Some Key Problems of Data Management in Army Data Engineering Based on Big Data

Xiao HongJu, Wang Fei, Wang FenMei, Wang XiuZhen

Department of Computer Science, Army Officer Academy, Hefei, China

Email:181594137@qq.com

Abstract—**This paper analyzed the challenges of data management in army data engineering, such as big data volume, data heterogeneous, high rate of data generation and update, high time requirement of data processing, and widely separated data sources. We discussed the disadvantages of traditional data management technologies to deal with these problems. We also highlighted the key problems of data management in army data engineering including data integration, data analysis, representation of data analysis results, and evaluation of data quality.**

Keywords-army data engineering; data management; data integration; data analysis; representation of data analysis results; data quality

I. INTRODUCTION

The current information society has entered the era of big data. In March 2012, the United States government issued a "big data research and development initiative" [1], launched a $200 million investment big data development plans. The U. S. government's plan is seen as another major move in the field of information after the u.s.. Big data, referred to by the conventional software tools in a certain range of time capture, management and processing of data, is the need for new processing mode to have a stronger decision-making ability, insight discovery and process optimization capabilities of massive, high rates of growth and diversification of information assets. Big data 5V features (IBM): Volume (large), Velocity (high speed), Variety (diversity), Value (low value density), Veracity (authenticity).

In the era of big data background, the construction of our army data project is faced with many opportunities and challenges. Along with the development of our army equipment and information technology, the army has produced a large amount of data, such as combat training data, weapons experimental data, equipment support data, personnel preparation data, weapons configuration data, etc.. And with the passage of time, the type of data will be more diverse, the amount of data will be greater, so the data is growing at a rapid pace and accumulation. It can be said that in the era of big data, the data is fighting. If you can't effectively manage these existing data and new data, you can't play the value of these data. At present, we put forward the following challenges to the data management of our army data project:

First,the data size is huge. With the series of new equipment, the extensive application of information technology, combat training in an unprecedented expansion of the scope of time and space, etc., Army data works in the amount of data far beyond the limits of traditional data management system can manage.

Second, the various types of data. Army data engineering usually contains a wide variety of data types, including traditional structured data, such as data stored in a relational database, and including text, images, sound, video and other multimedia data, even include a variety of network resources. These data in different structure, and some of the data model is not even mode. This brings a lot of inconvenience to the management of data.

Third, generation and fast data update frequency and timeliness of data processing strong. The frequency of data generation and updating is fast, and the timeliness of data processing is strong. With the improvement of the level of informatization of our army, arms are faster than ever to produce various types of data; In addition, military activities and requirements "soldiers rapidly", there is often a time limit for data processing. If you can not be completed within the prescribed time limit, it will lose valuable fighter.

Fourth, wide distribution of data sources. The Army is a huge military services, And the diversification of combat style, resulting in a wide range of data sources in the data engineering. In particular, relief troops maneuvering and deployment of travel or movement in the process of data, may change at any time, this has resulted in a wide range of data sources, in some cases, may not be able to timely collection and management.

Fifth is the lack of data and noise data. Due to the limitation of measuring equipment or data collection process vulnerabilities are likely to lead to problems of army weapon system operational data of the process, resulting in the army engineering data missing data and noise data etc.. Good data quality is the basis of data analysis and mining, and these imperfect data have a crucial impact on the results of data mining.

Therefore, in army data engineering, how to effectively manage the data, how to analyze the valuable information from the given data and how to present the results of the analysis, is an urgent problem that must be solved in the army data project at present. In fact，traditional data management technology, especially the database technology, which is represented by the relational database, has been difficult to meet the demand of data management in modern army data engineering. Specific performance in the following aspects:

First, can't meet the requirement of data management in army data engineering. Especially when the management of the data is distributed in a number of nodes, with the increase of the node size, the constraint of consistency and integrity in the relational database can greatly reduce the performance of the database, the data query response is slow and can not meet the needs of the actual use of the data.

Second, can't meet the heterogeneous needs of data management in army data engineering. Although the traditional relational database technology is very good at dealing with structured data, but the data engineering involves a large number of semi-structured, unstructured data is often inadequate.

Third, can not effectively deal with the data of the army's data in the absence of patterns and patterns of the situation. The traditional relational database system requires the data to be in the first mode before it can be stored. But for the army data project data, this limit is too strong, because many data patterns are missing or even no pattern.

In summary，we think, construction of the army data project to draw on the latest achievements of the current big data research. This paper will focus on the research status of data management in Army Data Engineering, focus on the following several key scientific issues based on large data perspective:(1) Data integration;(2) Data analysis;(3) Presentation of data analysis results;(4) Data quality evaluation.

