International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
©2017 INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

Indexing Techniques and Challenge in Big Data

Mamta Mittal*

G.B. Pant Govt. Engineering College, GGSIPU, New Delhi, India

Accepted 20 May 2017, Available online 25 June 2017, **Vol.7, No.3 (June 2017)**

Abstract

Evolution of unstructured and partially-structured data immersed us in the Information Age where vast amounts of data are available from internet sources. This abundant and heterogenous data should be analyzed properly. So, there is vital requirement of techniques which deal with such Big Data and emerged as new era to handle structured and unstructured data effectively and efficiently. Here, indexing techniques plays major role to access and process query at a faster rate. Major focus of this paper is to discuss various indexing techniques such as Database cracking, HAIL, LIAH, etc. as well as what are the challenges data analyst faces when such Big Data is to be analyzed.

Keywords: *Big Data, Indexing, Analytics, Hadoop*

1. Introduction

Everyday trillion, quintillion bytes of data are produced in the world. The rate of progress is exponential which necessitate us to analyze such a gigantic data effectively and efficiently. Except gigantic this data is heterogenous in nature also (Laney 2001). It is difficult to effectively extract useful information from all the available online information due to the volume and variety (structured/unstructured) of data. However, the foundation of a good analytical framework relies totally on the quality and quantity of data; if we have rich dataset then inferences would be better. Big Data analysis plays import role here (Sagiroglu *et al* 2013). Still there is need of techniques which can access and search data item speedily. Indexing strategy is used to design an access method to a searched item. It also describes how data is organized in a storage system so that information retrieval can be carried out (Chen *et al* 2013).

The remaining paper has been structured as follows. Indexing techniques for Big Data have been presented in Section 2. Challenges exist in handling Big Data have been presented in Section 3. A brief conclusion has been given in last section.

2. Indexing techniques for Big data

The idea of Big Data indexing is to fragment the datasets according to criteria that will be used in the query. The fragments are indexed with each value satisfying some query predicates.

A. Popular indexing technique in Big Data is *adaptive indexing* where indexes are created based on query processing. Here, the system creates the index for a given attribute based on a single incoming query.

An approach for adaptive indexing is *Database Cracking*. In Cracking, index is built adaptively during query processing based on the query. Thus, workload knowledge and idle time are not required. Database cracking uses THREE PIECE CRACK algorithm on the copy of a column for the first query and TWO PIECE CRACK algorithm for the rest queries. This has the benefits that the original column remains intact and there is no overhead of maintaining the complete table. Database cracking is depicted in Figure 1.

In Figure 1, following queries are there:

Query1: select * from R where A>10 and A<14
Query2: select * from R where A>7 and A>=16

As Figure 1, shows that on copied column three piece crack algorithm is used resulting in three pieces ranging A<=10, 10<A<14, A>=14 and next range predicate (Zoumpatianos *et al* 2014). Then, two piece crack algorithm has been applied resulting in 5 more pieces. This type of data cracking is known as Selection Cracking (Schuhknecht *et al* 2012).

*Corresponding author: **Mamta Mittal**

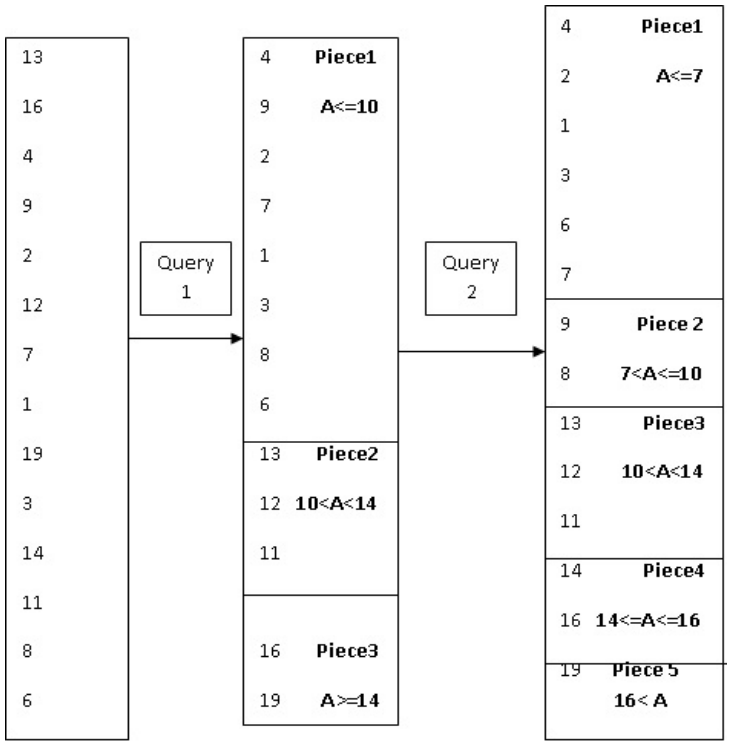

Figure 1 Database Cracking

A) Indexes in Hadoop

- *Hadoop ++:* Here, an index structure called trojan *index* is introduced and where 10 user defined functions of hadoop are modified. So, separate indexes for each HDFS block are created. The query runtimes of Hadoop++ are faster than Hadoop by a factor of 20. The time spent for creating these indices is very high (Gandomi *et al* 2015).
- *HAIL (Hadoop Aggressive Indexing Library):* A clustered index is built for each HDFS block. As each HDFS block has a imitation feature of three, So three different indices are present. From these, a suitable index can be found. HAIL has a disadvantage that its performance can only be improved if indexes on the correct attributes are constructed.
- *LIAH (Lazy Indexing and Adaptivity in Hadoop):* It creates adaptive index based on the query. Therefore no indexes are constructed beforehand. Indexing is carried out parallel with the ongoing map reduce tasks.

