November-December 2019
ISSN: 0193-4120 Page No. 2452 - 2455

A Study on Improvement of Data Analysis Process through Smart Compression

Cheolhee YOON[1], Jang Mook KANG[2] Jung joong KIM[3]
Police Science Institute, Asan Republic of Korea E-mail: bertter@police.ac.kr
Global Cyber University, Seoul, Republic of Korea. E-mail: honukang@gw.global.ac.kr
eGlobal Systems, Seoul, Republic of Korea. E-mail: john.kim@eglobalsys.co.kr

Article Info
Volume 81
Page Number: 2452 - 2455
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *12 December 2019*

Abstract
Smart Compression applies the appropriate compression algorithm for each column, taking into account the maximum compression and decompression rates. The compression algorithms used here are LZ4, RLE, PFOR, Delta encoding on top of PFOR, and Dictionary encoding. The vector - based compression algorithm used in this paper provides 4 to 6 times higher compression efficiency than the commonly used compression algorithm. To maximize Input and Output performance, we run CBM (Column Buffer Manager) which replicates the contents of the disk in the main memory area. In addition, The decompression is processed not in the memory but in the vector mode immediately before the data is processed in the CPU cache, and decompression proceeds. Therefore, it is efficient in terms of improving the data analysis process. Smart compression is derived to use large capacity manufacturing data in smart factory.

Keywords*: Smart compression, Algorithm, data analysis*

1. INTRODUCTION

Data from manufacturing sites are very valuable. In particular, the smart factory called the future manufacturing site presented a model that sought to improve productivity and increase efficiency through integration of the entire process, ranging from order and design to production distribution. In particular, manufacturing big data that will be mass-produced from Smart Factory is being used as an essential element through analysis and use. The large amount of data generated by all manufacturing processes is analyzed in real time and shall be intelligently optimized for manufacturing processes. Intelligent data analysis was attempted by vector analysis and big data analysis method called smart compression was designed and tried. Recently various efforts to improve manufacturing have been made worldwide, and the new manufacturing industry, which has been reorganized as a smart factory, is evolving into a data-analyzed information-driven system based on the latest technologies of big data and the Internet of Things. Various types of smart data compression functions were attempted by considering the data analysis requirements and efficiency that should be applied at smart factories through automated facilities and information service systems at manufacturing sites

2. DATA IN THE MANUFACTURING PROCESS

A. *Preparations of Smart Compression*

There are many sensors in the production facilities of smart factory. So, what data to collect from which sensors has emerged as an important issue. In addition, sensor data generated from sensors that manage the underlying equipment collect most of the data from all processes, and even ultra-precision data is collected to detect any slight variation in the process. That's why Smart Compression is important to collect large-scale data. These large-scale data are used as meaningful data in terms of efficiency, although the data analysis process should proceed in a scaled-down format. The various data generated by manufacturing phenomena are processed and processed in various ways and are used as important information to support decision making. In the paper, the organization of data infrastructure at the manufacturing level and management and data processing were more important than the role of such intelligence. Measures were taken to enhance data utilization from a smart-factoring perspective by supplementing problems and limitations with data collection, data utilization, and ICT infrastructure. The organization of software systems and related hardware devices was also considered for use. Figure 1 processes the process of finding

November-December 2019
ISSN: 0193-4120 Page No. 2452 - 2455

finally hidden information using smart compressions of data[1].

