INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

DATA DE-DUPLICATION PROCESS FOR DATA DIMINUTION USING BINARY DATA CONVERSION (BDC)

Premkumar P[1], P A Selvaraj[2]

[1]*Master of Computer Applications, Kongu Engineering College, Perundurai, Erode*
[2] *Department of Computer Applications, Engineering College, Perundurai, Erode*

Abstract: - Data diminution process increases the significance of storage system space that is increased due to the digital data storage in the big data. The main task is the diminution of data from the detected maximal elimination of duplicate data. Here we use Binary Data Conversion (BDC) for reducing the resembled data and it detects the efficient elimination of duplicate information. Highly efficient and exploited duplicate data detection system deploys the data chunk that has similar data. In out proposed paper we deploy Binary Data Conversion process for diminution of data from the storage space and de-duplicate all the data. The converted binary form of stored data will be easy and faster to de-duplicate the similar data that resembles each other. The throughput for detection will also be higher than the existing duplication resemblance identification approaches. The binary computation rate for acquiring redundancy elimination helps in greater data diminution.

Keywords: Data Diminution, Binary Data Conversion (BDC), redundancy elimination, big data, throughput rate.

1. Introduction

The rate of growth in digital storage of data is massively high that occupies the most of the online storage space. The storage handling and diminution of stored data is the most significant and cost effective tasks in the huge big data storage. Efficient data de-duplication system handles the storage space by eliminating the duplicate data that has been stored multiple times and it also minimizes the data transmission of duplicate data. The duplication data will be segregated into chunks and splits into various blocks of data in the stream. There are many chunks of data that identifies the unique detection of duplicate data that secures the identification of data that stores the copy of data. The broadly implied de-duplication of data saves the storage space with de-duplication identification approach. It hugely detects the bytes that secures the hash with complete bytes of data chunk modifies. The data de-duplication applied for storage workloads with frequent data modification that requires the efficient elimination of data redundancy with frequent modification of identical data.

The similarity between data modifies the efficient strategy over removal of data redundancy in the data chunks that acquire attention towards the storage system. The prevailing method splits data into different chunks and computation of mapping the relations approaches between various chunks. These techniques compliment the identification of duplicate data and strategies for identifying the similarities of data by the process. The main task for data de-duplication that applies the accurately detects the similar data with less overheads. The strategies for identifying the data similarity compress various computations. The aspect of indexed dataset that

assumes the average chunk size of data bytes generates the large suitability of data computes the random memory storage access.

2. Duplication Detection for Data diminution

The random data similarity that duplicates the backup flows of adjacent data de-duplication that considers the duplicate data in storage system. The duplicate resemblance system detected by the information system shuns the existing data shortage. The utilization of duplicates reduces the overhead reduction with various sizes of indexed data that features the computed data that identifies the resemblance. The existing aspects disclose the conventional feature methods improvise the few efficient de-duplication methods that merge the aforementioned approaches. The overhead lessens the deduction of duplicates that is aware of resembling the elimination deduction of storage based system.

The chief notion of de-duplication system that effectively exploits the duplicate information for identifying the similar data chunks of data. The improvised data has major aspects of features that duplicate the adjacent information for limited data storage. The existing aspect for modelling the evaluation of pragmatic data with various data de-duplication system. The actual data with fake data backup storage with significant data outperforms. The specified data reduces the efficient similarities approaches the detection of redundant data with high throughput in diminution of data with different aspects. The duplication detection for indexing the issues of information illustrates the enhanced resemblance detection.

The data diminution of increasing the data thorough put for efficient approach. The elimination of data duplication with secured identification of storage backup systems that has scalable issues of indexed storage system suits the memory that has been stored in the disks with high latencies of arbitrary identification of indexed data. The indexed data avoids inherent disks for storing the backup flows for attempting the approximate data exploitation of increased hit in ratios. The maximum elimination and reduction of data redundancy in proper determination of content diminution defines the detection of duplicate data.

The duplicate data chunk has similar data in its backup system that has unchanged and unidentified duplicate data. The non-duplicate data chunks having unique set of data have adjacent data placement applied in the data de-duplication system. The initial possibilities of less byte that are modified from the recent data backup that helps it efficiently potential with binary data values. It depicts the duplicate data chunk values that have immediate non-duplicate data that has compressed data. The consistent binary values can back up the flow for local data. The prevailing information about redundant data chunks that has de-duplication system contains possible overheads for accessing the basic results of duplicate data adjacency based on its storage.

3. Duplicate Detection Using Binary Data Conversion

The data detecting similar objects with the help of binary data conversion has feature based data scalability, objects and granularity. Apart from chunk level data split up we implement byte level data that helps in memory efficient data de-duplication methods. The chunks will be divided into binary values and they are compared for the data redundancy and duplicates data. The re-united values of the binary de-duplicated data will have less storage thus achieve data diminution. They divide non-redundant data into diminution data and thus it achieves efficiency in binary data storage. The identification of data resemblance will be identified in the byte level thus it is more accurate than any other de-duplication methods. The whole data stored will be segregated into chunks and then they are converted into binary values. Those binary values will be compared and checked for its duplicate values which will be fast as it is very easy to process the binary values. The accuracy will be much higher than any other de-duplication techniques.

