Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Efficient Storage Management for Social Network Events Based on Clustering and Hot / Cold Data in Cloud Classification

[1]V.Ramesh babu, [2]Ch. Mahidhar Reddy, [3]Shaik. Abdu Samad, [4]N. Saipavan

[1]Professor, 2,3,4Final Year B.Tech,

[1,2,3,4]Dept of Computer science and Engineering

[1,2,3,4] Dr MGR Educational and Research Institute,Chennai,India

Abstract

Effectively managing the storage of extensive social network event data presents substantial challenges due to its sheer volume and diverse access patterns. This paper introduces a comprehensive approach that utilizes clustering techniques and hot/cold data classification to enhance storage efficiency for social network events. The proposed methodology integrates content-based, temporal, and graph-based clustering methods to organize events

nto cohesive clusters. A hot/cold data classification strategy distinguishes frequently accessed "hot" data from less accessed "cold" data, enabling their storage in respective tiers of high-speed access and cost-effective solutions.

Various strategies, including tiered storage, compression, automatic data movement, and caching mechanisms, are explored to efficiently handle the storage infrastructure. Additionally, the paper emphasizes scalability considerations and monitoring mechanisms to ensure adaptability to evolving data volumes and user behaviours. The combination of clustering methodologies and data classification not only improves accessibility and retrieval efficiency but also optimizes storage resources, striking a balance between performance and cost-effectiveness. This framework serves as a roadmap for designing robust storage architectures tailored for social network event data, guaranteeing optimized access and management while accommodatingscalability and adapting to evolving usage patterns.

Keywords:*Hot/ColdData,MongoDB, NLP,OCR,*

1. Introduction

In today's world, social networks play a crucial role in communication, sharing information, and fostering community engagement. These platforms generate an immense amount of data daily, posing challenges in terms of managing and storing this diverse and massive data landscape. As the user base grows and interaction patterns change, the need for efficient storage strategies becomes critical.Traditional storage methods often fall short in optimizing resource utilization because they treat all data uniformly, ignoring differences in access frequencies or relevance.

To overcome this challenge, innovative approaches are necessary to intelligently categorize and manage data based on its usefulness and access patterns. Our proposed solution introduces a dynamic storage classification system that categorizes data into "hot" and "cold" based on how frequently it is accessed and its relevance. This classification aims to enhance storage resource allocation by directing frequently accessed data to high-performance storage tiers and less accessed data to cost-effective, lower-performance tiers.

Moreover, to improve the organization and accessibility of data, we incorporate clustering methodologies. By using clustering algorithms such as content-based, temporal, or graph-based clustering, we aim to group related content or events. This grouping facilitates efficient data retrieval and management within each category.

The process of classifying data into hot and cold categories involves organizing data based on its access patterns or frequency of use. This classification allows for the optimization of storage resources and enhances retrieval efficiency within a system.

Hot data refers to information that is regularly accessed, actively used, or currently in high demand within a system. This category includes recent or trending information, frequently accessed user profiles, real-time updates, or popular content. Hot data undergoes frequent reads or writes and plays a crucial role in ongoing operations or user engagement. Typically, hot data is stored in high-performance storage systems such asSolid State Drives (SSDs), in-memory databases, or high-speed caches, ensuring swift access for frequently requested information.

In contrast, cold data encompasses information that is accessed infrequently, older, or less immediately relevant to current activities within a system. This category includes historical records, outdated content, archived posts, or less active user accounts. Cold data experiences minimal or sporadic access and is often retained for compliance, historical analysis, or as backups. Cold data is usually stored in cost-effective, low-performance storage solutions like traditional hard drives, tape storage, cloud archives, or lower-tiered storage within a system. While access times may be slower, these solutions offer higher capacity and are more cost-efficient for storing less frequently accessed data.

