ISSN: 2710-0588

Research on Construction of Offline Data Warehouse for Ship Shore Power based on DolphinScheduler and Hive

Zewen Zhang*, Xin Zhang, Taizhi Lv

College of Information Engineering, Jiangsu Maritime Institute, Nanjing, Jiangsu 211170, China

Abstract

In today's networked, intelligent, and data-driven era, the shore power industry is facing the challenge of rapidly growing data. This paper presents the construction of a professional offline data warehouse system for shore power based on DolphinScheduler and Hive. Firstly, MySQL is adopted as the backend database, combined with Sqoop to synchronize business data to HDFS, ensuring data reliability and integrity. Secondly, the Flume-Kafka-Flume architecture is utilized to achieve real-time collection and caching of user behavior data, providing data support for subsequent analysis. Thirdly, HQL statements are written in Hive to clean, merge, and analyze shore power data, calculating key indicators such as electricity consumption and usage trends. Fourthly, data visualization is achieved through the integration of Superset, displaying data analysis results via a web interface. Fifthly, DolphinScheduler is employed for timed scheduling, ensuring dependency control among various tasks and the smooth operation of the project. This system fully leverages the replication mechanism of HDFS to enhance reliability, dynamically adds nodes to achieve system scalability, and fully utilizes the fault tolerance of the Yarn scheduler. It saves time and computational costs for the shore power industry, realizing higher value and benefits.

Keywords

Ship Shore Power; Offline Data Warehouse; Hadoop; Hive; DolphinScheduler.

1. Introduction

With the nation's vigorous deployment of 5G technology, the further maturation of the Internet of Things (IoT), and the continuous iteration of computer hardware and software, the internet industry has ushered in an unprecedented period of development. Alongside societal progress and technological advancements, ship shore power, as an environmentally friendly and efficient energy supply method, is gaining increasing favor from marine institutions and shipping companies [1]. The use of ship shore power not only helps optimize the energy structure and reduce ship emissions but also enhances energy utilization and lowers enterprise operational costs.

After thoroughly studying Hadoop ecosystem technology and data warehouse modeling theory, this paper designs and implements a ship offline data warehouse based on big data frameworks such as Hadoop and Hive. This project is a fusion of various major big data frameworks and demonstrates a deep understanding of reasonable data warehouse layering, creating a data warehouse characterized by high reliability, high scalability, and high fault tolerance. By writing HQL scripts to transform ship data in the data warehouse, data value is enhanced, providing data support for corporate decision-making. Simultaneously, these professional indicators also assist ship platform operators in controlling website operations, promptly addressing and rectifying related issues, and creating higher utilization value for the website. This truly realizes the promotion of related enterprises and technological development through technology.

ISSN: 2710-0588

2. Overall Design

The system provides efficient solutions for data collection, storage, management, and analysis. By synchronizing business data from MySQL to HDFS using the Sqoop tool, the system ensures data reliability and consistency. To achieve efficient processing of large-scale data, the system adopts HDFS as the underlying storage, fully leveraging its distributed storage and replication mechanism to enhance data reliability and fault tolerance. Through the use of HQL statements in Hive for data cleaning, merging, and statistical processing, the system is able to calculate key business indicators such as ship shore power utilization and maintenance status, supporting enterprise operational optimization and decision analysis.

The system also features a multi-layered data warehouse architecture, including the ODS (Operational Data Store), DWD (Data Warehouse Detail), DWS (Data Warehouse Summary), and ADS (Application Data Store) layers, each serving different levels of data processing and storage needs. This layered architecture not only enables effective management and storage of massive amounts of data but also allows the system to flexibly respond to various data analysis requirements [2].

2.1. Data Acquisition

The system utilizes the Sqoop tool to achieve batch synchronization of business data. The business data, primarily stored in a MySQL database, contains crucial information such as ship records, ship shore power piles, and ship shore power usage. By regularly executing Sqoop tasks, the system imports this business data into HDFS, ensuring data integrity and consistency.

2.2. Data Storage

In the ship shore power offline data warehouse system, data storage is a key aspect of ensuring data security, reliability, and efficient access. To meet the storage demands of massive amounts of data, the system adopts the Hadoop Distributed File System (HDFS) as the underlying storage solution. HDFS provides high reliability and high availability through its unique distributed architecture and replication mechanism, enabling large-scale data storage and management on inexpensive commercial hardware [3]. Specifically, when data is written to HDFS, it is divided into multiple blocks and replicated across different DataNodes, ensuring that data can still be recovered from other nodes even if a certain node fails. This design greatly enhances data security and system fault tolerance.

