International Journal of Data Warehousing and Mining
Volume 21 • Issue 1 • January-December 2025

Research on Application of Electromagnetic Environment Data Warehouse Based on Big Data

Jie Wang
22nd Research Institute of China Electronics Corporation, China

Zhenxing Li
https://orcid.org/0009-0005-7528-8146
22nd Research Institute of China Electronics Corporation, China & Xiamen University, China

Liping Zhou
Qingdao Preschool Education College, China

ABSTRACT

The widespread use of electromagnetic space utilization technology across various fields—including maritime, terrestrial, aeronautical, orbital, electrical, and telecommunications—has generated vast amounts of electromagnetic environment data. To manage challenges, such as the storage of large raw datasets, data discrepancies, isolated data across multiple pathways, and low data value density, a big data-based electromagnetic environment data warehouse is proposed. This warehouse standardizes data from diverse sources, integrates and reconstructs it according to business themes, and uses a mix of relational and non-relational databases for storage. It meets the needs for high data reliability, fast access, and massive storage capacity, offering a solution to data overload while supporting data mining and knowledge discovery in the electromagnetic field.

KEYWORDS

Electromagnetic Environment, Data Warehouse, Big Data

INTRODUCTION

With the continuous improvement in the informatization of electromagnetic environment applications and the long-term operation of related businesses, various departments have accumulated vast amounts of electromagnetic environment data (Chen & Zhao, 2020). This data has played an important role in radio management, spectrum control, electronic warfare, and other areas (Cheng et al., 2019; Ding, 2015; Ha & Jia, 2015). However, there are also many challenges in terms of comprehensive storage and application of the data (Group, 2017):

1. Difficulty in storing massive raw data: Electromagnetic environment sensing devices have powerful data collection capabilities, with fast data output frequencies and diverse data types. Over long periods, these devices accumulate large amounts of data, and existing information systems are unable to meet the storage demands.
2. Large differences in data elements: Currently, China has been building electromagnetic environment sensing devices for different business applications. Since these devices are produced

DOI: 10.4018/IJDWM.373715

by different manufacturers and follow different data transmission protocols, the resulting electromagnetic environment data varies in source and structure.

3. Data silos: The existing electromagnetic environment data is non-integrated, with massive amounts of data stored independently in each business system. This leads to inefficiencies in data sharing and interoperability, and there is a lack of integrated data applications.
4. Low data value density: Short-term data is insufficient to reflect the state of the electromagnetic environment, and effective analysis can only be achieved using long-term, large-scale data. Therefore, developing a standardized solution with fast storage and highly efficient processing for electromagnetic environment data has become an urgent need to enhance the value of these data applications.

As a subject-oriented, integrated, stable, and time-variant collection of data used to support management decisions, a data warehouse enables the integration of multiple heterogeneous data sources. It can reorganize data according to themes to meet online analytical processing (OLAP) needs, supporting data mining and management decision-making (He et al., 2022). A data warehouse based on Oracle and its related components was constructed to address the storage and processing requirements of radio data, as detailed in relevant studies (Hu, 2017). This data warehouse design was well-suited for scenarios characterized by low data volume, structured data storage, and stable business needs. A solution based on Oracle database and data warehouse technologies had been proposed, specifically tailored for the efficient storage and processing of radio monitoring data (Imran et al., 2021). This solution integrates functionalities, such as clustering analysis, unknown signal prediction, and pattern mining, utilizing intelligent data mining techniques to improve the comprehensive coverage of signals. Furthermore, by incorporating OLAP analysis within a Browser/Server model, the system enables more intuitive and visual data presentation, offering robust support for decision-making. However, with the increasing variety and volume of monitoring data, relying on relational databases, such as Oracle, can result in excessive database load. Additionally, the table structure design may become suboptimal, posing challenges in effectively supporting diverse data retrieval and processing demands.

Big data storage technologies excel at managing data with high timeliness, fast storage speeds, large capacities, and varying quality. One study designed a distributed storage system architecture for massive radio monitoring data, leveraging the Hadoop cloud computing platform and the HBase distributed database to address the storage, retrieval, and analysis requirements of radio regulatory agencies (Inmon, 2006). Another approach employed a combination of Redis and MongoDB, two non-relational databases, to achieve distributed storage for large-scale spectrum data (Lu, 2019). An efficient data storage solution leveraging the distributed characteristics of HBase has been proposed, enabling dynamic design of storage table structures based on query requirements and offering the flexibility to meet diverse storage needs (Pan et al., 2023). Another approach combines MongoDB's replica sets and sharding technology to create a distributed storage architecture that efficiently stores spectrum data across multiple storage server nodes (Tian et al., 2017). Additionally, an integration of HBase with Elasticsearch was introduced to address the challenges of storing and querying large volumes of electromagnetic environment data. This solution provided a robust and efficient framework for managing complex data requirements (R. Wang, 2023). Moreover, the application of cognitive radio technology and spectrum sensing algorithms, such as GSRED, was proposed to enhance the spectrum management process in smart grids, offering significant improvements in real-time data processing and decision-making in dynamic environments (Y. Wang, 2020). A comprehensive framework for an electromagnetic environment data warehouse was proposed, addressing the challenge of managing large-scale electromagnetic data through a multidimensional model (Q. Yang et al., 2020). One study explored the use of open-source NoSQL columnar databases to build data warehouse solutions in such environments, further supporting the need for flexible, scalable storage technologies (X. Yang, 2023).

