Multiple system fusion research Based on Hadoop and ETL

Lili Zhang[1,a,*], Yunwu Wan[2,b] , Lei Wang[1,c] and Hui Liu[1,d]

[1] Anhui Sun Create Electronic Co., LTD, Hefei, Anhui 230088, China.
[2] Hefei Taihe Optoelectronic Technology Co., Ltd, Hefei, Anhui, China.
[a] zhanglily.1225@163.com, [b] yunfan619@163.com, [c] 67685463@qq.com, [d] 121791411@qq.com
*corresponding author

Keywords: Hadoop, ETL, Big data, data cleaning

Abstract: In recent years, urban road transport vehicle dynamic supervision and management of in-depth, constantly strengthen continuously, but also has a lot of problems and insufficiency. the transportation management office for different business management requirements have different management systems, Most of these systems and database are independent of each other, Each system is unable to share data, so We can't make full use of each system data to complete the dynamic regulation of road transport. Based on Hadoop and ETL technology, through the establishment of big data platform and data exchange sharing platform and data mart to implement the data cleaning and processing business. Data cleaning and processing business priorities to finish the unification of the huge amounts of data storage, management, information sharing and service to provide data resources. And as a support of application system, in view of the different business set up different topics. Establish a perfect architecture for data acquisition, loading, storage, analysis and application show, complete system data fusion, comprehensive dynamic regulation of road transport.

1. Introduction

In recent years, urban road transport vehicle dynamic supervision and management of in-depth, constantly strengthen continuously, but compared with the national policy and the current situation, there are many problems and shortcomings, mainly displays in: Administrative department of the lack of information management methods, passive management of transport enterprises, practitioners to avoid management, and vehicle dynamic loss in management problem is common. The existence of the above problem will severely affect and restrict the development of road transport safety management work, must cause enough attention, and take concrete and effective measures to solve them. At present most of the urban transportation management office for different business management requirements have different systems or platforms, most of these systems running independently, each system are independent of each other, the database is independent, the system of data cannot be Shared, so We can't make full use of each system data to complete the dynamic regulation of road transport.

This paper mainly introduces using Hadoop technology and ETL technology to complete the data fusion of multiple system, realize the dynamic regulation of road transport. Through the data cleaning and processing business priorities completed the unification of the huge amounts of data storage, management, information sharing and service to provide data resources, and as a support of application system, in view of the different business set up different topics, establish a perfect architecture for data acquisition, loading, storage, analysis and application show. Data cleaning and processing business is mainly through the establishment of big data platform, data exchange and sharing platform and data marts.

2. Based on Hadoop and ETL implement multiple system data fusion

Using Hadoop technology and ETL technology for multiple transportation administration system of the unity of the huge amounts of data storage, management, information sharing, realize the dynamic regulation of road transport, mainly to complete the data cleaning and processing business. Data cleaning and processing business priorities to finish the unification of the huge amounts of data storage, management, information sharing and service to provide data resources. And as a support of application system, in view of the different business set up different topics. Establish a perfect architecture for data acquisition, loading, storage, analysis and application show. Data cleaning and processing business is mainly through the establishment of big data platform, data exchange and sharing platform and data marts.

2.1. Data cleaning and processing

In order to deal with limited amount of structured data, choose the traditional ETL platform[6]. For the huge amounts of data or a large number of semi-structured data and unstructured data, ETL process using build ETL platform based on Hadoop[7]. ETL process includes the following categories: data cleaning, data conversion, data gathering.

Data cleaning: implementation of the standardization of business data, have to remove duplicate records, replace treatment and remove the invalid data, and other functions. To different sources of business data cleaning and transformation, converts data under different standard into the data that conform to the data standard and the data definition of data platform.

Data conversion: Unstructured data into structured data, through the business system in the interpretation of log files, implement the conversion of unstructured data to structured data, finally saved to the MPP database cluster. Low value density data conversion to high value density data. To low value density of unstructured data such as audio, video and images, through metadata extraction characteristics, the characteristics of the data saved to the MPP database cluster, so as to realize the conversion to high value density of structured data[4]. Under the action of metadata, By loading the extracted feature information, finally saved to the data warehouse, in order to achieve high performance of query analysis provides the basis.