To effectively manage and access structured data, the system leverages the Hive data warehouse tool. Hive provides a SQL-like query language (HiveQL), allowing users to easily query and analyze large-scale data stored in HDFS. Through Hive, the system can partition, bucket, and index data, thereby improving query performance and data management efficiency [4]. To save storage space, the system also applies compression techniques such as LZO during the storage process, which not only reduces disk space usage but also accelerates data transmission speeds [5].

2.3. Data Processing

Data processing is one of the core functions of the ship shore power offline data warehouse system, mainly including data cleaning, data integration, and data computation. To ensure data quality and consistency, the system first performs data cleaning during the data processing process, removing invalid, duplicate, or erroneous data. Using cleaning scripts written in HiveQL, the system can efficiently filter and transform the collected raw data, ensuring data accuracy and completeness.

In terms of data integration, the system unifies information scattered across different data sources into the data warehouse. Through Hive's ETL (Extract, Transform, Load) process, the system merges business data imported from MySQL with real-time data obtained from Kafka,

ISSN: 2710-0588

generating a unified, structured dataset. During this process, the system also summarizes and aggregates data based on business needs, generating key business indicators such as shore power pile usage rates and pile failure rates. To meet the processing demands of large-scale data, the system adopts the Spark big data processing framework. As an in-memory computing framework, Spark's high computational performance and rich API interfaces enable the system to perform more complex real-time data processing and analysis operations [6]. Through Spark SQL, users can directly query data in Hive, further improving the flexibility and efficiency of data processing.

2.4. Data Visualization

Data visualization is the final link in the ship shore power offline data warehouse system and the part that directly faces users. It aims to present analysis results to users in an intuitive and easy-to-understand manner, helping them make effective business decisions. Apache Superset is a powerful open-source data visualization and data exploration platform designed to provide data analysts and business users with elegant, intuitive tools for querying, visualizing, and analyzing data [7]. It supports connections to multiple data sources, including SQL databases, cloud services, log files, etc., and offers a rich variety of visualization types such as bar charts, line charts, pie charts, maps, scatter plots, as well as interactive dashboards and data slicing functions, allowing users to deeply explore data and discover insights.

The system achieves data visualization through Superset, displaying key business indicators and analysis results in the form of charts such as status charts, pie charts, bar charts, and maps. These charts not only visually reflect data distribution and trends but also allow users to customize queries and analysis needs through an interactive interface, further tapping into the potential value of the data.

3. Building a Data Warehouse Based on Hive for Ship Shore Power Data Analysis

This paper focuses on constructing a three-tiered architecture for ship shore power data analysis using Hive: the data operation layer, the data detail layer, and the data service layer. Hive, a data warehousing tool built on Hadoop, enables massive data extraction, transformation, storage, and querying. It provides an SQL query interface that translates SQL statements into MapReduce tasks for execution. Through data layering, a clear data structure is achieved, shielding anomalies in raw data and simplifying complex issues. External tables are created in the system, with some tables implementing partitioned processing. For instance, the shore power usage data table is partitioned by month, and the ship trajectory information table is partitioned by day to enhance query efficiency.

3.1. Data Operation Layer

In this paper, the original ship shore power data from the source MySQL database is stored intact, forming the data operation layer, also known as the preparation area. This data serves as the source for subsequent data processing. Tables in this layer are named using the format "O_SourceTableName". During import, all char and varchar types are converted to string types, number types are converted to double types, and date and TIMESTAMP types remain unchanged.

3.2. Data Detail Layer

The DWD (Data Warehouse Detail) layer primarily serves two purposes: (1) to clean data from the ODS (Operational Data Store) layer, and (2) to perform dimension degradation and dimension modeling on business data tables to reduce inefficiencies caused by multiple table joins during subsequent queries.

ISSN: 2710-0588

The data cleaning stage involves complementing and correcting fields with null values, incomplete content, and inconsistencies. This process is divided into five steps. Missing value cleaning is done by removing or filling in missing values based on their proportion. For example, if power usage data is missing, it is supplemented with 0. Format content cleaning includes unifying date formats, numbers, and other formats. Logical error cleaning involves removing unreasonable values, such as non-numeric or excessively high-power usage figures. The basic data is modeled using a star schema, involving business process selection, granularity declaration, dimension determination, and fact determination. Driven by business processes, fact tables such as the ship fact table and usage fact table are constructed.

3.3. Data Service Layer

The DWS (Data Warehouse Service) layer aggregates and summarizes data from the data detail layer based on different thematic perspectives to obtain statistical data. The characteristics of this summarization are: (1) determining fields based on the analysis requirements of ship shore power data, (2) aggregating data according to different dimensions, and (3) categorizing requirements and dividing them into multiple tables based on statistical themes. Tables in this layer are named using the format "T_Theme_TableName", with themes named according to specific applications.