However, these storage architectures are primarily focused on storing business data and do not establish data themes based on specific business needs, limiting their suitability for analytical data processing and decision-making support. To address this challenge, a fast storage solution based on a rotation model was proposed, aimed at significantly enhancing the storage throughput of electromagnetic big data streams. This solution effectively alleviated the pressure on storage systems caused by high-speed data flows, providing robust support for the real-time processing of electromagnetic environment data (W. Zhang, 2020). In addition, a Hadoop-based CTK clustering algorithm was developed for the analysis of radio monitoring spectrum data (Y. T. Zhang, 2024). This algorithm efficiently processed massive datasets within a distributed environment, addressing the lack of automated analysis in radio monitoring and offering new insights for accurate spectrum data analysis and real-time decision-making, especially in big data environments.

This research aims to address the challenge of managing and analyzing large-scale electromagnetic environment data by proposing a comprehensive framework for an electromagnetic environment data warehouse. The significance of this study lies in its ability to bridge the gap between traditional data management techniques and the growing demands of big data in the electromagnetic domain. Specifically, this paper first models the themes of the electromagnetic environment data warehouse, designing the subject areas and dimensions of the electromagnetic environment. On this foundation, a multidimensional model for electromagnetic environment big data is presented (see data model of the electromagnetic environment section). Then, a big data-based architecture for the electromagnetic environment data warehouse is proposed, utilizing big data storage and data warehouse technologies to store electromagnetic environment data. The architecture introduces data sources, and the processes of extraction, cleansing, and standardization. The storage structure of the electromagnetic environment data warehouse is designed using both relational and non-relational databases (see data warehouse architecture section). Finally, a multidimensional analysis of electromagnetic signal data is conducted to demonstrate the effectiveness of the electromagnetic environment data warehouse (see multidimensional analysis of electromagnetic environmental data section).

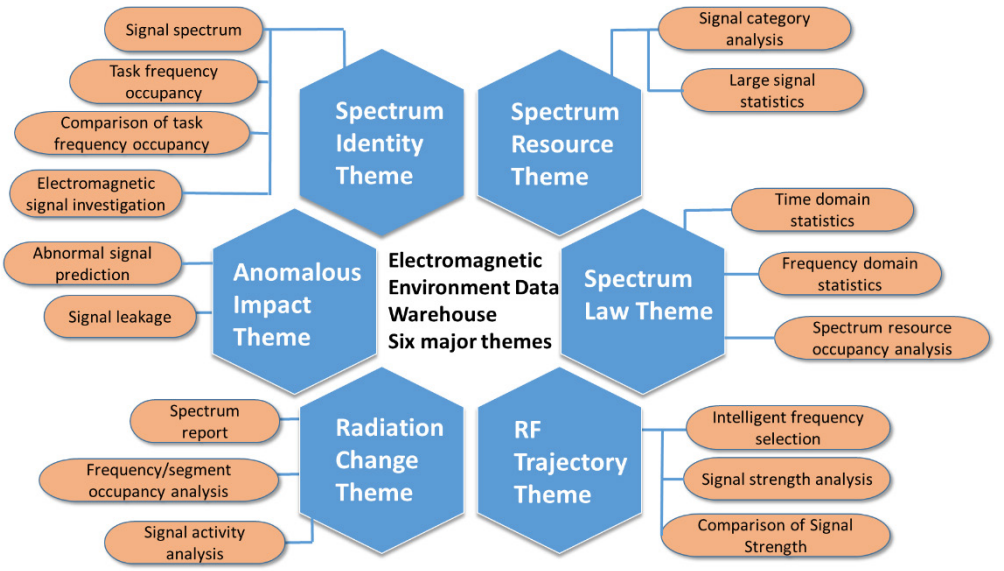
DATA MODEL OF THE ELECTROMAGNETIC ENVIRONMENT

There are various data warehouse modeling methods, such as the ER (Entity-Relationship) model, dimensional model, data vault model, anchor model, etc., among which the dimensional model is a commonly used modeling method (Zheng, 2017). Dimensional modeling involves extensive preprocessing for each dimension, which can greatly enhance the processing capabilities of the data warehouse, providing a significant performance advantage. Additionally, dimensional modeling is very intuitive, closely centered around the business model, and can directly reflect business problems without requiring particularly abstract processes to complete the modeling. Therefore, the dimensional modeling method is used here for modeling the electromagnetic environment data warehouse.

