International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.3, pp.129-131, June (2017)

Review Paper

E-ISSN: 2320-7639

Various Chunking and Deduduplication Techniques in Big Data

[1]Naresh Kumar and [2]Ishu Devi*

[1]CSE Department, UIET-Kurukshetra University, Kurukshetra, India
[2*]CSE Department, UIET, Kurukshetra University, Kurukshetra, Haryana, India

Corresponding Author: ishupunia81@gmail.com

Available online at: www.isroset.org

Received 12th May 2017, Revised 24th May 2017, Accepted 18th Jun 2017, Online 30th Jun 2017

***Abstract*—** In today's environment very huge amount of data is generated with duplication. This huge amount of data is called big data. To handle this kind of big data and reduce duplicity from data chunking and deduplication mechanism is used. In deduplication mechanism duplicate data is removed by using chunking and hash functions. In this paper an attempt has been made to converse different chunking and deduplication techniques. A comparative analysis of these techniques with different pros and cons has been presented.

***Keywords*—** Big Data, Chunking, Deduplication, FBC (Frequency Based Chunking) and CDC (Content Defined Chunking)

I. INTRODUCTION

In today's modern civilization everything or device is digitized. No one can survive without this digital space. This digitization creates lots of data in Exabyte. This data is in the form of unstructured or redundant data. The main issues in big data were that how to maintain this large amount of data, who maintains it and where to store this big data. Indeed, even today Facebook, twitter and other long range social communication site creates extensive measure of information in consistently. In this way, there is extensive degree on big data. As there are many methods and advances to break down the information. There are different methods to investigate the big data i.e Data de-duplication techniques. As when the information is expanding then there are more possibilities that information is rehashed or excess that necessities to prerequisite of huge stockpiling devices.

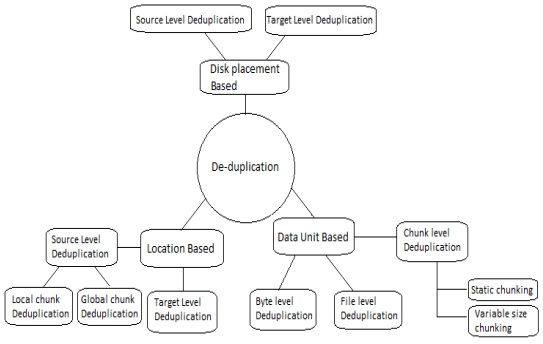

Fig1: Deddupliction Techniques [3]

It additionally expands the preparing time of the query. So there is an extensive degree to break down the information by information deduplication systems with the goal that can decrease the necessity of expansive data and diminish the handling time to execute the problem [2].

The process of elimination of related data or duplicate data is called deduplication process. It is divided into three parts data based, location based and disk based. They are further divided into subcategories as shown in fig 1 [3].

A. Types of Deduplication

Data Unit Based: this technique work on data units and it is further divided into three types. These are as follows:

- Byte Level De-duplication: In byte level deduplication data is processed in the network in the form of bytes. Each byte of data is to be checked separately to find duplicate data and if duplicity is find then that data will not be stored and discarded otherwise it will be stored.[4].
- File Level De-duplication: In this technique deduplication is processed on files. A hash value of each file is generated and if this hash value of two files is matching then one of them is duplicate and system will discard duplicate file and if both the files have different hash value then these file is to be said as unique files [5].
- Chunk Level De-duplication: In this type of deduplication data is divided into blocks called chunks. Here each chunk have its own unique id and if two chunks have same id then both of these chunks contains same data and one of them is necessary to be discarded [6].

Location Based: In location based deduplication redundancy of data can be eliminated depending on location of data.

- Source Level De-duplication: Here in source level deduplication removal duplicate data is performed at sender level from where data is to be send. It can be further divided into two subparts. These are:
- Local Chunk De-duplication: In this deduplication, dedupliation is performed on local sender before sending it to destination.
- Global Chunk De-duplication: In this de-duplication method, duplicate data is removed globally for all senders.
- Target Level De-duplication: When duplicate data is eliminated at destination from where data is to be stored then it is called target level data de-duplication. This process may increases processing time but meanwhile also increases bandwidth.

Disk Placement Based: when deduplication is performed on disks then it is called disk level deduplication and it is off two types forward reference or backward reference.

- Forward reference: In this technique current data chunks are maintained and the total old data chunks are associated with pointers that points forward to the current chunks.
- Backward reference: It creates more fragments of past chunks [6].

This paper is organized in four sections. Section I covers introduction of big data, deduplication and its types, in section II related work of various existing papers are presented with their drawbacks, section III shows comparative analysis of various chunking techniques at last section IV presents conclusion and future work of paper.

II. RELATED WORK

In [7] made a Content Based File Chunking model structure which melded a CPU chunking subsystem and GPGPU subsystem.