II. DATA INTEGRATION

Data integration is the basis of data management in army data engineering. The goal of data integration[2] is to query a variety of different data sources by providing a unified object model interface to meet the needs of the user's query. In the army data project, the data sources are often distributed in a number of different locations, and may use different data models, the heterogeneity of data makes it difficult to use and manage these heterogeneous distributed data. This requires a large number of heterogeneous distributed data integration. Data integration usually involves the following steps:

First, data preprocessing. Real world data is generally incomplete, dirty data inconsistencies, data mining can not be directly, or just passable mining results. In order to improve the quality of data mining, data preprocessing is needed.This including denoising, the noise data of the redundant data redundancy, and to clean the dirty data, to ensure the quality and reliability of data integration.

Second, schema mapping [3]. Schema mapping is a mapping between source and target patterns. It is usually required to give a set of mapping rules to define how the related elements in the source schema correspond to the related elements in the target schema. At the time of schema mapping, particular attention is paid to the processing of the data in the absence of patterns and the absence of patterns, semantic consistency and semantic conflict before and after model transformation.

Third , query processing. Data integration system is responsible for the user to the overall pattern of the query to convert multiple source pattern query, query rewriting. For lossless source schema mapping, the equivalence of query rewriting can be easily maintained. But in the army data project, this kind of ideal mapping is often not exist, more mapping is lossy, namely, the mapping is an incomplete relation from partial source mode to partial global mode. In this case, the data source and its mapping just provide an incomplete view. So, it is necessary to deal with the problem of query problem in lossy mapping efficiently.

In addition, in order to solve the problem of heterogeneous data in Army Data Engineering, it is also necessary to deal with the problem of uncertainty in data integration. According to the hierarchy of data integration，the main source of uncertainty[4] in the data integration in the army data engineering:(1)The uncertainty of the source data to be integrated. Such as battlefield sensor networks or radio frequency data generated, due to the impact of equipment or the environment, are often uncertain; Some of the data due to time, technical or policy reasons, can not get accurate data, such as data loss in the database, real-time data, sensitive data, by the relevant policies to protect the data. (2) Uncertainty of schema mapping. In many cases, it is difficult to give a mapping between the source and target models; In addition, with the development of data integration system, the change of the original source pattern, the new source mode and the need of application lead to the change of the target pattern, it is more difficult to maintain a deterministic schema mapping relationship. (3) Query transformation uncertainty. For the query submitted by the user, the data integration system needs to be converted to the query of the source pattern, in many cases, it will cause the incomplete matching between the initial query and the transformed query, which will cause the uncertainty of query transformation. To sum up, in the process of data integration, the data integration problem becomes more and more difficult and challenging because of the introduction of uncertain factors. How in the data integration processing of the data itself, the uncertainty source model and target model mapping and the uncertainty of the data integration system query conversion uncertainty is very challenging.

III. DATA ANALYSIS

Army data engineering construction of its purpose is to better use of existing data, and extract valuable information, auxiliary operations, training, management, security and other aspects of military activities. As the data in the army data project has the aforementioned characteristics, traditional data analysis techniques[5], such as data mining, machine learning, statistical methods, it is difficult to adapt to the new features of the data, therefore, it can not be directly applied to the analysis of the data in the army data engineering, it is needed to solve some common problems:

First, the traditional data analysis algorithm must solve the problem of performance in large data environment. For the traditional data analysis algorithm, the data size is not very large, its performance is often no problem. But with the increase in the size of the data processing, its performance tends to decline [6], especially in the case of large data processing army data, the problem is often more prominent. In addition, for real time application of army data

engineering, in the start time of the task, the completion time and so there are mandatory. Therefore, it is a new challenge that how to improve the traditional data analysis algorithm in order to meet the requirements of mass data processing in army data engineering. In this regard, we can learn from the latest results of large data research, as the new C-AMAT mathematical model[7], it is through the parallel memory to read the way, the computer's cache and memory system parallel processing, can greatly reduce the computer in the data read the delay.

Second is the traditional data analysis algorithm must solve the data noise problem in big data environment. In the construction of Army Data Engineering, data from multiple heterogeneous data sources are often integrated, these data in many aspects, such as patterns, formats, and other aspects are not the same, even different. In addition, in the process of data collection, due to the transmission, conversion, processing and other links, often can cause data inconsistency or conflict. The traditional data analysis algorithms are often helpless in the face of such heterogeneous and large amount of noise data. Therefore, how to improve the traditional data analysis algorithm in order to adapt to the large number of noise in the data environment, is also a new challenge.

IV. PRESENTATION OF DATA ANALYSIS RESULTS

Army data engineering is ultimately to serve the military decision-making and other military activities, for the results of data analysis, if not in a reasonable, easy to understand and use the presented to various decision-making personnel, will reduce the use of the results, or even misleading. The presentation of the results of the data analysis is mainly:

First is in the form of text. This form of presentation is mainly used to display some simple query results, such as weapons and equipment parameters, etc..