Data gathering: data gathering process includes data split and merged, make a preliminary summary for the different sources of data, form a complete data set. Each business system data merging, business system data with unstructured data conversion to merge, log analytical data merged with pictures, audio and video feature extraction data. Remove duplicate fields, break the paradigm, the original entity set together into a data set.

2.2. Big data platform

Data cleaning and data processing software need to build a big data platform, to integrate optimization inside and outside of all kinds of information resources, form the basis of the repository, and on the basis of the basic data repository construction, through secondary extraction, Indexed integration and logic, construction form application service repository. Big data platform architecture uses hadoop technology[5], through the hadoop distributed file system (HDFS) for data distributed storage, for business scenarios, using Hbase database to storage the data that need random access, real-time read and write. At the same time the use of parallel processing technology such as MapReduce/Storm for parallel computing[1]. Big data frame design as shown in figure 1.

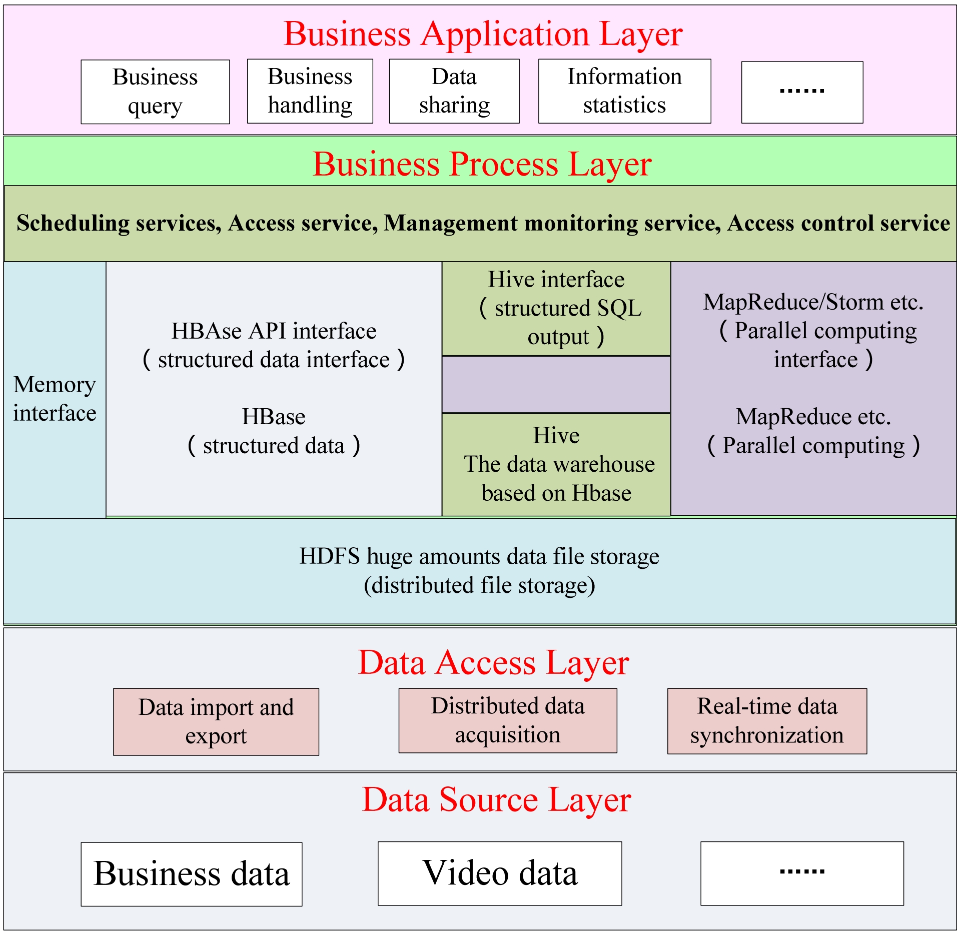
Business Application Layer
Business query
Business handling
Data sharing
Information statistics
......
Business Process Layer
Scheduling services, Access service, Management monitoring service, Access control service
Memory interface
HBAse API interface (structured data interface)
HBase (structured data)
Hive interface (structured SQL output)
Hive The data warehouse based on Hbase
MapReduce/Storm etc. (Parallel computing interface)
MapReduce etc. (Parallel computing)
HDFS huge amounts data file storage (distributed file storage)
Data Access Layer
Data import and export
Distributed data acquisition
Real-time data synchronization
Data Source Layer
Business data
Video data
......

