that involves tools, software, methodologies and organizations.

Further, the rate at which this data is being generated induces extensive challenges of data storage, linking, and processing. A data-intensive cloud provides an abstraction of high availability, usability, and efficiency to users. However, underlying this abstraction, there are stringent requirements and challenges to facilitate scalable and resourceful services through effective physical infrastructure, smart networking solutions, intelligent software tools, and useful software approaches.

Keywords— Big data, Data Storage , Abstraction, Cloud etc

INTRODUCTION

ISSUES & CHALLENGES WITH BIG DATA

A major challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both:

- Design appropriate systems to handle the data effectively
- Analyze it to extract relevant meaning for decision making.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

They suggest there are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues. Each of these represents a large set of technical research problems in its own right.

- Taxonomies, ontologies, schemas, workflow
- Perspectives – backgrounds, use cases
- Bits – raw data formats and storage methods
- Cycles – algorithms and analysis

BIG DATA ANALYTICS (KNOWLEDGE DISCOVERY)

As discussed above, DBMS fall short when query is asked. Processing big size data is more difficult for the DBMS than ingestion of the data. Moreover the data when considered as big data, it is unstructured as it‘s characteristics is variety, while DBMS knows working with the structured data. Unstructured data is heterogeneous and variable in nature and comes in many formats, including text, document, image, video, and more. Unstructured data is growing faster than structured data. According to a 2011 IDC study it will account for 90 percent of all data created in the next decade. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly, and jump on new business opportunities. Big data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners, and the business—and ultimately gaining competitive advantage. Emerging technologies such as Hadoop and MapReduce are designed to address the three Vs of big data. They also put significant demands on infrastructure to support the distributed processing of unstructured data analytics.

ANALYTICS OF BIG DATA

―Analytics using big data (as characterized by volume, velocity, and variety) within an enterprise architecture (across multiple functional areas) to support critical operational processes (as contrasted with one-time ad-hoc analyses).‖

Big data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. The primary goal of big data analytics is to help companies make better business decisions by enabling data scientists and other users to analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI)programsese other data sources may include Web server logs and Internet click stream data, social media activity reports, mobile-phone call detail records and information captured by sensors.

According to Gueyoung Jung recently, collected data can exceed hundreds of terabytes and continuously generated. Such big-data represents data sets that can no longer be easily analyzed with traditional data management methods and infrastructures

BIG DATA & DATA STRUCTURES

As we have discussed earlier DBMS uses different data structures while dealing with data. These data structures are used at various processes carried out by DBMS .

1. Query processing and optimization

Query is the knowledge discovery from big data for the purpose different analytical softwares like Hadoop & mapreduce, Teradata aster are used.

2. Efficient representation of data on disk

The data is represented on disk using different data structures like b-Tree

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

3. Switching of data between main memory & external storage

The DBMS, while switching data between main memory uses different paging algorithms.One of the efficient paging algorithms is shadow paging. Though with the big data analytics are used it is not possible for any high processing server also to store the required big data in main memory. On big data also one need to perform all norma database operations on data, at the same time user is expecting operations to be performed efficiently. And for the purpose data structures helps.

While selecting data structure from different available structures user needs to check applicability of structure to problem as well as data used in problem. At the same time efficiency of the algorithms and arrangement of data is also considered. During execution every algorithm uses different resources which are competitive in nature. Every process (algorithm in running state) needs resources for completion. Space (main memory) and time (processor‘s time) are these two resources, regarding which efficiency of algorithms and structure are considered. Efficiency of any structure is represented as a mathematical function, which is executed depending on size of space and number of iterations to be executed while completing the process, and is termed as ‗Complexity of Structure‘. Traditionally for analyzing any structure and its algorithms asymptotic notations were used. The asymptotic notations considers upperbound and lower bound values for algorithm. These notations helps to calculate complexity by considering single operation. In case of combination of multiple operations it is better to consider complexity of all algorithms executed sequentially. Amortized analysis is another technique which helps to calculate the average case complexities. For the structure developed by the researcher, goal is to represent big data and making it available to the process with minimum searching time and memory usage. When these two challenges are considered, the data is divides in two parts, Main Structure and Pool Structure. Both the structures are designed using different available traditional structures. Space complexity for any data structure is denoted by the size of space or memory occupied while its implementation. Here space occupied by different components of the process is counted in terms of measures used to count memory. These components are given as –