Theme Design

Based on the comprehensive business requirements of the electromagnetic environment, six major themes have been designed for the electromagnetic environment data warehouse: spectrum identity, spectrum resources, signal patterns, radio frequency (RF) trajectory, radiation variation, and anomalous impact. Each major theme includes several sub-themes to facilitate related data analysis, as shown in Figure 1.

Figure 1. Thematic design of the electromagnetic environment data warehouse



Note. RF = radio frequency.

Spectrum Identity Theme

The spectrum identity theme aims to aggregate and integrate environmental monitoring data with radiation source direction-finding data. Comprehensive analysis of electromagnetic signal characteristics within a specific frequency band functionality includes statistical reporting and generation of task-specific frequency occupancy reports, comparative and trend analysis of signal frequency occupancy, and signal mapping to present characteristic information of specific devices or signals, including field strength variation and coverage. Additionally, the spectrum identity theme encompasses baseline analysis of signal detection and field strength variation for designated devices within specific timeframes and frequency bands, offering a clear representation of signal occurrence frequency and field strength fluctuation trends.

Spectrum Sources Theme

The spectrum resource theme aims to comprehensively describe and analyze the spectrum resources and related information within the electromagnetic environment. This theme encompasses a multidimensional analysis of data on frequency usage by stations, spectrum parameters of frequency-using equipment, frequency resource data, and spectrum sensing data. It enables detailed statistical analysis of signal categories, including information on signal types, occurrence frequency, and relationships with equipment, as well as signal frequency and direction. Additionally, the spectrum resource theme focuses on the statistical analysis of major signals within specific timeframes and frequency bands, recording basic signal information, occurrence frequency, and their relationships with equipment.

Signal Pattern Theme

The signal pattern theme aims to reveal the temporal and spatial distribution characteristics and variation patterns of electromagnetic signals through multidimensional analysis in the time domain, frequency domain, and spatial domain. This theme focuses on conducting correlation analysis and mining of signal data from the perspectives of time, frequency, and space, assisting users in gaining a comprehensive understanding of signal behavior patterns.

In terms of time domain statistics, this theme integrates data from the same frequency band across different tasks and performs statistical analysis on signal occupancy, field strength variation, and other metrics at time granularities, such as hours, days, and months, generating time domain reports that reveal the variation patterns of signals within specific time periods. Frequency domain statistics focus on aggregating multitask data within the same frequency band, analyzing changes in signal frequency distribution, field strength, and occupancy, and generating signal occupancy analysis reports. Spatial analysis can be further expanded by combining data from different geographical locations to illustrate the coverage and intensity variation trends of signals in different areas.

RF Trajectory Theme

The RF trajectory theme aims to reveal the spatial movement trajectory, intensity change, and spectrum utilization of signals through a comprehensive analysis of RF signals received by electromagnetic environment sensing devices. This theme focuses on describing the dynamic changes of signals in the spatial and temporal dimensions, which provides important support for target behavior and attribute adjudication. In the RF trajectory theme, core data, such as signal generation time, movement path, frequency characteristics, and influence range, are integrated and analyzed. Through the intelligent frequency selection function, the results of the band scanning task can be utilized to identify the free frequency bands within a specific time range and the occupation of signals, providing a reliable basis for spectrum management and resource allocation. Signal strength analysis generates a detailed signal strength report by processing the monitoring data, showing the change of signal strength in a specific frequency band and time. Signal strength comparison supports accurate comparison of field strength changes of different signals through cross-task and cross-device data summarization.

Radiation Variation Theme

The radiation variation theme is dedicated to an in-depth analysis of the dynamic changes in equipment radiation as revealed by long-term monitoring results. It uncovers the behavioral patterns of signals and the utilization of spectrum resources by evaluating signal activity across different time periods, frequency occupancy, and spectrum usage. This theme encompasses several key areas of analysis, including signal activity analysis, which focuses on signal occupancy during both daytime and nighttime to understand performance variations at different times; frequency occupancy statistics, which identify and analyze usage patterns of frequency points, particularly the occupancy levels of frequently transmitted signals; spectrum report generation, which provides detailed information on the usage of specific frequency bands within designated time frames; and frequency band occupancy statistics, which illustrate the overall occupancy of frequency bands and their variations over different time granularities. These analyses facilitate the effective management and optimization of spectrum resources, ensuring their efficient utilization while also enabling the assessment of trends in changes to the electromagnetic environment.