Second is in the form of graphics rendering. For some of the more complex analytical results, such as the use of a variety of weapons and equipment, the combat style of the multi agent system, the results of the text in the form of analysis is not only lengthy, but also difficult to understand. If you can use the graphical form, the image is intuitive, so that people at a glance. But how to present the analysis result with the graphical form, is facing many technical problems, such as the construction of graphics, rendering, updating and how to reasonably determine the complexity of the graphics, these problems are often closely related to time and space, and need to pay special attention to their time efficiency and space requirements.

Third is to provide an interactive way, so that the command and decision-making and other types of users can interact with data analysis results. The user can operate on the original analysis results, for example, the replacement of a war force, increase a kind of weapons and equipment, change a combat style, etc., in order to visually observe the modified analysis results. The results of this analysis are particularly suitable for the use of the use of personnel gradually clear and detailed a strategic and tactical expectations.

V. DATA QUALITY EVALUATION

How to evaluate the quality of data in army data engineering is a very important problem, this is related to the quality of the army data engineering construction. If you can give a set of evaluation indicators of the quality of the data, it will provide scientific guidance and guidelines for data engineering. In this regard, research on evaluation indicators and research and development of standard test tools should be intensified. Based on the evaluation of traditional data management, we can give the following evaluation criteria from the qualitative aspect:

First is army data engineering data management system to contain the ability of the source data. That is to say, after the integration of the data, the data is just as direct query the original data, data without omission, no repeat. Data management system to provide users with a transparent, completely to the mode mapping, query, the results of the work of shielding the merger.

Second is the Army data engineering data management system to reduce the redundancy of the data. That is, the less data in the data management system, the less the better, less redundancy has an important role in the maintenance of data, especially for the consistency and integrity of the data. We believe that the army data engineering, according to its support for the different applications, but also need the following evaluation criteria: integrity, consistency, accuracy, timeliness. To assess whether the data to achieve the desired quality requirements, you can judge these four aspects.

Integrity refers to the existence of missing data information, the lack of data may be missing the entire data record, data may also be a field of information records missing. Incomplete data can draw on the value will be greatly reduced, data quality is one of the most basic evaluation criteria.

Consistency is whether the data follow a unified specification, whether the data set is maintained in a uniform format. The consistency of data quality is mainly reflected in the data records and the data is consistent with the logic.

Accuracy is the index according to record the information whether there is abnormal or incorrect. And consistency is not the same, the accuracy of the data is not only the problem of inconsistent rules. The most common error data accuracy as garbled. Secondly, the abnormal large or small data is also inconsistent with the conditions of the data.

Timeliness refers to the data from the production to be able to view the time interval, also known as the delay of data. Timeliness for the data analysis itself is not high, but if the data analysis cycle plus the data set up for too long, it may lead to the conclusion of the analysis has lost its significance.

Third is the army data engineering data management system to be easy to use. Army data engineering management system is the service at all levels of combat, training and security personnel,is for people to use, therefore, it must be easy to use the man-machine interface, easy to operate, easy to maintain and so on.

Fourth is the army data engineering data management system security is better. Army data engineering management system application environment is often harsh, this is not only manifested in the physical and natural environment of the harsh, but also performance in the face of it will be the time to face the physical destruction of the enemy, technical attacks, etc.. Therefore, it must be robust, secure and highly fault-tolerant.

VI. Concluding Remarks

Army data engineering construction is related to our army and even the whole army fighting a major event, is a systematic project. So in the construction process, only by using the latest scientific research methods and the latest research results, can we build and use well. This article follows the current development trend of information technology, especially the large data technology in data management of the latest research trends and achievements, in view of the challenges faced by data management in Army data engineering, such as large-scale data, diverse data types, fast data generation and updating, high timeliness of data processing, and wide distribution of data sources, this paper discusses some key issues of data management based on large data perspective, including data integration, data analysis, the presentation of data analysis results and data quality evaluation. We believe that the study of this paper can give the army data engineering construction, especially in the data management to provide some ideas of construction and management of the relevant research has some reference.

REFERENCES

[1] Big Data Initiative-The White House. www.whitehouse.gov/sites/.

[2] M. Lenzerini. Data integration: a theoretical perspective. In: Proc. of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems, ACM, New York, NY, USA, 2002, pp. 233-246.

[3] Lv Teng, Wang Qunshan. Journal of Artillery Academy, 2010, 30 (6): 101-103.

[4] X. Dong, A. Halvey and C. Yu. Data Integration with Uncertainty. The VLDB Journal, 2009, 18(2): 469-500.

[5] Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Technologies and Challenges. Computer Research and Development, 2013,50 (1): 146-169

[6] A. Rajaraman, J. Ullman. Mining of massive datasets. http://infolab.stanford.edu/~ullman/mmds.html

[7] Chinese American experts proposed a substantial increase in computing speed of the new model. http://news.xinhuanet.com/tech/2014-03/31/c_1110023038.htm