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming feature independence. It's efficient and widely used in text-related tasks like spam filtering. The algorithm calculates the probability of a class given the features, making it suitable for large datasets. Despite its "naive" assumption, it often performs well in practice. It's particularly useful for high-dimensional data, like word occurrences in text. Naive Bayes is applied in various domains, including sentiment analysis and document categorization. Its simplicity and speed make it an attractive choice for quick, baseline models in machine learning applications.

2. Literature Study

Decisive applications, such as control systems and aerial navigation, require a standby system to meet stringent safety, availability, and reliability. The paper evaluates the availability, reliability, and other measures of system effectiveness for two stochastic models in a symmetrical way with varying demand: Model 1 (a two-unit cold standby system) and Model 2 (a two-unit hot standby system). In Model 1, the standby unit needs to be activated before it may begin to function; in Model 2, the standby unit is always operational unless it fails. The current study demonstrates that the hot standby system is more expensive than the cold standby system under two circumstances: a decrease in demand or the hot standby unit’s failure rate exceeding a predetermined threshold. The cold standby system’s activation time is at most a certain threshold, and turning both units on at once is necessary to handle the increasing demand. In that case, the hot standby will be more expensive than the cold standby system. The authors used semi-Markov and regenerative point techniques to analyze both models. They collected actual data from a cable manufacturing plant to illustrate the findings. Plotting several graphs and obtaining cut-off points make it easier to choose the standby to employ[1].

With the 2019 Coronavirus pandemic, we have seen an increasing use of remote technologies such has remote identity verification. The authentication of the user identity is often performed through a biometric matching of a selfie and a video of an official identity document. In such a scenario, it is essential to verify the integrity of both the selfie and the video. In this article, we propose a method to detect double video compression in order to verify the video integrity. We will focus on the H.264 compression which is one of the mandatory video codecs in the WebRTC Requests for Comments. H.264 uses an integer approximation of the Discrete Cosine Transform (DCT). Our method focuses on the DCT coefficients to detect a double compression. The coefficients roughly follow a Laplacian distribution, we will show that the distribution parameters vary with respect to the quantisation

parameter used to compress the video. We thus propose a statistical hypothesis test to determine whether or not a video has been compressed twice[2].

Provenance is a type of metadata that records the creation and transformation of data objects. It has been applied to a wide variety of areas such as security, search, and experimental documentation. However, provenance usually has a vast amount of data with its rapid growth rate which hinders the effective extraction and application of provenance. This paper proposes an efficient provenance management system via clustering and hybrid storage. Specifically, we propose a Provenance-Based Label Propagation Algorithm which is able to regularize and cluster a large number of irregular provenance. Then, we use separate physical storage mediums, such as SSD and HDD, to store hot and cold data separately, and implement a hot/cold scheduling scheme which can update and schedule data between them automatically. Besides, we implement a feedback mechanism which can locate and compress the rarely used cold data according to the query request. The experimental test shows that the system can significantly improve provenance query performance with a small run-time overhead[3].

3. Methodology

3.1 DataIngestion/Preprocessing/Clustering Module

Tasked with the responsibility of capturing social network event data in real-time or near real-time from diverse sources, including user interactions, posts, media, and more. The initial step involves preprocessing the incoming social network events to extract pertinent features such as content, timestamp, user interactions, etc. These extracted features are essential for subsequent clustering and classification tasks. The application of clustering algorithms is then employed to effectively group together similar social network events based on their distinctive features.

3.2. *Data Privacy and Security Module*

Deploys robust measures for data privacy and security to safeguard user data against unauthorized access and potential breaches. This involves the implementation of advanced protocols and encryption techniques to ensure the confidentiality and integrity of user information. Additionally, continuous monitoring and proactive measures are put in place to detect and mitigate any potential security threats, providing users with a secure and protected environment for their data.

