Transactions on Social Science, Education and Humanities Research
ISSN: 2960-1770, eISSN: 2960-2262 | Volume 3
ERSS 2023

# On the Application of Big Data in Smart Transformation Enterprises

Shan Ge

Liaoning Communication University, 110136, China

**Abstract.** The arrival of the digital informationization era heralds the issue of technological innovation in the intelligent transformation of enterprises, for different scenarios and different cases, the application of big data plays an important role in the accumulation of massive data, the issue of the reliability of the data and the form of data transformation with the transformation of the enterprise has undergone a significant change in the data processing software platform, and strive to deal with the accuracy and efficiency of the data is a multitude of Enterprises plan to achieve the goal, under the general software of big data, the introduction of data information software platform is undoubtedly an important initiative in today's society in order to meet the massive data research. This paper summarizes the development and application of data information software platforms on the basis of awards, data as a prerequisite, put forward for intelligent platforms to monitor the systematic approach to the integration of the Sqoop framework in the Hadoop platform for the enterprise to formulate an intelligent data monitoring indicators of the processing model.

**Keywords:** Data Analysis; Data Information Software Platform; Information Interaction; Data Iteration.

## 1. Introduction

In view of the multifaceted characteristics of the data information software platform, the issue of multiple data conversion for transforming enterprises will also become an important issue for future development. The compatibility of software platforms commonly used for big data and the terminal display module has always been an important research direction for real-time interaction between data information software platforms and terminals, and the heavy and redundant nature of data information in the processing of real-time terminal display and extraction and cleaning process, the speed of the received can not be improved, which also brings new challenges to the terminal display of real-time information.

## 2. Development and Application of Data Information Software Platforms

### 2.1. Big Data and Cloud Platforms

In the wave of big data, with the help of a variety of software platforms, the collection and mining of data sources make intelligent equipment have a new meaning, in many software, the cloud platform operation mode based on the superposition of the Internet-related services is particularly important, the use of servers and network data delivery mode is usually dependent on big data software platforms and cloud computing platforms to complete the overall operation of the industry chain. The cloud represents the virtualized resources of equipment storage in the network, and the data filtering of big data completed under the support of the cloud platform will make use of various modes such as data centers and virtual machine equipment. In the current cloud platform delivery end, often relying on a variety of payment forms to organize and store data, only to provide on-demand network browsing and access, configure the overall computer resources and equipment sharing pool, the speed of the server to achieve a large number of data iteration, in the virtualization of the equipment service terminal, the use of multi-port network client media, to complete the uploading and downloading of the server, for the capacity of the storage device. The setting of the relevant parameters should be followed up in conjunction with the comprehensive operational requirements of intelligent information, in which the application software architecture for big data is usually based on the

business model of the enterprise to complete the setting of the Hadoop platform architecture on the original basis, the processing of the interface for the cloud platform, making the cloud platform on the operating side more suitable for the output content of the Hadoop platform, which fundamentally solves the problem of sharing the content of the interface of the software framework platform. This solves the problem of sharing the interface content of the software framework platform fundamentally. After the configuration of the underlying equipment, the user experience of the cloud platform can reach 10 trillion iterations per second computing speed, Hadoop platform with its own unique framework advantage to show its charm of processing massive data. In the process of trillions of iterations, the cloud platform utilizes the processing advantages of virtual machines and virtual space to complete the sharing of data, in addition, while the data has a life, it provides the details of the work records for the user-designed service provisioning work, which makes it possible for big data to be processed with only a very small number of staff members who can complete the great amount of work, and the mode of management is greatly simplified, which also reduces the number of users and the number of users. The management model is greatly simplified, which also reduces the chance of interaction between data and service providers, and basically realizes the exchange of data and information with less interaction.