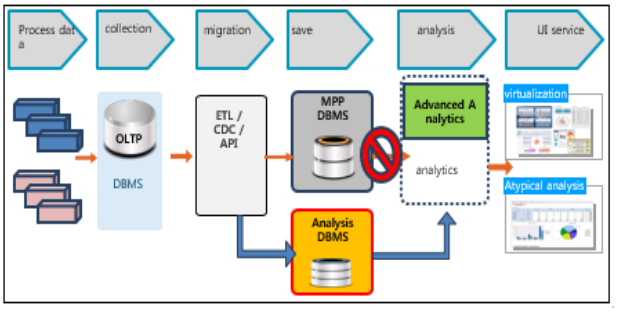

Figure 1. Process of analysis

B. *Establishment of process for application of Smart Compression*

Depending on the analysis information requirements at the site, large volumes of data can be designed to correspond to various analysis methods. The high-speed analysis results are then derived and immediately used as processing data used for OLAP, statistical packages. In other words, without creating a data mart and without generating an index for performance improvement, SQL enables collection storage data to be analyzed in seconds in a manufacturing state and the resulting environment to be derived. This configuration will allow analysis of manufacturing information, integrated contact history and various information collected from numerous channels to be compressed and continuously provided with products suitable for the situation. Additional technical elements are data processing via SIMD processing and on-chip cache computing by executing a single command on a data set of x86 chips through Vector Processing. This will allow compression to be performed within the CPU to maximize throughput through Smart Compression[2]

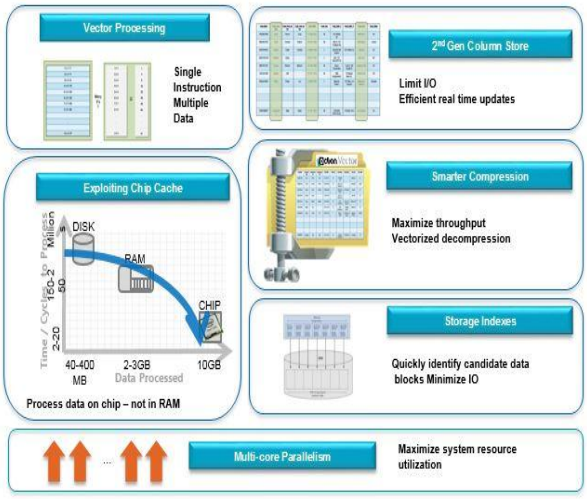

Figure 2. Data Compression

3. MANUFACTURING PROCESS BY SMART COMPRESSION

A. *Smart Compression*

Smarter Compression can apply the appropriate compression algorithm in each column considering the maximum compression rate and decompression rate, and the compression algorithm used is LZ4, RLE, PFOR, Delta encoding on top of PFOR, and Dictionary encoding. Typically, the compression algorithm used is highly compression efficient, and to maximize IO performance, will operate a column buffer manager (CBM) that duplicates the contents of the disk in the main memory area.

For analysis, decompression is performed by Vector processing just before data is processed in CPU Cache, rather than when data is in memory, making it a high-speed analysis in environment. Fig3,4is a configuration for smart compression

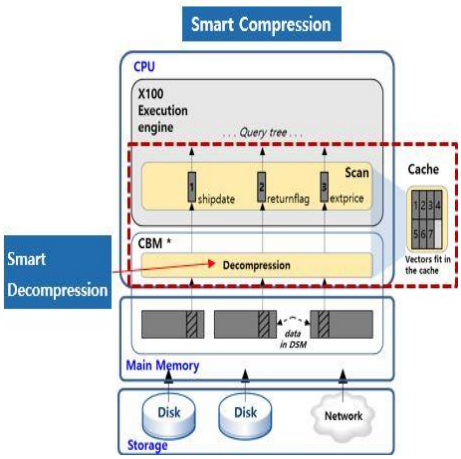

Figure 3. Smart Compression

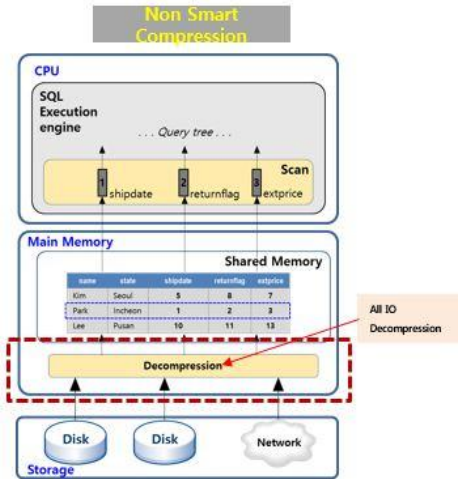

Figure 4. Non Smart Compression

This generally results in performance differences in all aspects of the system being decompressed and analyzed from

November-December 2019
ISSN: 0193-4120 Page No. 2452 - 2455

memory. Figure 3 "No Smart Compression" shows a process that is uncompressed in main memory area as a model that does not apply Smart Compression.[4][

B. *Test of Smart Compression*

The table below shows the compression ratio of 10 billion data capacities for Line-item in TPC-H. The capacity of 10 billion files and vectors was compared to Line-item of TPC-H. The size of the file is 1,706GB, while the capacity when it was compressed by a vector is 380.1GB, indicating that it was compressed to about 22.28% of the total file size.