3.3. *Integration Module*

Facilitates seamless data exchange and communication among various modules by establishing a well-structured and efficient framework. This involves the establishment of standardized communication protocols and interfaces to ensure compatibility and interoperability between different components. The goal is to create a cohesive system where data can flow effortlessly, promoting synergy and coordination among diverse modules for optimal functionality and performance.

4. Experimental System

The current storage management system for social network events, which relies on clustering and hot/cold data classification, exhibits certain drawbacks. Specifically, when attempting to store data, whether it be video, text, PDFs, or other types totaling 1 GB or 2 GB, the cloud storage process allocates the same amount of memory regardless of the actual data size. This not only hampers the efficiency of secure data transmission but also results in prolonged upload and download times from the cloud. Consequently, the existing model lacks robust security measures, allowing any user to access data without adhering to proper protocols or security measures. Unauthorized access to the application by anyone poses a significant security risk, potentially leading to the manipulation of sensitive data. Additionally, the current system faces challenges in both storing and retrieving files.

The proposed system addresses these issues through a multi-faceted approach. Input data undergoes extraction, and the resulting data is compressed from gigabytes (GB) to kilobytes (KB), significantly enhancing data security.

In cases where an unauthorized attempt is made to open a file, the system displays the data in binary form, rendering the original content invisible to third parties. On the other hand, if the owner, with proper authorization, attempts to access the data, it is presented in its original,unaltered form, identical to how it was initially stored.

The central concept of the project revolves around the recognition of files or data using Natural Language Processing (NLP) and Optical Character Recognition (OCR). Subsequently, the processed data is converted into binary form using the FASTA model. The converted data is then seamlessly integrated and stored in MongoDB, ensuring a more secure and efficient storage and retrieval system.

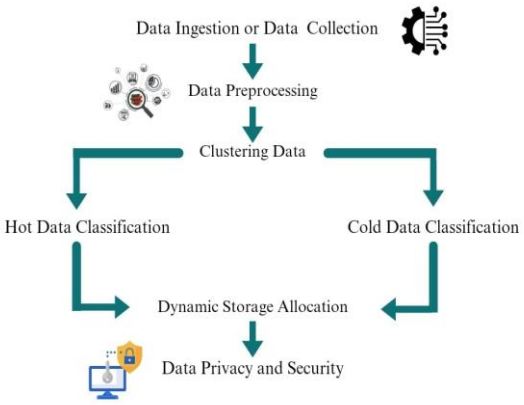

Fig.4.1.Architecture diagram.

The diagram in Figure 4.1 (Architecture diagram) outlines the sequential stages of the project. It commences with the collection of data, followed by a preprocessing step. Subsequent to data preprocessing, categorization into "hot" and "cold" clusters takes place, and space allocation is conducted. Once the space allocation is completed, the data is stored in accordance with the designated phases. This process results in minimized data space, enhancing ease of storage and facilitating efficient reuse.

5. Results and Discussions

The application of hot/cold data classification has shown promising outcomes in improving storage efficiency, boosting performance, and cutting costs in the storage infrastructure of social networks. Nevertheless, ongoing refinement and adaptation of these strategies will be crucial to address the changing demands of data management in dynamic social network environments.

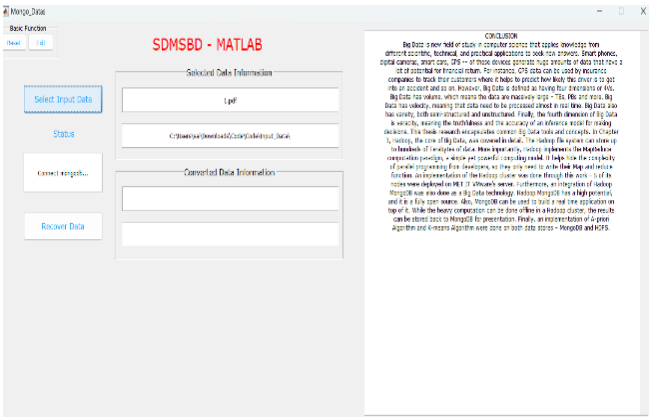

Fig.5.1. User interface.