### 2.2. Study of the Generalizability of Data Software

Data processing is basically a server-centered operation mode. Compared with the previous mode, hundreds of personal computers and BladeCenter and System x servers provide 1,600 processors, but there are still problems such as slower processing speeds for the multiple sources of data and the diversity of formats, and based on BladeCenter and System x servers , the system can support multiple open code platforms, including Linux, Hadoop and other diverse platforms. Support for multiple open source code platforms, including Linux, Hadoop and other diverse platforms, and in the 0penStack open source code was introduced into the system integration, Windows Server was injected with new vitality, followed by Ubuntu's version update. In the study of the generality of the new version, the portability of the architecture is often ignored, borrowing virtual devices in a specific environment for the import of Ubuntu, so as to run the massive data resources of big data, which, from the ground up, solves the problem of transferring the data, in the virtual invocation of the data, in the massive data pushed to accumulate, the establishment of the basic framework of the cluster version of the data, in the processing of the data, the phenomenon of congestion in the data will be generated.

### 2.3. Pre-disturbance of "Dirty" Data

In the process of processing data, "dirty" data based on the source data has a certain degree of representativeness, the data source in the collection process, resulting in incomplete, erroneous and duplicate phenomenon of the data from time to time, and this type of data in the processing of the internal data collation will make the final source of missing values or attribute dependence on the conflict and other phenomena, which in the This can lead to misleading duplication of records and affect the quality of the data when investigating the trend prediction of the data after real-time data collection. Generally for such problems will be single data source, multiple data sources and special data source processing to classify and discuss, single data source of the instance analysis layer produces more of this problem, generally manifested in the problem of having similar duplicate record problems, and the existence of individual missing values in the problem, and for the study of the pattern in a single data source, with integrity constraints problem becomes the goal of efforts, a large number of data attribute dependence has an uncertainty in the results. The most difficult is the special data source problem, in this kind of data source for the pattern layer and instance layer division, the data itself attribute problem there is diversity compatibility and discrimination form of difference, and then attached to the integrity constraint problem so that the scope of the data source to lose the expression of the data pattern, the data of the information storage is missing the corresponding restrictions, then a large number of data sources will produce a variety of anomalies, such as inconsistency in spelling. This is far from being able to satisfy the fundamental processing classification task of data sources in the attribute dependency problem.

### 2.4. Research Methodology for Handling "Dirty" Data on Data Information Software Platforms

In order to further eliminate "dirty" data sources, data source legitimacy test is one of the functions added to the data information software platform, in the data source has a certain amount of information to provide, the software platform for the data source specifications for the reliability of the standard test, the basis for judgment based on the intelligent transformation of the enterprise content requirements for the function of the setting, mainly With the data source format standards, data source storage range, data source enumeration list and data source relevance for unified definition, in line with the normative format of the data source should be backed up and fully carried out attribute records, to ensure the effectiveness of the data when calling and cleaning; on the other hand, in the process of being detected, the data source field data whether the intelligent transformation of the enterprise to give the key fields for proofreading, if the deviation of the field problem to be carried out. On the other hand, in the process of being detected, whether the field data of the data source is proofread in the key fields given by the intelligent transformation enterprise, if there is any deviation of the fields, the filtering and decomposition should be carried out; if there is any case that the fields do not match the key fields, the basic program should be called to reload the data; finally, the relevance of the collected data source should be detected in a planned step-by-step manner, in the event of the relevance of the data source and correlation are determined, the value of the relational function of the data source should be filtered out higher than the general estimation of the value of the data fields. Then analyze the data source according to the attributes of the data fields.

## 3. Innovation in Big Data Software Technology

### 3.1. Changes in Data Cleansing

Data sources in the overall extraction process data analysis process requires data cleaning work effectively, and for the huge group of data sources, more and more data sources have the characteristics of multi-source, in the difficulty of data cleaning needs to be further innovated, data cleaning algorithms have also been improved, as shown in Figure 1, due to the problem of the timeliness of the data, the data source in the primary call will generally use the Cluster fast processing method to complete the data organization and cleaning.