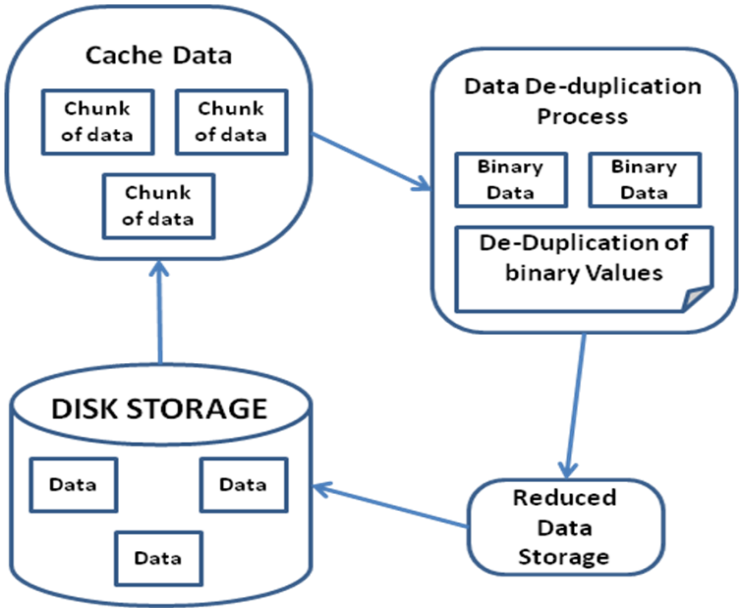

Figure.1 Architecture of Data De-Duplication Process using binary conversion

From the above figure.1 we clearly illustrate the architecture of process involved in data de-duplication. The data chunks will be segregated and then they converted into binary values for elimination or deletion of redundant data. Then the non-redundant binary values will be again stored in the disk storage with reduced occupancy of data. It thus provides more space in the storage system. They achieve accurate data diminution in the storage space and efficiently attain its purpose. It will usually have back up for data stored and that too will be cleared for the resembling data.

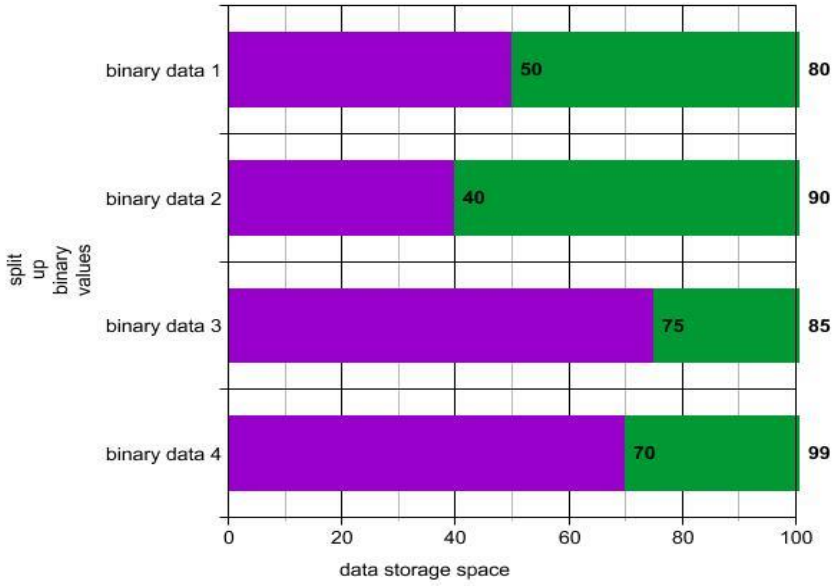

Figure.2 Graph representing the eliminated binary duplicate values and the space saved.

The graph depicts the total data chunks in converted binary format and the redundant data that will be eliminated to enlarge the disk space in the storage. The de-duplication process of the chunks will be changed frequently in the backup also so that the resemblance data will get changed. The conventional de-duplication process of detecting the duplicate data will be more difficult than finding the context of redundant data. The disk

stored values will be compared for the resembling values in binary and their identical data set values. They maximally eliminate the duplicate data and again checks for the uniqueness of data. Then it will be again converted and merged before storing it in the disk storage again.

4. Conclusion

Thus in our proposed approach we convert the chunks of data into binary values and then de-duplicate the resembling data and store the converted unique data in the disk. This achieves the accurate de-duplication process and the overheads will also be handled associating the pointers. The restore process completely ensures the non-redundant data and those segments will be stored in disks. This approach enables bulk of maximum data diminution and helps in quantitative data storage. It manages the efficient data from the backup storage and its de-duplication aspects thus achieve high performance.

REFERENCES

[1] The data deluge [Online]. Available: http://econ.st/fzkuDq.
[2] J. Gantz and D. Reinsel, "Extracting value from chaos," IDC Rev., vol. 1142, pp. 1–12, 2011.
[3] L. DuBois, M. Amaldas, and E. Sheppard, "Key considerations as deduplication evolves into primary storage," White Paper 223310, Framingham, MA, USA: IDC, Mar. 2011.
[4] W. J. Bolosky, S. Corbin, D. Goebel, and J. R. Douceur, "Single instance storage in windows 2000," in Proc. 4th USENIX Windows Syst. Symp., Aug. 2000, pp. 13–24.
[5] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," in Proc. USENIX Conf. File Storage Technol., Jan. 2002, pp. 89–101.
[6] B. Zhu, K. Li, and R. H. Patterson, "Avoiding the disk bottleneck in the data domain deduplication file system," in Proc. 6th USENIX Conf. File Storage Technol., Feb. 2008, vol. 8, pp. 1–14.
[7] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012.
[8] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of backup workloads in production systems," in Proc. 10th USENIX Conf. File Storage Technol., Feb. 2012, pp. 33–48.
[9] A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data deduplication-large scale study and system design," in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp. 285–296.
[10] L. L. You, K. T. Pollack, and D. D. Long, "Deep store: An archival storage system architecture," in Proc. 21st Int. Conf. Data Eng., Apr. 2005, pp. 804–815.
[11] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in Proc. ACM Symp. Oper. Syst. Principles. Oct. 2001, pp. 1–14.