Anomalous Impact Theme

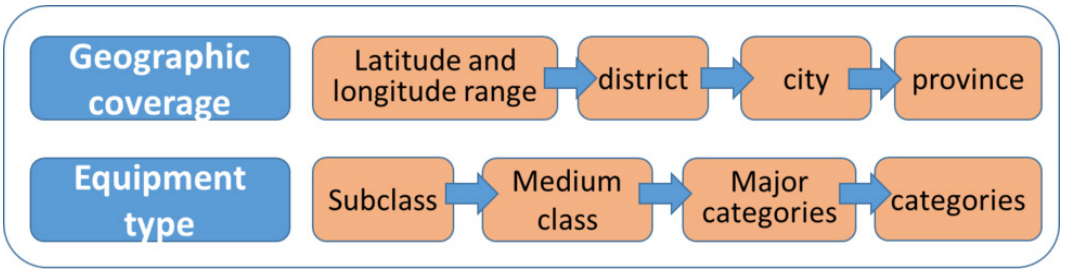
The anomaly impact theme focuses on analyzing and assessing signals that may have a detrimental impact by integrating long-term monitoring results of anomalous signals and signal characteristic data to identify and predict potential threats. This theme encompasses a comprehensive analysis of anomalous electromagnetic environment data, anomalous signal data, and threat alert analysis results. The primary focus is on the prediction of anomalous signals, which involves analyzing data from a past period to forecast the likelihood of unknown, illegal, or illicit signals emerging in the future, along with labeling and alerting these signals. This analysis incorporates the historical field strength and occurrence frequency of signals, as well as signal mapping, to accurately predict future anomalous signals. Another critical aspect is the analysis of signal leakage, which involves comparing electromagnetic environment monitoring data of the same specifications collected at different times. By selecting reference and

comparison subjects and merging and comparing data within specified timeframes and frequency bands, this analysis helps identify missing signals between the reference and comparison subjects, providing a clear comparison of signals and revealing potential signal leakage issues.

Dimensional Design

The design of data analysis encompasses five dimensions: time, device, task, frequency, and signal. The time dimension primarily includes seconds, minutes, hours, days, months, and years. The device dimension mainly includes name, model, latitude and longitude, coverage area, and the upper and lower frequency limits monitored by the device. The task dimension mainly includes associated devices, task start time, task end time, and task parameters. The signal dimension primarily encompasses various aspects, including signal name, signal frequency, signal bandwidth, signal field strength, modulation method, signal type, site unit, licensed station, and assigned station. By synthesizing these dimensions, one can establish granularities for time, geographical range, frequency, and type. Smaller data granularity leads to a more detailed description of the scenario (Hu, 2017). Time granularity ranges from seconds to years; geographical range granularity varies from province, city, district to specific location; frequency granularity includes frequency points, channels, and frequency bands. The type granularity is divided into device type and signal type, with both types encompassing categories, such as major categories, subcategories, and detailed classifications. Figure 2a illustrates the granularity of the device dimension, while Figure 2b presents the granularity of the signal dimension. The granularity of the device dimension may include geographical range and device type, while the granularity of the signal dimension includes time granularity, geographical range granularity, frequency granularity, and type granularity.

Figure 2. Design of particle granularity



Multidimensional Data Model of the Theme

Definition 1

Multidimensional modeling of big data for the electromagnetic environment: Define a seven-tuple $(T, G, F, Y, H, M, \Delta)$, in which:

1. T represents a collection of times;
2. G denotes the set of geographic locations;
3. F denotes the set of frequencies;
4. Y denotes the set of signal types;
5. $\mathrm{H} = \{H_T, H_E, H_M, H_F, H_Y\}$, and H_T, H_E, H_M, H_F, H_Y represent the conceptual hierarchy of time, device, task, frequency and signal type, respectively;
6. M denotes the set of metrics;

7. $\Delta = \{\delta_1, \delta_2, ..., \delta_m\}$ is a collection of functions such that $\delta_i: T \times E \times M \times F \times Y \rightarrow M$, indicating that, given a specific time, device, task, frequency, and signal type, a unique metric value can be obtained.

Definition 2

Conceptual hierarchy: Let the set of hierarchies $\mathrm{L} = \{L_1, L_2, ..., L_n\}$, such that there exists a mapping function $f_i: L_i \rightarrow L_{i+1}$, such that, for $l_i \in L$, there exists a $l_{i+1} \in L_{i+1}$ that satisfies $l_{i+1} = f_i(l_i)$, and then the concept hierarchy can be expressed as (L, F), and $\mathrm{F} = \{f_1, f_2, ..., f_{n-1}\}$.

