Battlefield Situation Recognition Technology and Application Trend Based on Data Lake

Lejiang Guo*, Wenjie Tu, Fangxin Chen and Lei Xiao

The Department of Early Warning surveillance intelligent, Air Force Early Warning Academy, 430019, Hubei, China
Email: radar_boss@163.com*, Fangxin@163.com, Shaoxiong@263.com, Yahui@263.com

Abstract. Battlefield situation is the basis of cognitive consistency of battlefield situation understanding. In the service of situation information organization, displaying and distribution is necessary to solve the problem of situation information difference between users at different levels and different combat tasks. Through the research of data Lake technology combined with advanced data science and machine learning technology, the situation recognition and processing architecture based on data lake is proposed to solve these problems. It includes the network architecture, the situation cognition and processing methods which can improve various types of data storage and big data processing problems.

1. Introduction
Battlefield situation map is an organic combination of battlefield situation elements. It is based on the time synchronization and spatial registration of multi-source situation information. It integrates a unique and unambiguous situation. It ensures the consistency of situation information in the process of description, storage, update, query and distribution by a certain multidimensional organization and management mode. The common problems in battlefield situation map include incomplete information, information overload, general rough up and down, and state without potential.

The data lake is a centralized repository that allows all structured and unstructured data to be stored at any size. It can store data without first having to structure the data and run different types of analysis. From control panel and visualization to big data processing, the real-time analysis and machine learning to guide better decision-making. When there is enough computing power to process and analyze all types of data, the analyzed data will be stored for users to use. Process structured data and convert them into multidimensional data or report to meet the requirements of advanced report and data analysis.

2. The Current Situation of Battlefield Situation Generation Technology
The generation of battlefield situation including discovery, collection, processing to system integration is actually a process from point to line and line to surface, from single and local images to common tactical map and then to user-defined operational diagram, so that the battlefield situation has the sharing, consistency, overall situation and accuracy. With the development of technology, there are some new problems.

The data is high-dimensional and high-density. The use of new protocols and the invention of new modes lead to the unstructured data becoming higher dimensional. The situation information from different information sources should be consistent in the description process to avoid possible inconsistent description.
Data storage consistency is difficult. In the process of organizing and storing massive situation information, the storage error of hardware, system, environment and accidental human should be controlled within the allowable range.

Data is nonstandard isolated. The image data comes from the nonstandard of equipment and operation, and the difference between the data is very big. There is a so-called distribution drift phenomenon.

Consistency of situation information is update. Due to factors such as time, update mechanism, synchronization mode, it may lead to update error of situation information in the process of situation information updating.

The annotation is sparse and noisy. Image data annotation is time-consuming and expensive which will result in the sparsity of annotation at the same time. Different tasks need different forms of annotation. The labeling is usually different from person to person. There is a big difference between different people. The labeling is usually noisy. The establishment of the gold standard is also an outstanding issue.

The samples is different and uneven. In the labeled data samples, whether positive samples or negative samples, individual samples have large differences and their probability distribution presents typical multimodality. At the same time, the proportion of positive samples and negative samples is extremely unbalanced.

The tasks are complex and diverse. Unstructured data computing has many tasks. At the technical level, there are reconstruction, enhancement, restoration, classification, detection, segmentation, registration, etc. these technologies combined with the diversity of image modes and tasks produce endless application scenarios and high complexity tasks.

Consistency of situation information distribution is various. To avoid the influence of distribution mode design, distribution scheme implementation error, distribution situation update mechanism and other factors lead to non-uniformity of situation information distribution.

3. Big Data Solution Based on Data Lake
Data lake is a kind of system or storage that stores data in natural or original format, usually object block or file. Data Lake includes structured data, semi-structured data, unstructured data and binary data from relational database. Data swamp is a kind of degraded and lack of management data lake. Data swamp is either inaccessible or unable to provide enough value for users. Focusing on the key nodes and important targets in the battlefield space, it provides the targeted situation products for users through reasonable organization of situation information.

Data inflow is the beginning of the whole data Lake construction, including the inflow of metadata and the inflow of business data. Metadata inflow includes data source creation and metadata fetching which will eventually form data resource directory and generate corresponding security settings and
access control policies. Determine the data source to access, complete the data extraction and incremental access.

As the centralized storage of the whole data lake, data precipitation can be expanded on demand or counted by usage including data source, data type, data form, data mode, total data and data increment.

The starting mode of tasks can be divided into manual triggering, timing triggering and event triggering. Through the use of depth generation model and confrontation training, deep learning opens up a new way for the research of complex image synthesis tasks. With the birth of large and diverse image data sets, the usability of imaging research has been increasing, which is attributed to the deep learning technology, which not only improves the accuracy and versatility but also reduces the reasoning time and the need for complex preprocessing. In particular, convolutional neural network realizes effective network parameterization and spatial invariance and reduces the demand for knowledge in specific fields. This is very important in processing high-dimensional neural image data, and has been proved effective in a series of prediction and analysis tasks. It is suggested that batch processing requirements and SQL processing capacity should be considered in the calculation engine.

![Data Lake Component Reference Framework](image)

**Figure 2.** Data Lake Component Reference Framework

In addition to providing basic batch computing mode, various external computing engines are used to provide rich computing mode support and interactive batch processing capability based on computing capabilities, including stream computing capability and machine learning capability.

