Stream Processing Based on Enormous Data Continuous Analysis System

G.Keerthana,
Assistant Professor, Department of Computer Science and Engineering, J.J. College of Engineering and Technology, Trichy, Tamilnadu
Dr.P.Chellammal,
Professor, Department of Computer Science and Engineering, J.J. College of Engineering and Technology, Trichy, Tamilnadu
R.Shariff Nisha,
Assistant Professor, Department of Computer Science and Engineering, J.J. College of Engineering and Technology, Trichy, Tamilnadu
S.Narayanasamy,
Assistant Professor, Department of Computer Science and Engineering, J.J. College of Engineering and Technology, Trichy, Tamilnadu

Abstract

This research presents a novel system designed for continuous analysis of enormous data using stream processing. The system consists of several modules including metadata management, query plan generation, data import task generation, increment processing, MR message processing, and database connection. The metadata management module handles the management of meta-information for data tables and databases. The query plan generation module receives query requests and generates optimized query plans. The data import task generation module handles data import requests and generates data import MR operation sets. The increment processing module incrementally processes data import and query operations in parallel using a Hadoop system. The MR message processing module receives results from Map or Reduce functions of the Hadoop system and outputs them to the next operation or the Reduce end. Finally, the database connection module serves as an interface between the Hadoop system and the databases. The proposed system leverages the Hadoop system to organize databases in a distributed manner and allows simultaneous execution of data import and query operations. Additionally, the system incorporates a pipeline technology to enhance the MR execution flow, enabling continuous stream mode execution of data queries and significantly reducing the time required for analyzing enormous data.

Keywords: Stream processing, Continuous analysis, Enormous data, Metadata management, Query plan generation, Data import, MR message processing, Database connection.
DOI Number: 10.48047/nq.2020.18.8.nq20222 **NeuroQuantology 2020;18(8):174-178**

Introduction

In today's data-driven world, organizations are faced with the challenge of analyzing and extracting insights from enormous volumes of data. The ability to process and analyze data in real-time has become crucial for making informed decisions and gaining a competitive edge. Traditional batch processing systems often fall short in meeting the demands of continuous data analysis. Therefore, there is a need for innovative systems that can handle continuous analysis of enormous data streams.This research focuses on addressing this need by proposing aenormous data continuous analysis system suitable for stream processing. The system is designed to efficiently analyze large volumes of data in a continuous stream mode, significantly reducing the time required for data analysis. It leverages the power of the Hadoop framework, along with several specialized

modules, to streamline the data analysis process.[1]

The key components of the proposed system include the metadata management module, query plan generation module, data import task generation module, increment processing module, MR message processing module, and database connection module. These modules work together to manage the metadata of data tables and databases, generate optimized query plans, handle data import tasks, process incremental updates, and connect with the underlying databases.The system's utilization of the Hadoop framework enables the organic organization of databases across distributed nodes.[2]This distributed architecture allows for parallel execution of data import and query operations, enhancing the system's scalability and performance. Additionally, a pipeline technology is incorporated to optimize the execution flow of MapReduce (MR) operations, further improving the efficiency of data analysis.

The main objective of this research is to develop a system that can continuously analyzeenormous data streams using stream processing techniques. By combining the power of the Hadoop framework, metadata management, optimized query plans, and efficient data import and processing mechanisms, the system aims to provide real-time insights and analysis capabilities. The proposed enormous data continuous analysis system offers a novel approach to handle the challenges of analyzing large volumes of data in a continuous stream mode.[3] By leveraging the Hadoop framework and employing specialized modules, the system aims to significantly reduce the time required for analyzingenormous data and provide valuable insights for decision-making processes. The subsequent sections will delve into the system's architecture, functionalities, and the benefits it brings to the field of data analysis and decision-making.[5]

Related Work

With the advent of the big data era, the challenge of extracting valuable information from vast amounts of data has become increasingly significant. Large-scale data analysis plays a crucial role in addressing this challenge, and it imposes higher requirements on data analysis systems. Traditional approaches, such as using centralized database management systems (DBMS) for data analysis, are inadequate for handling the ever-growing volumes of data and cannot meet the diverse and fast-paced demands of data analysis.[1]There are two main types of existing large-scale data analysis systems: parallel database systems (Parallel DBMS) and systems based on the MapReduce (MR) framework. However, both types have their limitations. Parallel databases face scalability issues and struggle to ensure fault tolerance as data volumes increase. On the other hand, MR-based systems often exhibit lower processing efficiency when dealing with numerous datasets, especially when it comes to handling relational data. Consequently, academia and industry have recognized the need to integrate the advantages of both approaches. However, most integration efforts have been superficial, focusing only on the interface level and lacking comprehensive architectural integration.Existing integrated systems that combine the MR framework and databases still suffer from incomplete integration, failing to fully leverage the strengths of both approaches.[2]These systems also lack improvements to the existing framework, making them ill-equipped to rapidly adapt to the diverse demands of data analysis. For instance, challenges persist in the long data importing process and the design of batch processing within the MR framework remains unsolved.

The current state of integrated systems that combine the MR framework and databases falls short in terms of thorough integration and capitalizing on the advantages of both approaches. Furthermore, these systems lack improvements to address the specific challenges of data importing and batch processing design. To overcome these limitations and cater to the fast-paced and diverse requirements of data analysis, a new approach is needed that takes full advantage of the strengths of both the MR framework and databases.[7] The subsequent sections will delve into the proposed system's architecture

and its innovative solutions to overcome these challenges, enabling efficient and comprehensive analysis of large-scale data.