3. Various Challenges in Big Data

The large volume of Big Data brings lots of challenges like data acquisition, storage, management, analysis and data security. Traditional Data management systems are used for management and analysis of structured data and not for unstructured or semi structured data (Khan *et al* 2014). Also RDBMS is not able to handle the high volume and heterogeneous structure of Big Data. In Big Data, large amount of various data types must be analysed for decision making (Mukherjee *et al* 2016).

The challenges with Big Data can be categorised into: data, process, and management, analytics and security. The following section describes these challenges in detail:-

A. Data Challenges

- *Data representation*: In Big Data, datasets have some level of heterogeneity in type, semantics and structure. For user interpretation and computer analysis, data needs to be
- meaningful and understandable. If data representation is improper, this will decrease the value of the data and will further affect the data analysis.
- *Reducing redundancy and data compression*: The sources of Big Data are sensors, mobile phones and social networks. Hence the data from these sources is highly redundant, which should be compressed. Compression is needed so that the value of the data is not affected and this will reduce the indirect cost of the system.
- *Storage of data*: Sensors, social media, mobile devices and other systems generate data at unprecedented rate. Due to this massive amount of data, there are number of challenges. One of the

challenges is that the storage systems used currently cannot store such large amount of data. So decision needs to be taken which data is to be kept and which data needs to be discarded.

- *Heterogeneity:* Sources of data are smart devices, social networking sites, sensors etc. A lot of this data is raw, unstructured, and semi-structured. Data mining algorithms can be used for analysis f or structured data. To perform the analysis of unstructured or semi-structured data, it should be properly structured before its analysis. For example, in hospitals each patient can have number of tests reports of which can be in different formats.
- *Scalability:* Other Challenges in data analysis includes the storage and analysis of large amount of data and the speed at which data is increasing. As a result, for Big Data analysis lot of navigation over a large search space is required.

B. Process Challenges

Finding the right model for analysis of Big Data is necessary. The process challenges include-

- Data acquisition from different sources.
- Proper alignment of data generated from different sources as it is important to resolve when two objects are same.
- Transforming or filtering the data in to a form which is suitable for analysis.
- Modeling of data, using some simulation technique
- Understanding the output, visualizing the output and sharing of results (Roberto 2012).

C. Management Challenges

Many datasets contain sensitive data. To access such data, there are many legal and ethical concerns. The challenge for accessing such data is to abide by its relevant laws. The main challenges for management are

- Data privacy
- Ethical concerns
- Security of Personal data
- Legal concerns

D. Knowledge of various techniques for Analyzing Big Data

To analyze data effectively and efficiently knowledge of various techniques is required. These techniques are: Association Rule Learning, Classification Tree Analysis, Genetic Algorithms, Machine Learning, Regression Analysis.

E. Security

Another challenge in Big Data is the security of data. As Big Data is growing, the security and privacy problems are also increasing. There is large collection of heterogeneous data which is shared amongst scientists, businesses, government and citizens. With the growing size and variety of data from online services and mobile devices, there is a need to control of access to such information. The three privacy mechanisms that need to be addressed are access control, auditing, and statistical privacy. However, the tools and technologies developed to manage this data do not incorporate adequate security or privacy measures, thus there is a need to update the current approaches to prevent data leakage.

Conclusion

Big Data requires knowledge of multiple disciplines and is used in every field like computer, astrophysics, business, food industry, telecom industry, health sector and many more. In this paper, the authors have discussed Indexing techniques and challenges have been discussed. Indexing techniques are always be helpful to access data speedily, thus when volume of data is so high, indexing plays important role in accessing data timely. Existing challenges motivates the readers to manage, store and analyse this massive amount of data with various security issues and techniques.

References

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, *2014*.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Mukherjee, S., & Shaw, R. (2016) Big Data–Concepts, Applications, Challenges and Future Scope. *International Journal of Advanced Research in Computer and Communication Engineering Vol.* 5, Issue 2.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*, 70.

Zoumpatianos, K., Idreos, S., & Palpanas, T. (2014). Indexing for interactive exploration of big data series. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 1555-1566). ACM.

Roberto V. Zicari. (2012) Big Data: Challenges and Opportunities. *Big Data Computing Pages* 104-128

Schuhknecht, F. M., Jindal, A., & Dittrich, J. (2013). The uncracked pieces in database cracking. *Proceedings of the VLDB Endowment*, *7*(2), 97-108.