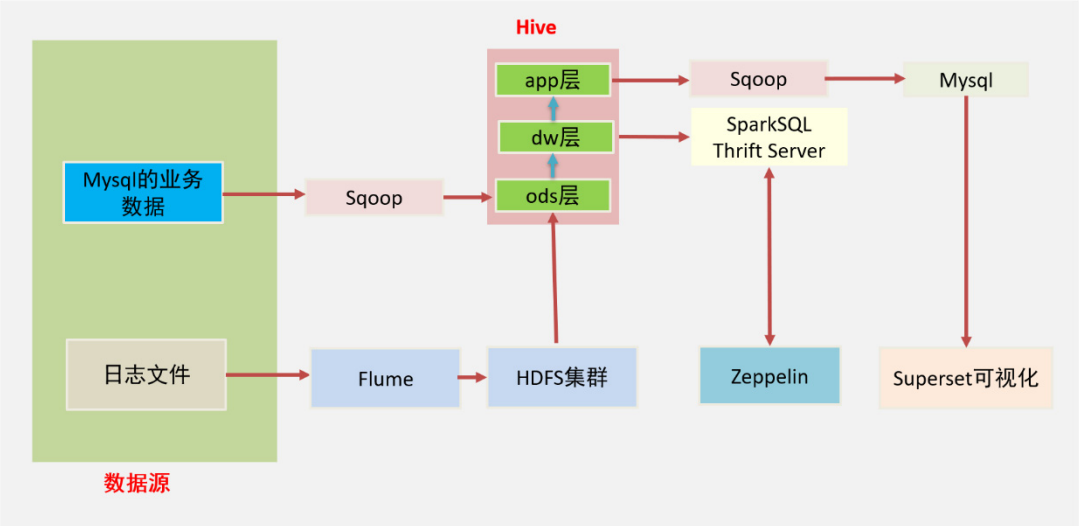


**Fig 1.** The process of calling the module from the data source

Usually in the methods used, searching out abnormal cases is the first step in practicing data source cleaning, the automatic detection function of data has been more mature ideas in the algorithms for data cleaning, but for multi-source data, the advancement of detecting duplicate records in the case of log files being recorded, new algorithms are needed to go along with the completion of the process,

in recent years, clustering algorithms for the classification of multi-source data have made new Advancement, after the improvement of the algorithm, the use of program order detection method, can quickly determine the attribute error of the date source, the applicability of the algorithm is better in the case of the detection of the key field change. Whether it is two data or one data resulting from merging, the algorithm can accurately locate the duplicate records of a real entity and make effective hints when matching. In the experiment, the effect of the matching algorithm with recursive fields is better, in addition to the detection of duplicates problem, try to use the key fields in the intelligent transformation of the enterprise for the multi-mode matching algorithm of the field is also more effective.

## 4. Scenario Analysis of Big Data Application in Smart Transformation Enterprises

### 4.1. Example of a Date Server of Smart Transformation Companies

In a smart transformation enterprise, a large number of data sources are extracted and cleaned and processed, often yielding the expected planning direction for enterprise content transformation, for example, in the results of a survey conducted for aging employees in a smart transformation enterprise as shown in Table 1.

**Table 1.** Rospondents' ideal expectations for pension property

| Items | form | people | percentage (%) |
|---|---|---|---|
| | economic capacity | 78 | 81.3 |
| purchase-oriented | living facilities | 92 | 95.8 |
| | medical care | 89 | 92.7 |
| factor | basic conditions | 95 | 98.9 |
| | supporting conditions | 94 | 97.9 |
| Surrounding | shopping | 81 | 84.3 |
| | entertainment | 72 | 75 |
| | medical care | 88 | 91.7 |
| size of the house | $< 70\ m^2$ | 62 | 64.5 |
| | $\geq 70\ m^2$ | 34 | 36.5 |
| price factor | < 6,000 yuan | 76 | 79.2 |
| | ≥6,000 yuan | 20 | 20.8 |

### 4.2. Discussion of Methods for Processing Case Data

In the raw data under investigation, an attempt is made to perform attribute separation i.e. extraction of key values in its attribute fields, the attributes of the raw data source of the survey results include a wide range of information, where each of them covers a meaning that can in turn be defined as a refined attribute. When the data is cleaned, the duplicate records of the data are screened. If in the process of collecting information, the intelligent data information software platform finds duplicates in the fields or errors that need to be corrected, the platform clustering algorithm is used to follow up and eliminate the wrong data fields. In order not to let the data lose its validity, and there is no problem of omission of data fields, after the data source is eliminated, the data can be rewritten in accordance with the repetition of the algorithm call, the data rewriting step, before this, the need for intelligent data information software platform to import the code word dictionary, and the dictionary can be automatically completed with the query and spelling detection function, which, to a certain extent, perfects the subsidiary functions of the intelligent data information software platform.