In [8] proposed a novel chunk-based de-duplication methodology which could stay away from a summary in RAM and in this way evade all RAM use basic to hold an once-over. They kept up various canisters in plate, which basically contained distinctive chunk IDs.

In [9] concentrated on indicating chunking and its effect on recovery, learning, and adequacy.

In [10] displayed deduplication gathering. The Approach of joining comparability with area was related with the deduplication gathering.

In [11] introduced DARE, a deduplication-cautious, low-overhead likeness disclosure and end plot for delta weight on the most essential reason for deduplication on stronghold datasets.

In [12] proposed POD, an execution orchestrated deduplication arrangement, to update the execution of central stockpiling structures in the Cloud by utilizing information deduplication on the I/O way to deal with evacuate excess make demands while moreover sparing storage room.

III. COMPARATIVE ANALYSIS

Here in this section comparison of different chunking techniques has been presented. In table1 mapping of different chunking techniques are represented. To map these chunking techniques various application of big data are used named File synchronization, backup, storage and data retrieval. This mapping can be done by reviewing the various research papers of these techniques. And conclude that FBC used for back up, storage and data retrieval.

Table 1: Mapping of chunking techniques to Big Data application[13]

Techniques	**Applications**			
	File Synchronization	Backup	Storage	Data Retrieval
FBC(Frequency Based Chunking)	No	Yes	Yes	Yes
CDC (Content Defined Chunking)	No	Yes	No	Yes
Byte-Index Chunking	No	Yes	Yes	No
Multi –level Byte Index chunking	Yes	Yes	No	No

V. CONCLUSION

Big data contains data in unstructured form and have redundant data. To manage this unstructured data and reduce duplicity from data various chunking techniques and deduplication techniques has been used. In this paper different deduplication techniques with their pros and cons has been discussed. After that a comparative analysis of different chunking techniques in perspective of application areas of big data has been presented. In future try to propose a novel mechanism based on existing mechanism to improve deduplocaton ratio and reduce chunk generation time.

REFERENCES

[1] M. Dirk, "*Advanced data deduplication techniques and their application*", Ph.D. dissertation, Universit¨ at sbibliothek Mainz, pp.1-6, 2013.

[2] M. Dirk, K. J¨urgen, B. Andre, C. Toni, K. Michael, K. Julian, "*A study on data deduplication in hpc storage systems*", in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, USA, pp.1-7, 2012.

[3] Chi Yang, Jinjun Chen, "*A Scalable Data Chunk Similarity Based Compression Approach for Efficient Big Sensing Data Processing on Cloud*", IEEE Transactions on Knowledge and Data Engineering, China, pp.1144-1157, 2017.

[4] R. Tuchinda, C. Knoblock, P. Szekely, "*Building data integration queries by demonstration*", Proceedings of the 12th international conference on Intelligent user interfaces, USA, pp. 170-179, 2007.
[5] Q. He, X. Zhang, Z. Li, "*Data deduplication techniques*", 2010 International Conference on Future Information Technology and Management Engineering (FITME), , CA, pp. 430-433, 2010.
[6] A. Banu and C. Chandrasekar, "*A survey on deduplication methods*", International Journal of Computer Trends and Technology, vol.3, no.3, pp. 364-368, 2012.
[7] Zhi Tang, Youjip Won, "*Multithread Content Based File Chunking System in CPU GPGPU Heterogeneous Architecture*", 2011 First International Conference on Data Compression, Communications and Processing, China, pp. 58-64, 2011.
[8] Zhike Zhang, Zejun Jiang, Zhiqiang Liu, Cheng Zhang Peng, "*LHS: A Nobel Method Of Information Retrieval Avoiding An Index Using Linear Hashing With Key Groups In Deduplication*", Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, China, pp.1312-1318, 2012.
[9] Duane F. Shell, Leen-Kiat Soh, Vlad Chiriacescu, "*Modeling Chunking Effects on Learning and Performance using the Computational-Unified Learning Model (C-ULM): A Multiagent Cognitive Process Model*", IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing, India, pp. 77-85, 2016.
[10] Xingyu Zhang, Jian Zhang, "*Data Deduplication Cluster Based on Similarity- Locality Approach*", IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, CA, pp.2168-2173, 2013.
[11] Wen Xia, Hong Jiang, Dan Feng, Lei Tian, "*Combining Deduplication and Delta Compression to Achieve Low-Overhead Data Reduction on Backup Datasets*", Data Compression Conference, France, pp. 203-212, 2014.
[12] Bo Mao, Hong Jiang, Suzhen Wu, Lei Tian, "*Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud*", IEEE Transactions on Computers, NY, pp.1-14, 2015.
[13] Sonali D. Chaure, M. U. Kulkarni and Pankaj M. Jadhav, "*Web based ETL Approach to Transform Relational Database to Graph Database*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.7, pp.92-97, 2015.