Battlefield situation consistency should first ensure that the comprehensive battlefield situation information obtained by all nodes is consistent including accurate configuration information, action information, target information and task status of all combat forces. secondly, situation data should be shared in the whole network, updated in real time and granularity. At the level of strategy, campaign and tactics, the granularity is adjusted in real time through the aggregation and decoupling of situation information.

The data Lake solution covers all functions except quality management and data governance. In fact, the work of quality management and data governance is strongly related to the organizational structure and business type of the enterprise. Therefore, it is understandable that the general solution does not cover this content. If there is a strong demand for quality management and data governance, it can be customized and developed.

4. The Architecture Features and Key Technologies

The consistency generation technology of battlefield situation map mainly includes standardizing the situation information customization requirements of users or combat units from the aspects of information content, granularity and timeliness, generating unique and unambiguous situation
information through multi-source fusion processing, ensuring the situation information in describing, storing, updating and checking multi-dimensional organization and management mode Consistency of inquiry and distribution.

The service model based on data Lake gives users three capabilities including Data capitalization, analysis modeling and service customization.

Data capitalization capability is that the data can be continuously precipitated by using the data lake and the customized requirements of users or combat units for situation information can be described in terms of information content, granularity, timeliness. The unique and unambiguous situation information of the system can be generated, and the consistency of situation information in the process of description, storage, update, query and distribution can be ensured by means of multi-dimensional organization and management.

Data lake also provides data asset management capabilities, businesses can not only manage the original data but also store the processed process data and result data separately which greatly improves the value of buried point data.

Analysis of modeling ability is not only original data in the data lake but also models of buried point data. The buried point data model reflects the abstraction of business logic in the global data intelligent service platform. It takes cognition as the basic clue to establish the internal relationship among data, information, knowledge and wisdom, which is a deep integration of situation information.

Service customization capability includes the data integration and data development capabilities provided by the data lake based on the understanding of the buried point data model. The original data is processed iteratively and valuable information is extracted from the data. The value beyond the original data analysis service is obtained.

Starting from the data application can be used by business in the process of data ETL. At the same time, form the data model, index system and corresponding quality standards. Data Lake emphasizes the storage of original data, the exploratory analysis and application of data, the understanding and abstraction of business will greatly promote the development and application of data lake. Data Lake technology makes data processing and modeling retain great agility, and it can quickly adapt to the development and change of business.

5. Conclusion and Prospect

This paper introduces the related concepts and main characteristics of battlefield situation system. It describes the situation processing process and key technologies for the development of battlefield situation analysis and construction research ideas. As a new generation of big data analysis and processing infrastructure, data Lake needs to go beyond the traditional big data platform. In the following aspects, it is the possible development direction of data Lake solution in the future.

Cloud native architecture of storage and computing are separated, computing power and storage capacity can be independently expanded. multimodal computing engine support, SQL, batch processing, streaming computing, machine learning, etc. Server state services are provided to ensure sufficient flexibility and support pay on demand. Strong data management capabilities. Data Lake needs to provide more powerful data management capabilities including data source management, data category management, processing flow arrangement, task scheduling, data traceability, data governance, quality management, authority management, etc. The management and support of various heterogeneous data sources, the support of full incremental migration of heterogeneous data, and the support of various data formats are the directions that need to be improved continuously. Deep integration and integration with business. The composition of typical data Lake architecture has become a consensus in the industry. The key to the success of the data Lake solution lies in the data management. Whether it is the management of the original data, the management of the data class purpose, the management of the data model, the management of the data authority or the management of the processing tasks, all of which are inseparable from the adaptation and integration with the business.

In the future, more and more industry data Lake solutions will emerge, with data scientists and data analysis development and interaction. The analysis model and customized algorithm in the data Lake solution may be a key point in the future differentiation competition of the data Lake field.
6. References

[1] Jayesh Patel, "The Democratization of Machine Learning Features", Information Reuse and Integration for Data Science (IRI) 2020 IEEE 21st International Conference on, pp. 136-141, 2020.

[2] Dirk-Jan van Veen, Ravi S. Kudesia, Hans R. Heinimann, "An Agent-Based Model of Collective Decision-Making: How Information Sharing Strategies Scale With Information Overload", Computational Social Systems IEEE Transactions on, vol. 7, no. 3, pp. 751-767, 2020.

[3] Elisabeta Zagan, Mirela Danubianu, "Data Lake Approaches: A Survey", Development and Application Systems (DAS) 2020 International Conference on, pp. 189-193, 2020.

[4] Lauri Loven, Ella Peltonen, Abhinay Pandya, Teemu Leppanen, Ekaterina Gilman, Susanna Pirttikangas, Jukka Riekki, "Towards EDISON: An Edge-Native Approach to Distributed Interpolation of Environmental Data", Computer Communication and Networks (ICCCN) 2019 28th International Conference on, pp. 1-6, 2019.

[5] Faisal Ahmad, Anirban Sarkar, Narayan C Debnath, "QoS lake: Challenges design and technologies", Recent Advances in Signal Processing Telecommunications & Computing (SigTelCom) International Conference on, pp. 65-70, 2017.