Figure 1 Big data frame design

The data source layer is mainly for the connection of front-end of various data, including the business data, video data.

The data access layer for different types of data sources using different data acquisition strategy, realize the import and export of structured data, unstructured data and semi-structured data.

The business process layer implements the data distributed storage and parallel computing, and provide unified resources scheduling services, access services, management monitoring service and access control service, support the transportation management office each business application.

The business application layer implement business query, business statistics, data sharing, statistical information and other services.

The Hadoop distributed file system: realize the distributed storage of data, hide the lower load balance and the redundant copy details, for the upper application provides a unified file system API interface. HDFS doing a special optimization against huge amounts of data characteristics, including: access of large files, read and write operation scale is too large, the PC easily happened fault caused by the node failure, etc. HDFS split the file into the size of the pieces (adjustable, such as: 64 Mb), distributed on the clusters machine, use file system for Linux or Windows. At the same time each file at least three or more redundant (customizable). Center is a management node (NameNode), according to the file index, search for files block of data nodes(DataNode) [3].

Hbase large database: using distributed, stored in columns, multidimensional table structure of the distributed real-time database. It can provide large amount of structured and unstructured data high speed read and write operations, designed for high-speed data services online.

Utilizing the framework such as MapReduce/Storm implementations parallel processing of complex tasks, such as most distributed computing can be abstracted as MapReduce operation. The Map is resolve the Input into the middle of the Key/Value pairs,

Using MapReduce/Storm framework realization of parallel processing of complex tasks, such as most distributed computing can be abstracted as MapReduce[2]. The Map is resolve the Input into the middle of the Key/Value pairs, The Reduce compound the Key/Value into the final output. Lower facilities distribution Map and Reduce operations on the cluster is running, and the result is stored in the HDFS.

2.3. Data exchange and sharing platform

Data exchange and sharing platform for the different business system database, file system and other data sources provide data integration for extraction, conversion, transmission and load operations. Automatic, convenient and fast to implement the data extraction cleaning, finish application integration based on data.

ETL data exchange platform can be used to quickly realize the information sharing and exchange between different business systems, to realize the integration of the application[8]. ETL platform through efficient data extraction, data cleaning, data conversion, data loading, etc. complete the location data, business data, management data and other business data from a data source to target data warehouse process.

2.4. The data mart

Data cleaning and processing software through the ETL exchange sharing platform will be collected data extraction to the base data center, including the business system of business data and other data of industry inside and outside system, formation of all kinds of data mart[9], according to the type of business is divided into basic database, business database and subject database.

The basic database storage people, cars, homes, and other basic information. The business database storage business is dealt with and approval information. The subject database storage the query results and all kinds of report data that the system generation. The stored data through data management platform implementation inquire and display, through the data analysis platform implementation decision analysis. Data flow is shown in figure 2.