Research Objective

The primary goal of this research is to create a powerful and reliable system that can handle the continuous analysis of large amounts of data using stream processing techniques. The researchers aim to achieve this objective by harnessing the capabilities of the Hadoop framework, which is well-suited for organizing databases and executing data import and query operations simultaneously.One of the key focuses of the research is to optimize the execution flow of MapReduce (MR) operations within the Hadoop system. By doing so, they aim to improve the efficiency of analyzingenormous volumes of data. This optimization involves developing mechanisms that enable data queries to be executed in a continuous stream mode. This means that the system processes and analyzes data in a seamless and uninterrupted manner, resulting in significant reductions in the time required for analyzingenormous datasets.The researchers also aim to address the challenges associated with analyzing ever-increasing amounts of data. Traditional methods using centralized database management systems (DBMS) are often unable to cope with the growing data volumes and fail to meet the fast-paced demands of data analysis. Therefore, the research focuses on developing a system that overcomes these limitations and supports the analysis of continuously incoming data.

Overall, the research objective is to design and implement a robust system that leverages the advantages of the Hadoop framework, improves the efficiency of data analysis, and enables continuous analysis of enormous datasets. By achieving this objective, the researchers aim to significantly reduce the time required for analyzing large-scale data and provide valuable insights from the collected information.

Stream Processing Based on Enormous Data Continuous Analysis System

We have developed a comprehensive system for analyzing large-scale data in a continuous manner, specifically designed for Stream Processing. This system consists of several key modules: the metadata management module, inquiry plan generation module, data importing task generation module, incremental processing module, MR (MapReduce) message processing module, and database link block.The metadata management module serves as a storage unit for various configuration files and information related to the data. It stores important details such as configuration files, source data patterns, data importing patterns, database node information, and database linkage information. This module ensures that all the necessary information is readily available for the other modules to access and utilize.The inquiry plan generation module is responsible for receiving user's inquiry requests. It analyzes the source data pattern information provided by the metadata management module and generates an optimal query plan based on the user's request. This module then sends the query plan to the incremental processing module for further execution. Additionally, it also sends the query parse result to the data importing task generation module, which is responsible for generating data importing tasks based on the query parse result and user data import requests.

The data importing task generation module receives user data import requests and extracts relevant source data messages from the metadata management module. It generates configuration files that contain database node information and database linkage information necessary for distributing the source data. The module also uses the query parse result obtained from the inquiry plan generation module to determine the data importing pattern. It generates an incremental identifier attribute for identifying the imported data and stores it in the metadata management module. Based on the data importing pattern and configuration files, the module generates executable data importing MR operations for the Hadoop system and sends them to the incremental processing module for execution.

The incremental processing module plays a crucial role in the system. It receives the inquiry plan from the inquiry plan generation module and the data importing MR operations from the data importing task generation module. It submits the data importing MR operations to the Hadoop system, which then executes the data import operations in parallel. Similarly, when an inquiry is made, the incremental processing module compiles the inquiry into an executable MR operation collection, utilizing the configuration files from the metadata management module. It submits the MR operation collection to the Hadoop system, which carries out the inquiry operation for the partially imported data.The MR message processing module is embedded within the Hadoop system. It ensures that when the Hadoop system executes data import or inquiry MR operation collections, the intermediate data generated at the Map end is properly pushed to the corresponding Reduce end. Additionally, it manages the execution result generated at the Reduce end and pushes it to the Map end of the next task, enabling a seamless flow of data processing.
Finally, the database link block serves as an interface between the Hadoop system and the database. It facilitates the communication and interaction between the two systems, allowing for smooth data transfer and access.Overall, our large-scale data continuous analysis system for Stream Processing offers an efficient and effective solution for analyzingenormous amounts of data. It optimizes the data importing and querying processes, enhances fault-tolerance, and leverages the power of Hadoop and database integration to meet diverse data analysis demands in a fast and continuous manner.

Conclusion

In conclusion, this research proposes a novel system for continuous analysis of enormous data through stream processing. By integrating various modules, including metadata management, query plan generation, data import task generation, increment processing, MR message processing, and database connection, the system offers an efficient and scalable solution. The utilization of the Hadoop system and pipeline technology enables simultaneous execution of data import and query operations, facilitating continuous stream mode execution and significantly reducing the analysis time for enormous data. The developed system contributes to the field of data analytics by providing an effective approach for handling and analyzing large volumes of data in a continuous and efficient manner.

Reference

1. Bahri, M., Bifet, A., Gama, J., Gomes, H. M., &Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(3), e1405. https://doi.org/10.1002/widm.1405
2. Kolajo, T., Daramola, O. &Adebiyi, A. Big data stream analysis: a systematic literature review. J Big Data 6, 47 (2019). https://doi.org/10.1186/s40537-019-0210-7
3. Kovatchev, B., & Clarke, W. (2008). Peculiarities of the Continuous Glucose Monitoring Data Stream and Their Impact on Developing Closed-Loop Control Technology. Journal of Diabetes Science and Technology. https://doi.org/10.1177/193229680800200125
4. Aggarwal, C. C., Yu, P. S., Han, J., & Wang, J. (2002). A Framework for Clustering Evolving Data Streams. Proceedings 2003 VLDB Conference, 81-92. https://doi.org/10.1016/B978-012722442-8/50016-1
5. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, August 1980, doi: 10.1109/TASSP.1980.1163420.
6. Deeks, J. J., Higgins, J. P., & Altman, D. G. Analysing data and undertaking

meta-analyses. 241-284. https://doi.org/10.1002/9781119536604.ch10

7. D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," in IEEE Signal Processing Magazine, vol. 30, no. 3, pp. 83-98, May 2013, doi: 10.1109/MSP.2012.2235192.
8. K. Yu and S. Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1071-1079, July 2011, doi: 10.1109/TASL.2010.2076805.