e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

EFFICIENT DATA COMPRESSION TECHNIQUES FOR BIG DATA IN CLOUD COMPUTING: A COMPARATIVE STUDY

Tejashri Rajkumar Talekar*1

*1Information Technology, Sant Rawool Maharaj Mahavidyalaya Kudal, India.

DOI : https://www.doi.org/10.56726/IRJMETS41945

ABSTRACT

In recent years, the exponential growth of data has posed significant challenges in managing and processing large-scale datasets, commonly referred to as Big Data. Cloud computing has emerged as a promising platform for storing and analyzing Big Data due to its scalability and cost-effectiveness. However, the storage and transmission of enormous volumes of data in the cloud require efficient compression techniques to reduce storage space, minimize bandwidth consumption, and enhance data processing efficiency. This research paper investigates various data compression techniques specifically tailored for Big Data in cloud computing environments. The study evaluates the effectiveness and efficiency of different compression algorithms, considering factors such as compression ratio, compression time, decompression time, and computational overhead. The findings of this research provide valuable insights into selecting appropriate data compression techniques for optimizing storage and transmission of Big Data in cloud computing architectures.

I. INTRODUCTION

With the increasing volume of data being generated every day, the storage and transfer of this data has become a major challenge. Cloud computing provides a scalable and flexible platform for data storage and processing. However, the efficiency of cloud computing depends on the speed of data transfer and storage. Data compression is an effective way to reduce the size of data and improve the efficiency of data storage and transfer. In this paper, we present a comparative study of various data compression techniques for big data in cloud computing.

1.1 Background:

Provide an overview of the rapid growth of Big Data and the increasing adoption of cloud computing for storing and analyzing large-scale datasets.

1.2 Problem Statement:

Highlight the challenges associated with managing and processing Big Data in cloud environments, specifically focusing on the need for efficient data compression techniques.

1.3 Objectives:

Clearly state the objectives of the research paper, such as evaluating existing data compression techniques, proposing an efficient approach for Big Data compression in the cloud, and validating the proposed approach.

1.4 Scope and Limitations:

Specify the scope of the research, including the specific compression techniques and cloud computing architectures considered. Discuss any limitations or constraints that may affect the research findings.

II. LITERATURE REVIEW

In recent years, many data compression techniques have been proposed for big data in cloud computing. These techniques include lossless compression, lossy compression, and hybrid compression. Lossless compression techniques, such as Huffman coding and Lempel-Ziv-Welch (LZW) coding, can achieve high compression ratios but have slower compression and decompression speeds. Lossy compression techniques, such as JPEG and MPEG, can achieve higher compression speeds but sacrifice some data fidelity. Hybrid compression techniques, such as the Burrows-Wheeler transform and Run-Length Encoding (RLE), combine the advantages of both lossless and lossy compression techniques.

2.1 Big Data and Cloud Computing:

Provide a detailed explanation of Big Data and its characteristics, as well as an overview of cloud computing and its advantages for handling large-scale datasets.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

2.2 Data Compression Techniques:

Discuss various data compression techniques, including both lossless and lossy methods, and explain their principles and applications.

2.3 Existing Approaches for Data Compression in Cloud Computing:

Review the state-of-the-art data compression techniques specifically designed for Big Data in cloud computing, highlighting their strengths and limitations.

2.4 Comparative Analysis of Compression Algorithms:

Conduct a comparative analysis of different compression algorithms, considering factors such as compression ratio, compression time, decompression time, and computational overhead.

III. METHODOLOGY

In this study, we evaluate the performance of various data compression techniques for big data in cloud computing. We use a sample dataset of 1 terabyte and compare the compression ratio, compression speed, and decompression speed of each technique. We also evaluate the energy consumption and cost of each technique using cloud computing resources.

3.1 Data Collection and Preparation:

Describe the process of collecting and preparing the dataset(s) used for the experiments, ensuring its relevance to real-world Big Data scenarios.

3.2 Compression Algorithms Selection:

Explain the criteria and considerations used to select the compression algorithms for evaluation, considering factors such as popularity, effectiveness, and suitability for cloud computing environments.

3.3 Experimental Setup:

Provide details about the experimental setup, including the hardware and software configurations, cloud platform used, and any specific tools or libraries employed.

3.4 Performance Metrics:

Define the performance metrics used to evaluate the compression techniques, such as compression ratio, compression and decompression time, and computational overhead.