According to definition 2, in definition 1, H_T, H_G, H_F, H_Y can be expressed, respectively, as:

$$H_T = (\mathrm{L}_T, \mathrm{F}_T)\mathrm{L}_T = \{\text{year, quarter, month, day}\}(math.)\mathrm{F}_T = \{f_{day \rightarrow month}, f_{month \rightarrow quarter}, f_{quarter \rightarrow year}\},$$

$$H_G = (\mathrm{L}_G, \mathrm{F}_G)\mathrm{L}_G = \{\text{province, city, district, location}\}(math.)\mathrm{F}_G = \{f_{location \rightarrow district}, f_{disctict \rightarrow city}, f_{city \rightarrow provinve}\},$$

$$H_F = (\mathrm{L}_F, \mathrm{F}_F)\mathrm{L}_F = \{\text{band, frequency}\}(math.)\mathrm{F}_G = \{f_{frequence \rightarrow band}\},$$

$$H_Y = (L_Y, F_Y)\mathrm{L}_Y = \{\text{category, sub}category, type\}(math.)\mathrm{F}_Y = \{f_{type \rightarrow subcategory}, f_{subcategory \rightarrow category}\}.$$

Definition 3

Topic: Given a six-tuple $(T, G, F, Y, H, M, \Delta)$, the theme $S = (\Psi, \Omega)$, where $\Psi \subseteq \{\mathrm{T, G, F, Y}\}$ and $\Omega \subseteq \mathrm{M}$, Ψ is called the dimensions of the theme, and Ω is the metric of the theme.

Definition 4

Spectrum identity theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{Y}\}$. $\Omega = \{\mathrm{RM, RD, RSC}\}$, where RM, RD, RSC denote the indicators of radiation signal monitoring, radiation signal lateralization, and calculation of radiation source identity data, respectively, which mainly include occupancy, maximum/small field strength, number of signal occurrences, signal coverage, field strength trend, and signal change trend, and so on.

Definition 5

Spectrum resource theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{F}\}$. $\Omega = \{\mathrm{FS, DSP, FR, SS}\}$, where FS, DSP, FR, SS denote the metrics calculated with frequency station data, frequency equipment spectrum parameter data, frequency resource data, spectrum perception data, etc., respectively, including the number of signal categories, the number of signal occurrences, the signal frequency, the azimuthal angle, the historical field strength, the signal wavelength, the frequency of large signals, the number of large signals, and so on.

Definition 6

Signal pattern theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{F}\}$. $\Omega = \{\mathrm{SOP}, \mathrm{SOR}, \mathrm{SIV}\}$, where SOP, SOR, SIV, respectively, represent the signal appearance time period data, signal appearance range data, signal strength change data and other calculated metrics, including time domain\frequency domain signal strength, time domain\frequency domain frequency occupancy, the number of times the signal appeared, the spectral bandwidth, the signal interference rate, the airspace coverage rate, and so on.

Definition 7

RF trajectory theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{F}\}$. $\Omega = \{\mathrm{MSF}, \mathrm{IQ}, \mathrm{RFSC}, \mathrm{SML}, \mathrm{SGT}, \mathrm{SIR}\}$, where MSF, IQ, RFSC, SML, SGT, SIR denote the indicators calculated by monitoring signal frequency data, IQ data, RF signal characteristics data, signal movement and positioning data, signal generation time data, and signal influence range data, respectively, including signal frequency and strength, signal movement direction, signal speed, path length, propagation time, path signal strength change, signal strength in the coverage area, and interference level.

Definition 8

Radiation variation theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{Y}\}$. $\Omega = \{\mathrm{RVS}, \mathrm{RVA}\}$, where RVS, RVA denote the metrics calculated for radiation law change statistics, radiation law change analysis data, including: daytime\nighttime occupancy, frequent signal identification, granularity occupancy, and number of signal occurrences, respectively.

Definition 9

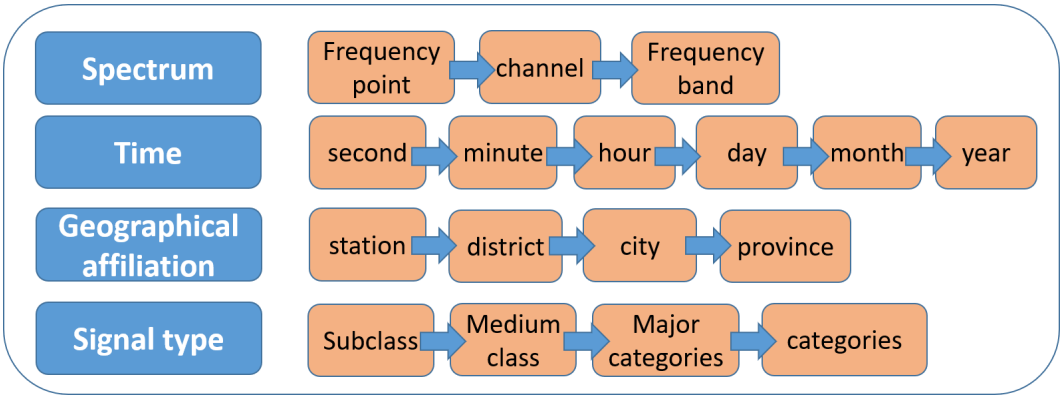
Anomalous impact theme: $\Psi = \{\mathrm{T}, E, M, \mathrm{Y}\}$. Ω=$\{\mathrm{AEM}, \mathrm{AS}, \mathrm{TAA}\}$, where AEM, AS, TAA denote anomalous electromagnetic environment data, anomalous signal data, and threat alarm analysis result data indicators, including prediction time, signal type, prediction probability, historical field strength, alarm identification, leakage signal identification, reference signal characteristics, and comparison signal characteristics, respectively.