Table 1. Performing compression on 10 billion records

The number of records	file capacity	vector capacity	Ratio (Vector/File) * 100
10 billion cases	1,706GB	380.1GB	22.28%

It was also possible to verify that the post-compression capacity of the log data for testing is compressed as shown in the following table. That is, the daily data logging volume is about 890 million, the file size is 722GB, and the size on the Vector Type DBMS is 106GB. This is stored at a rate of about 14.68% and corresponds to an average compression rate of seven-times.[3]

Table 2. Performing compression of log data

Log data	Data Size (Apply Vector)
722GB	106GB
5TB	711GB
10TB	1.4TB
20TB	2.8TB
30TB	4.2TB
50TB	7TB

4. CONCLUSION

Smart Compression is already a necessity, not an option, for all manufacturers, and the various data generated from manufacturing phenomena are processed and processed in various ways to serve as important information to support the decision making of the manufacturer. [5] For the utilization of these data, the problems and limitations of data collection, data utilization and infrastructure composition have been put forward in this paper. Data is utilized through smart compression, and all data in the manufacturing process, such as quality defect or equipment failure, can be collected and identified, and the results can be efficiently processed. The Smarter Compression presented in this paper can apply appropriate compression algorithms in each column to account for maximum compression rate and decompression speed, and to operate a column buffer manager (CBM) that duplicates the contents of the disk in the main memory area to maximize IO performance. For analysis, decompression is performed by Vector processing just before data is processed in CPU Cache, rather than when data is in memory, making it a high-speed analysis. It is considered to be a technology that must be used for future process real-time monitoring, product defect analysis, and facility-related data analysis for abnormalities detection[7-16]

In this paper, we designed and tried a big data analysis method called smart compression. Various efforts to improve manufacturing have been made worldwide in recent years, and the new manufacturing industry, which has been restructured into a data-analyzed information-driven system based on the latest technologies of big data and the Internet of Things, has attempted various forms of smart data compression by considering the site. Smart compression was derived to derive a smart factory model for analysis and use of large-capacity manufacturing data in context, and measures were proposed to accelerate the efficiency of data analysis by providing analysis scenarios for utilizing the platform [18-20].

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP). (No.2018-0-00705, Algorithm Design and Software Modeling for Judge Fake News based on Artificial Intelligence)
This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03038738, Algorithm Design & Software Architecture Modeling to Judge Fake News based on Artificial Intelligence)

REFERENCES

1. Chan mo Jeon et al. "Developement of integrated operation technology for smart factory application linked to large-scale manufacturing data", Korean CDE Conference, 2016.1, 219-230(12 pages)
2. Yang Jin-kyung, Lee Dong-hee, Lee Cho-hee and Kim Kwang-jae.. "A study on the selection of data-based core semiconductor manufacturing processes" Proceedings of the Korean Institute of Industrial Engineers Conference, 2017, 1229-1248
3. Park Hoon-seok, Oh Kyu-hyeop, Kim Ae-kyung, Berny Alfonso, Josue Obregon."Development of manufacturing big data analysis library for upgrading smart factories". Korean Society of Management Science Conference Proceedings,2018, 1527-1541
4. Peter Boncz, Marcin Zukowski, and Niels Nes.MonetDB/X100: Hyper-pipelining query execution. In CIDR, 2005.