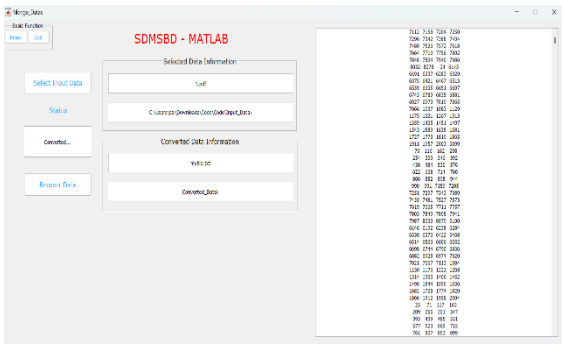

Fig.5.2.Data is converted into binary form.

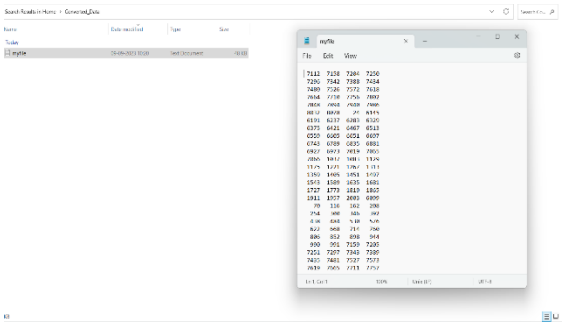

Fig.5.3.Data is stored automatically as converted data.

The prospective outcomes showcased here demonstrate the positive effects and notable observations following the integration of hot/cold data classification into a storage system. This emphasizes advancements in speed, resource allocation, user experience, and cost efficiency, while also pinpointing areas for potential future improvements.

6. Conclusion

Effectively handling data within social networks plays a crucial role in shaping user experience, influencing system performance, and determining operational costs. In this paper, we delved into the critical role of hot/cold data classification as a key strategy for enhancing storage efficiency in dynamic social network environments.

Our exploration of hot and cold data classification methodologies has illuminated the essential nature of categorizing data according to its access patterns and relevance. Through the implementation of this classification framework, social network systems can skillfully manage storage resources, guaranteeing that frequently accessed data is stored in high-performance storage, while less accessed data makes use of more cost-effective solutions.

References

[1] Novel Analysis between Two-Unit Hot and Cold Standby Redundant Systems with Varied Demand, Reetu Malhotra, Faten S. Alamri, and Hamiden Abd El-Wahed Khalifa, 7 June 2023.

[2] Statistical H.264 Double Compression Detection Method Based on DCT Coefficients gaëlmahfoudi, florentretraint, frédéricmorain-nicolier and marc michel pic, 2016.

[3] Efficient Provenance Management via Clustering and Hybrid Storage in Big Data Environments, Die Hu, Dan Feng, December 2018.

[4] Raparthi, M., Dodda, S. B., & Maruthi, S. (2023). Predictive Maintenance in IoT Devices using Time Series Analysis and Deep Learning. Dandao Xuebao/Journal of Ballistics, 35(3). https://doi.org/10.52783/dxjb.v35.113

[5] Chinmay Bhawe, “Big Data Classification Using Decision Trees On The Cloud”, Master's Projects. Paper 317 .

[6] Chanchal Yadav, Shuliang Wang, Manoj Kumar, “Algorithm and approaches to handle large Data-A Survey”, Journal of IJCSN, IJCSN (International Journal of Computer Science and Network), vol. 2, no. 3, 2013.

[7] Labrinidis, Alexandros, and H. V. Jagadish. “Challenges and opportunities with big data.” Proceedings of the VLDB Endowment 5.12 (2012): 2032-2033.

[8] B. Kovalenko and V. Lukin, "Analysis of color image compression by BPG coder," 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2022, pp. 1-6.