ELECTROMAGNETIC ENVIRONMENT DATA WAREHOUSE ARCHITECTURE

Data Warehouse Architecture

This section presents the architecture of the big data-based electromagnetic environment data warehouse, which is organized into four layers. The data sources layer collects various types of electromagnetic environment data; the data collection layer is responsible for data extraction, cleaning, and standardization; the data warehouse layer integrates and organizes data into different levels for further analysis; and the data application layer provides services for data analysis and visualization. Each layer plays a crucial role in ensuring the effective storage, processing, and utilization of the electromagnetic environment data. Figure 3 illustrates the architecture of a big data-based electromagnetic environment data warehouse.

Figure 3. Architecture of electromagnetic environment data warehouse based on big data



Note. OLAP = online analytical processing; OLTP = online transaction processing; RF = radio frequency; hdfs = Hadoop distributed file system.

Data Source

The data warehouse processes various types of electromagnetic environment data, such as spectrum monitoring data, frequency usage station data, and frequency usage equipment data. Based on the data acquisition time and method, these data sources can be categorized into batch historical data and real-time data.

Data Collection Layer

Bulk historical data needs to be loaded into the data warehouse through extraction, cleaning, and standardization, and real-time data is loaded into the data warehouse by data sources, such as sensing devices and business libraries, in the form of message subscription and data standardization processing.

Data Warehouse Layer

Massive and multi-source data must be integrated and computed to be effectively utilized for mining, thereby unleashing the potential value of the data to empower electromagnetic environment application requirements. In the face of the multi-source heterogeneity and computational complexity of electromagnetic environment data, the existing layered architecture of data warehouses is followed. The electromagnetic environment data warehouse is designed to include an operational data store (ODS), a data warehouse detail, a data warehouse summary, and a data warehouse application. The transformation of electromagnetic environment data assets into electromagnetic environment information assets is achieved through the processing between different layers of the data warehouse. The data processed in the ODS is essentially the same as that handled in the data acquisition layer, with the aim of preserving the original data and decoupling data sources. The detail data layer stores the standardized and processed electromagnetic environment data at the smallest granularity, providing unified, standardized, and clean data for subsequent processing. The summary data layer organizes data by subject, constructing multidimensional model data according to business requirements for data integration, splitting, and summarization within related subject areas. The application data layer constructs multidimensional model data based on business needs, and the resultant data is directly used for analysis and presentation.

Data Application Layer

Based on the well-integrated data from the data warehouse, secondary processing is carried out, providing various data services externally through interfaces, including OLAP, online transaction processing, data mining, and data visualization.