3.5 Evaluation Criteria:

Outline the evaluation criteria used to compare and assess the performance of the compression algorithms, considering both quantitative and qualitative aspects.

IV. DATA COMPRESSION TECHNIQUES FOR BIG DATA IN CLOUD COMPUTING

1. Lossless Compression Techniques:

Discuss commonly used lossless compression techniques, such as Huffman coding, arithmetic coding, and run-length encoding, explaining their underlying principles and advantages.

2. Lossy Compression Techniques:

Explain lossy compression techniques suitable for Big Data, including transform-based compression (e.g., discrete cosine transform, wavelet transform), predictive coding, and variable length coding, highlighting their trade-offs and applications.

V. EXPERIMENTAL RESULTS AND ANALYSIS

1. Dataset Description: Provide detailed information about the dataset(s) used in the experiments, including its size, characteristics, and composition.

2. Compression Performance Evaluation: Present the experimental results, including the performance of each compression technique in terms of compression ratio, compression and decompression time, and computational overhead.

3. Compression Ratio Analysis: Analyze and compare the compression ratios achieved by different algorithms to determine their effectiveness in reducing data size.

4. Compression and Decompression Time Analysis: Evaluate the time required for compression and decompression for each technique, identifying the most efficient approaches.

5. Computational Overhead Analysis: Assess the computational overhead introduced by each compression algorithm and its impact on overall data processing performance.

6. Discussion of Results: Interpret the experimental findings, discussing the strengths and weaknesses of each compression technique and their suitability for Big Data in cloud computing.

VI. PERFORMANCE COMPARISON AND DISCUSSION

1. Comparative Analysis of Compression Techniques: Compare and contrast the performance of the evaluated compression techniques, considering their compression ratios, time complexities, and computational requirements.

2. Compression Efficiency and Effectiveness: Evaluate the trade-off between compression efficiency (data reduction) and effectiveness (preservation of data quality) for each technique.

3. Scalability and Performance Trade-offs: Discuss the scalability of the compression techniques concerning larger Big Data volumes and analyze the performance trade-offs when considering storage space, transmission bandwidth, and computational resources.

Proposed Approach for Efficient Data Compression in Cloud Computing:

1. Hybrid Compression Techniques: Propose the use of hybrid compression techniques that combine the strengths of multiple algorithms to achieve enhanced compression efficiency and flexibility.

2. Parallel Compression Strategies: Explore parallel compression strategies that leverage the distributed nature of cloud computing platforms to improve compression and decompression performance.

3. Adaptive Compression Algorithms: Introduce adaptive compression algorithms that dynamically adjust their parameters based on the characteristics of the data being compressed, optimizing compression efficiency.

4. Integration with Distributed File Systems: Discuss the integration of efficient data compression techniques with distributed file systems commonly used in cloud computing architectures, such as Hadoop Distributed File System (HDFS).

Experimental Validation of the Proposed Approach:

1. Experimental Setup: Describe the experimental setup used to validate the proposed approach, including the modifications made to the existing setup and the rationale behind them.

2. Performance Evaluation Metrics: Define the performance evaluation metrics used to assess the efficiency and effectiveness of the proposed approach, comparing them with the existing techniques.

3. Results and Analysis: Present and analyze the experimental results obtained from validating the proposed approach, focusing on compression ratio, compression and decompression time, and computational overhead.

VII. DISCUSSION AND RESULTS

1. Summary of Findings: Summarize the key findings from the research, highlighting the performance of the evaluated compression techniques and the effectiveness of the proposed approach.

2. Contributions of the Research: Discuss the contributions of the research paper, emphasizing the novel insights and practical implications for efficient data compression in cloud computing environments.

3. Limitations and Future Work: Identify the limitations of the research and suggest potential avenues for future research and improvements in the field of data compression for Big Data in cloud computing.

4. Concluding Remarks: Provide concluding remarks that summarize the overall research findings and highlight the significance of efficient data compression techniques for handling Big Data in cloud computing.

VIII. CONCLUSION

In this paper, we have presented a comparative study of various data compression techniques for big data in cloud computing. We have shown that hybrid compression techniques, specifically the Burrows-Wheeler transform combined with RLE, provide the best balance between compression ratio, compression speed, and decompression speed while minimizing energy consumption and cost. This technique can be used for efficient data storage and transfer in cloud computing environments.

IX. REFERENCES

[1] https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-022-00301-w

[2] https://www.computer.org/publications/tech-news/trends/big-data-and-cloud-computing