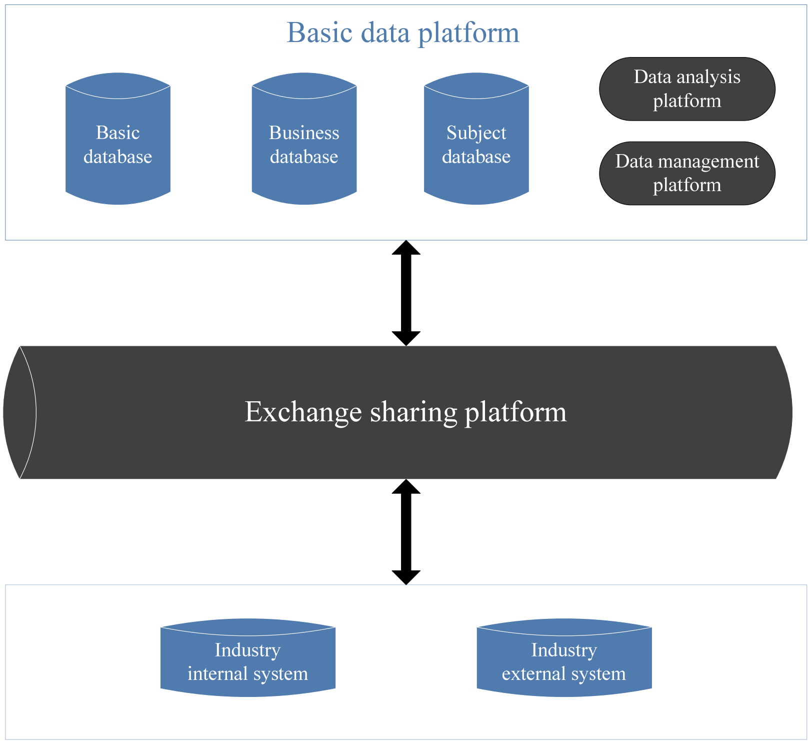

Figure 2 Data flow diagram

Data mart based on business application requirements in terms of construction, including fixed statements, ad-hoc query, OLAP, data mining, etc., to store the data mainly for analytical data. Data mart is a data collection that the data warehouse is the only data source, that is specific analysis of the application, that in a certain way to reorganization, is a subset of the data warehouse. Based on data warehouse is a data mart to create, to the needs of different business units and different analysis of the application of data storage.

3. Conclusions

This article is based on Hadoop technology and ETL technology, through the data cleaning and processing business for multiple transportation administration system of the unity of the huge amounts of data storage, management, information sharing, and realize the dynamic regulation of road transport. Data cleaning and processing business priorities to finish the unification of the huge amounts of data storage, management, information sharing and service to provide data resources. And as a support of application system, in view of the different business set up different topics. Establish a perfect architecture for data acquisition, loading, storage, analysis and application show. Data cleaning and processing business is mainly through the establishment of big data platform, data exchange and sharing platform and data marts. On the big data resource layer, for highways, safety, public security and other relevant government departments and the public to provide all kinds of rich business support and data analysis support, such as peak industry regulations, auxiliary decision-making, illegal operation route, traffic situation, correlation analysis and multi-dimensional analysis, complete the comprehensive dynamic regulation of road transport.

References

[1] Jeffrey Dean, Sanjay Ghemawat. MapReduce:simplified data processing on large cluster[C]. OSDI. 2004,137-150.

[2] Yang H C, Dasdan A, Hsiao R L. Map-Reduce-Merge:Simplified Relation al Data Processing on Large Clusters[C]. International Conference on Management of Data Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, 1029-1040.

[3] Dhruba Borthakur. The Hadoop Distributed File System:Architecture and Design.Apache Software Foundation. 2007.

[4] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh. Bigtable: A Distributed Storage System for Structured Data. In Proc. OSDI. 2006, 205-218.

[5] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google File System. Proceedings of the 19th ACM Symposium on Operating Systems Principles. 2003, 20-43.

[6] Skoutas D. Designing ETL Processes Using Semantic Web Technologies[C]. Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP. New York: ACM, 2006, 67-74.

[7] Vassiliadisl P, Simitsis A, Georgantas P, et al. A Framework for the Design of ETL Scenarios[C]. Proceedings of Conference on Advanced Information Systems Engineering (CSiSE). Klagenfurt: CAiSE, 2003: 520-535.

[8] Tziovara V, Vassiliadis P, Simitsis A. Deciding the Physical Implementation of ETL Workflows[C]. Proceedings of the ACM 10th International Workshop on Data Warehousing and OLAP. New York: ACM, 2007: 49-56.

[9] Zhang Xufeng, Sun Weiwei, Wang Wei, et al. Gemerating Incremental ETL Processes Automatically[C]. Proceedings of the First International Multi-symposiums on Computer and Computational Sciences. Picataway, NJ: IEEE, 2006: 516-521.