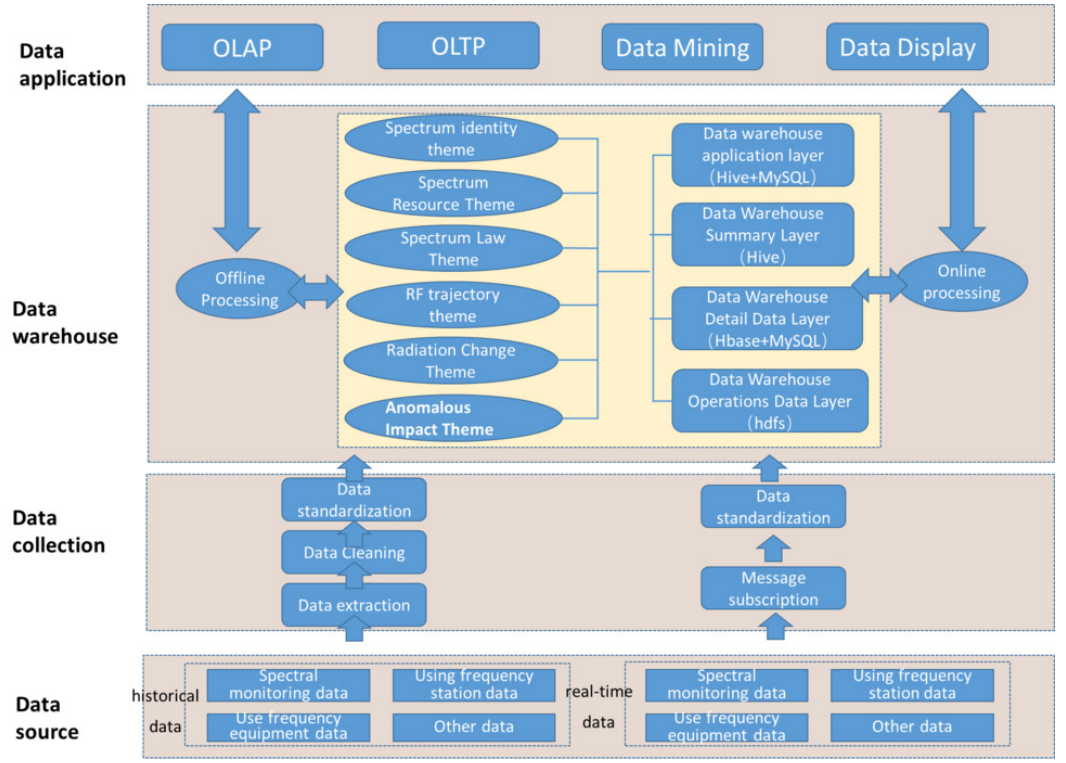
Data Sources

Among the data aggregated in the electromagnetic environment data warehouse, bulk historical data come from existing operational databases and offline electromagnetic environment data files, and real-time data mainly come from electromagnetic environment sensing equipment that performs sensing tasks. Among them, there exist operational databases, such as a radio station database, frequency-using equipment database, frequency-using station database, or monitoring system database; offline electromagnetic environment data files, such as short-wave detection data files, short-wave or ultra-short-wave band monitoring data files, or signal direction measurement data files; electromagnetic environment sensing equipment, such as spectrum monitoring equipment, detection equipment, and direction measurement equipment. Electromagnetic environment data follow different formats or protocol standards, formats, such as XML, CSV, JSON, txt; and protocols, such as real-time messaging protocol specification device protocols, R&S vendor device protocols, Thales device protocols, and so on.

Data Acquisition Process

This section outlines the data collection process for the electromagnetic environment data warehouse, focusing on data extraction, cleaning, and standardization. The process is illustrated in Figure 4, and it includes methods for handling both batch historical data and real-time data. The extraction phase utilizes various tools and techniques depending on the data type, followed by cleaning procedures that ensure data accuracy and consistency. Finally, the standardized data is processed and loaded into the data warehouse for further use. The data collection process is illustrated in Figure 4, where the main tasks involve data extraction, cleaning, and standardization.

Figure 4. Data acquisition process



Note. DB = database; XML = extensible markup language; CSV = comma-separated values; JPG = joint photographic group; RMTP = real-time messaging protocol; R&S = Rohde & Schwarz; ftp = file transfer protocol; MQ = message queuing; HDFS = Hadoop distributed file system.

First, data is extracted from the data sources. Given the diversity of sources for the electromagnetic environment data warehouse, different methods are employed for data extraction based on the data type. For batch historical data, extraction tools, such as Sqoop, Kettle, and FTP, are utilized. When loading data into the warehouse, incremental extraction methods are used for data sourced from existing operational databases, while full extraction is applied to offline electromagnetic environment data files. For real-time data, message subscription methods are employed, utilizing platforms, such as Rabbit message queuing (MQ), Redis, and Kafka. Simultaneously, during the data extraction process, specified data collection dates, geographic information, and frequency domain information are appended to the extracted data to facilitate unified management. Next, the extracted data undergoes a cleaning process. The data loaded into the ODS is cleaned according to established rules, covering three types of checks: data format, data consistency, and the rationality of business logic. Data format checks primarily address errors, missing data, out-of-range values, or illegal data formats. Data consistency focuses on resolving primary and foreign key reference conflicts and handling data redundancy. The rationality of business logic ensures that the data complies with relevant rules pertaining to the business context. Finally, the cleaned data undergoes a standardization process. This involves converting data of various formats and protocols into a unified standard, which includes standardized naming, formatting, value assignments, and the removal of duplicate data. Ultimately, the standardized data is loaded into the data warehouse.

DESIGN OF ELECTROMAGNETIC ENVIRONMENT DATA WAREHOUSE STORAGE BASED ON BIG DATA

Relational database has several advantages, such as the relational model being easy to understand, user-friendly, easy to maintain, and having a low probability of data redundancy and inconsistency. However, due to the diverse types of electromagnetic environment data and the large volume of data, using relational databases can lead to excessive database load. In contrast, non-relational databases, like the Hadoop distributed file system and HBase, offer advantages, such as diverse data storage formats, ease of scalability, and high-performance concurrent read and write capabilities, making them suitable for the efficient storage of long-term, multi-granularity, and diverse types of data in electromagnetic environment data warehouses. Therefore, combining the strengths of both database types, the storage design for the electromagnetic environment data warehouse is as follows.