November-December 2019
ISSN: 0193-4120 Page No. 2452 - 2455

5. Lee Dong Yoon, Joo Sung Yoon, and Sung Keun Lee." Value and Use of Manufacturing Data". Mechanical Journal, 2017, 57 (8), 49-53.
6. M. Nam, and S. Lee, "Bigdata as a Solution to Shrinking the Shadow Economy", The E-Business Studies, Vol. 17. No. 5. pp. 107-116, 2016. DOI: https://doi.org/10.20462/TeBS.2016.10.17.5.107.
7. S.H. Kim, S. Chang, and S.W. Lee, "Consumer Trend Platform Development for Combination Analysis of Structured and Unstructured Big Data", Journal of Digital Convergence, Vol. 15. No. 6. pp. 133-143, 2017. DOI: https://doi.org/10.14400/JDC.2017.15.6.133.
8. Y. Kang, S. Kim, J. Kim, and S. Lee, "Examining the Impact of Weather Factors on Yield Industry Vitalization on Bigdata Foundation Technique", Journal of the Korea Entertainment Industry Association, Vol. 11. No. 4. pp. 329-340, 2017. DOI: https://doi.org/10.21184/jkeia.2017.06.11.4.329.
9. S. Kim, H. Hwang, J. Lee, J. Choi, J. Kang, and S. Lee, "Design of Prevention Method Against Infectious Diseases based on Mobile Bigdata and Rule to Select Subjects Using Artificial Intelligence Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 174-178, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18603.
10. I. Jung, H. Sun, J. Kang, C.H. Lee, and S. Lee, "Bigdata Analysis Model for MRO Business Using Artificial Intelligence System Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 134-138, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18593.
11. S. Kim, S. Park, J. Kang, and S. Lee, "The Model of Bigdata Analysis for MICE Using IoT (Beacon) and Artificial Intelligence Service (Recommendation, Interest, and Movement)", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 314-318, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.21192.
12. S.H. Kim, J.K. Choi, J.S. Kim, A.R. Jang, J.H. Lee, K.J. Cha, and S.W. Lee, "Animal Infectious Diseases Prevention through Bigdata and Deep Learning", Journal of Intelligence and Information Systems, Vol. 24. No. 4. pp. 137-154, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.4.137.
13. S. Lee, and I. Jung, "Development of a Platform Using Big Data-Based Artificial Intelligence to Predict New Demand of Shipbuilding", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 19. No. 1. pp. 171-178, 2019. DOI: https://doi.org/10.7236/JIIBC.2019.19.1.171.
14. H. Hwang, S. Lee, S. Kim, and S. Lee, "Building an Analytical Platform of Bigdata for Quality Inspection in the Dairy Industry: A Machine Learning Approach", Journal of Intelligence and Information Systems, Vol. 24. No. 1. pp. 125-140, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.1.125.
15. Y. Shon, J. Park, J. Kang, and S. Lee, "Design of Link Evaluation Method to Improve Reliability based on Linked Open Bigdata and Natural Language Processing", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 168-173, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18601.
16. T. Minami and K. Baba, "A Study on Finding Potential Group of Patrons from Library's Loan Records", International Journal of Advanced Smart Convergence, Vol. 2, No. 2, pp. 23-26, 2013. DOI: https://doi.org/10.7236/IJASC2013.2.2.6
17. Hussain, H.I., Kamarudin, F., Thaker, H.M.T. & Salem, M.A. (2019) Artificial Neural Network to Model Managerial Timing Decision: Non-Linear Evidence of Deviation from Target Leverage, International Journal of Computational Intelligence Systems, 12 (2), 1282-1294.
18. Chukurna, O., Nitsenko, V., Kralia, V., Sahachko, Y., Morkunas, M., & Volkov, A. (2019). Modelling and managing the effect of transferring the dynamics of exchange rates on prices of machine-building enterprises in Ukraine. Polish Journal of Management Studies, 19 (1), 117-129.
19. Prakash, G., Darbandi, M., Gafar, N., Jabarullah, N.H., & Jalali, M.R. (2019) A New Design of 2-Bit Universal Shift Register Using Rotated Majority Gate Based on Quantum-Dot Cellular Automata Technology, International Journal of Theoretical Physics, https://doi.org/10.1007/s10773-019-04181-w.