## 5. Technical Indicator Analysis of Cases

After the realization of the data collection of the data source, in the traditional database dependence, intelligent data information software platform to call Sqoop to complete the task of data migration, combined with the above enterprises in the research project, the purchase of oriented living facilities accounted for a high percentage of the welfare policy for the enterprise, for the purchase of part of the aging enterprise employees living facilities in the procurement of spare parts data as shown in Figure II.

```
--订单明细事实表采集
/export/server/sqoop-1.4.7/bin/sqoop import  \
--connect jdbc:mysql://node1:3306/itcast_shop \
--username root \
--password 123456 \
--table itcast_order_goods  \
--target-dir
/user/hive/warehouse/itcast_ods.db/itcast_order_goods/ \
--delete-target-dir \
--fields-terminated-by '\t' \
--m 1

--订单退货表数据采集
/export/server/sqoop-1.4.7/bin/sqoop import  \
--connect jdbc:mysql://node1:3306/itcast_shop \
--username root \
--password 123456 \
--table itcast_order_refunds \
--target-dir
/user/hive/warehouse/itcast_ods.db/itcast_order_refunds \
--delete-target-dir \
--fields-terminated-by '\t' \
--m 1
```

**Fig 2.** Data on procurement of spare parts in living facilities

In the case, HDF8 is used to migrate the indicators of the data source to the relational database, according to which the data statistics form is built to complete the statistics and cycle call, which is based on the use of Sqoop, but also combines the migration of statistical data to the MySQL prognostic table.

## 6. Expectations and Prospects for Smart Big Data

With the enhancement of the cleaning ability of the data source and the iterative development of a variety of big data software platforms, there will be further algorithmic improvements in the intelligent data functions of big data, the use of clustering algorithms for data classification and cleaning methods, combined with Sqoop's means of data migration to complete the data records and attribute classification, which will help enterprises in the period of intelligent transformation in response to the arrival of the era of big data in a variety of data challenges issues.

## References


[1] Liu Piao, Cheng Donghui, Gao Qiqi, Lu Chen. Recommendation system based on big data job analysis[J]. Smart City, 2021,7(10):13-14.

[2] YANG Yibo, WANG Femxin, RANG Hao. Research on metwork big data technology for cloud computing and internet of things[J]. Computer Knowledge and Technology,2021,17(24):03-04.

[3] Xue Yan, Tang Tuo. Evaluation of online-offline hybrid teaching effectiveness with big data analysis[J]. Information Technology, 2021,(08):70-74+80.

[4] Shuqi Chen, Ran Zhao, Shuwen Ren. Big data analysis of chicken market based on multiple evaluation scales[J]. Academic Journal of Engineering and Technology Science, 2021,4(4).

[5] Lv Xiaozhan, Cai Xiaojing, Zhou Ping. Visualization and analysis of big data research collaboration networks in the context of technology globalization[J]. Science and Technology Managenent Research, 2021,41(10):20-30.

[6] Song Lining. Implementation and application of big data center based on cloud computing[A]. Tianjin Electronics Society, Tianjin Instrumentation Society. Proceedings of the 33th China (Tianjin) 2021'IT, Network, Information Technology, Electronics, Instrumentation Innovation Conference[C]. Tianjin Electronic Society, Tianjin Instrumentation Society: Tianjin Electronic Society, 2021:5.

[7] Sun Ye. Big data analysis platform technology for manufacturing enterprises[J]. Electronic Technology and Software Engineering, 2021.(16):178-179.

[8] He Yi. Research on big data relevance mining technology[J]. Computer Knowledge and Technology, 2021, 17 (23):23-24+31.

[9] Wen Haolin. Exploration of economic innovation management strategy of enterprises in the mew period[J]. Finance and economics, 2021,(23):57-38.

[10] Li Yan. Analysis of data security issues based on big data cloud computing network environment[J]. Wireless Intercommuting Technology, 2021,18(15):19-20.

[11] Huang Botao. Strengthening platform regulation to promote data security management [N]. Economic Reference News, 2021-08-10(008).

[12] Qin Zhian. An overview of the informatization construction of big data application in human society system[J]. Finance and economics,2021,(22):65-66.