- Variables used in code
- Heap size used by program to store data and different function calls
- Size of input data
- Data

With respect to every component of the process memory required for it is discussed in this section. As discussed about the platform, ‗c‘ has been used for implementation. When size of code developed using ‗c‘ is considered, it needs very less space. For time complexity following code blocks are considered –

- Number of comparisons to be carried out
- Iterations for searching required data

Finally, The structure was found to be efficient as it is based on traditional structures. The structure has less space requirement as wee as less time complexity. It is proved that the data represented in the pool structure is loss less representation of original data

REFERENCES

G. Mohamed and N. B. Ithnin, “Survey on Representation Techniques for Malware Detection,” System American Journal of Applied Sciences, 2017.

Praveen Sundar, P.V., Ranjith, D., Vinoth Kumar, V. et al. Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. Int J Speech Technol (2020). https://doi.org/10.1007/s10772-020-09686-y.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V. et al. New and robust composite micro structure descriptor (CMSD) for CBIR. International Journal of Speech Technology (2019), doi:10.1007/s10772-019-09663-0.

Karthikeyan, T., Sekaran, K., Ranjith, D., Vinoth kumar, V., Balajee, J.M. (2019) "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques", International Journal of Web Portals (IJWP), 11(2), pp.41-52

Vinoth Kumar, V., Arvind, K.S., Umamaheswaran, S., Suganya, K.S (2019), "Hierarchal Trust Certificate Distribution using Distributed CA in MANET", International Journal of Innovative Technology and Exploring Engineering, 8(10), pp. 2521-2524.

Maithili, K , Vinothkumar, V, Latha, P (2018). "Analyzing the security mechanisms to prevent unauthorized access in cloud and network security" Journal of Computational and Theoretical Nanoscience, Vol.15, pp.2059-2063.

V.Vinoth Kumar, Ramamoorthy S (2017), "A Novel method of gateway selection to improve throughput performance in MANET", Journal of Advanced Research in Dynamical and Control Systems,9(Special Issue 16), pp. 420-432

Dhilip Kumar V, Vinoth Kumar V, Kandar D (2018), "Data Transmission Between Dedicated ShortRange Communication and WiMAX for Efficient Vehicular Communication" Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654.

Kouser, R.R., Manikandan, T., Kumar, V.V (2018), "Heart disease prediction system using artificial neural network, radial basis function and case based reasoning" Journal of Computational and Theoretical Nanoscience, 15, pp. 2810-2817.

Shalini A, Jayasuruthi L, Vinoth Kumar V, "Voice Recognition Robot Control using Android Device" Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2197-2201

Jayasuruthi L,Shalini A,Vinoth Kumar V.,(2018) " Application of rough set theory in data mining market analysis using rough sets data explorer" Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2126-213

E. Bou-Harb, M. Debbabi, and C. Assi, "Cyber scanning: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 16, no. 3, pp. 1496–1519, 2014. \

N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen, "SocialHelix: visual analysis of sentiment divergence in social media," Journal of Visualization, vol. 18, no. 2, pp. 221–235, 2015.

Deepak Gupta and Rinkle Rani, "Big Data Framework for Zero-Day Malware Detection", Cybernetics and Systems, DOI: 10.1080/01969722.2018.1429835,2018. Sitalakshmi Venkatraman andMamounAlazab, "Use of Data Visualisation for Zero-Day Malware Detection", Security and Communication Networks, Article ID 1728303, 13 pages, 2018.

ISBN: 978-93-5782-920-5, www.aetsjournal.com "Handbook of research in Big data Analytics, Artificial intelligence and Machine learning"