4. Automated Scheduling based on DolphinScheduler

For the efficient management and utilization of the ship shore power data warehouse, this paper proposes an automated task scheduling solution based on Apache DolphinScheduler to achieve full-process automation from data collection, storage, processing to visualization. This solution aims to build a flexible and scalable data processing pipeline by integrating various open-source tools and platforms.

4.1. Automated Data Acquisition

Leveraging DolphinScheduler task scheduling capabilities, Sqoop tasks are triggered at scheduled times to automatically extract the required data from the source database of the ship shore power system. DolphinScheduler supports complex dependency management and scheduled task scheduling, ensuring the accuracy and timeliness of data collection tasks.

4.2. Automated Data Storage

After data collection, the data is automatically stored in Hive. DolphinScheduler seamlessly connects the data exported by Sqoop to Hive by configuring the corresponding task flow, realizing automated data storage. This step not only simplifies the data storage process but also enhances data security and manageability.

4.3. Automated Data Processing and Visualization

Following data storage, Spark SQL is utilized for data processing and analysis. The system defines a series of Spark SQL tasks through DolphinScheduler task orchestration functionality to perform operations such as cleaning, transformation, and aggregation on the data in Hive, catering to various data analysis needs. Once processing is complete, the data is visualized using SuperSet, an open-source data visualization tool. SuperSet offers a wide range of chart types and interactive interfaces, making data visualization more intuitive and convenient. DolphinScheduler links data processing and visualization tasks together to form a complete automated process, significantly improving the efficiency and accuracy of data analysis and application.

Through DolphinScheduler automated task scheduling functionality, the system achieves full-process automation for the ship shore power data warehouse, spanning data collection, storage, processing, and visualization. This not only simplifies the data processing workflow but also

ISSN: 2710-0588

enhances the efficiency and accuracy of data processing, providing robust data support for the intelligent management and decision-making of the ship shore power system.

5. Conclusion

Based on DolphinScheduler and Hive, the construction of an offline data warehouse for ships possesses numerous advantages that traditional data warehouses lack, holding significant importance for improvements in the shipping industry and people's lives. By setting HDFS replicas, high data reliability is ensured; by increasing the number of cluster nodes, the issue of insufficient cluster processing capacity is addressed; and by utilizing the Yarn component, cluster resources are effectively managed. Through deep mining of ship shore power data, enterprises can better understand ship operations, promoting the green and intelligent development of the shipping industry.

With the continuous development of the shipping industry, future business is bound to change, and data analysis will also require more detailed granularity. Coupled with limited time and energy, there are still many areas for improvement and refinement in the system. Firstly, to address more new business requirements, ad-hoc queries should be added, allowing for the quick implementation of temporary customer demands. Secondly, a stream processing function module should be incorporated, as a comprehensive e-commerce system should not only include offline batch processing but also possess stream processing capabilities to better empower data and drive industry development. Thirdly, metadata governance and permission management should be added to ensure data security and integrity.

Acknowledgments

This work was financially supported by the funding of the Students Innovation and Entrepreneurship Training Program of Jiangsu Province Insitute (G-2023-0409), and the Excellent Teaching Team for QingLan Project of the Jiangsu Higher Education Institutions of China (Big Data Technology Teaching Team with Shipping Characteristic).

References

[1] Zou, Yujuan, Peiyi Tang, and Taizhi Lv. "Design and implementation of ship shore power data analysis system based on Doris data warehouse." 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2022: 367-370.

[2] Chen, Juntao, Jinmei Zhan, and Fei Tian. "Research on the Construction of a Data Warehouse Model for College Student Performance." International Conference of Pioneering Computer Scientists, Engineers and Educators. Singapore: Springer Nature Singapore, 2023: 408-419.

[3] Zhai, Yanlong, et al. "Hadoop perfect file: A fast and memory-efficient metadata access archive file to face small files problem in hdfs." Journal of Parallel and Distributed Computing 156 (2021): 119-130.

[4] Małysiak-Mrozek, Bożena, et al. "High-efficient fuzzy querying with hiveql for big data warehousing." IEEE Transactions on Fuzzy Systems 30.6 (2021): 1823-1837.

[5] Mantri, A. "Optimizing HDFS Storage and Managing TTL for Unused Hive Tables: Strategies for Improved Data Efficiency." J Artif Intell Mach Learn & Data Sci 2023 1.4: 680-683.

[6] Sleeman IV, William C., and Bartosz Krawczyk. "Multi-class imbalanced big data classification on spark." Knowledge-Based Systems 212 (2021): 106598.

[7] Qiu, Yuanhui, et al. "TsQuality: Measuring Time Series Data Quality in Apache IoTDB." Proceedings of the VLDB Endowment 16.12 (2023): 3982-3985.