Operation data layer: Since the raw sensing data are mainly binary files, combined with the electromagnetic environment sensing task execution cycle, the file size is generally more than 100 MB, so the Hadoop distributed file system is used to store the raw data extracted from the data source.

Detailed data layer: The standardized data stored in this layer contains structured, semi-structured and unstructured data, MySQL is used to store structured data, and HBase is used to store semi-structured and unstructured data; MySQL stores structured electromagnetic environment data, such as frequency station data and frequency division data; HBase stores semi-structured and unstructured data at second-level granularity; semi-structured HBase stores semi-structured and unstructured data at second-level granularity, semi-structured data, such as spectral field strength data, signal measurement parameter data, direction finding and localization data, and IQ data, and unstructured data, such as raw sensory data, sound, binary files, and spectral snapshot images.

Aggregate data layer: According to business requirements, this layer can be manually/timed to detail data layer data processing into a variety of data granularity of data, and these data in accordance with business needs use Hive database for storage.

Application data layer: MySQL is primarily utilized to store statistical summary result data related to the electromagnetic environment, including station information, signal recognition results, automated analysis results, target recognition results, and threat alert analysis results. In contrast, Hive is used to store vast amounts of supporting data, including electromagnetic background noise, signal parameter information, target parameter information, and threat alert information.

Construction of the Thematic Model for the Electromagnetic Environment Data Warehouse

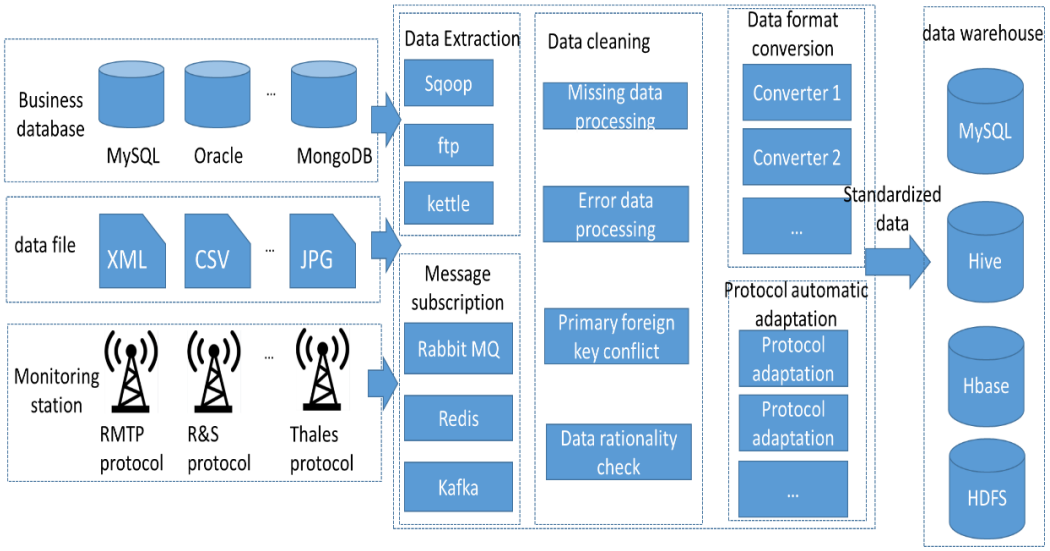
Each theme is associated with a fact table and dimension tables. The dimension table stores the associated attributes of the objects in the fact table, while the fact table mainly stores electromagnetic environment data. The relationships between the dimension tables and fact tables can be categorized as “star,” “snowflake,” or “hybrid” models. For the dimension table data, dimension changes are captured based on the primary key. If changes occur, a slowly changing dimension process is applied to extract the changed data into the data warehouse; if there are no changes, the process ends. For the fact table data, field changes are captured based on the primary key. If there are changes to the fields, the corresponding changed data is extracted into the data warehouse. If there are no changes, the newly added data is extracted into the data warehouse.

When designing dimension tables and fact tables, the index design for cross-database access is conducted in conjunction with business requirements. The dimension tables at the detail data layer are designed based on time domain, frequency domain, and spatial domain information. The fact tables are designed with primary and foreign keys based on the information from the dimension tables. The summary data layer extracts and integrates data according to themes, creating indexes and data storage at different granularities, which can be linked to the detail data layer through indexes in the

tables. The application data layer can associate data from the summary data layer's themes based on tasks, devices, and time.

Taking the spectrum resource occupancy analysis under the theme of signal patterns as an example, the spectrum resource occupancy analysis primarily implements the analysis of the occurrence patterns of electromagnetic signals in the area of interest within a specified time range. Based on this requirement, the data relationships for spectrum resource occupancy analysis are designed using dimension tables and fact tables in a star schema model. Based on data granularity, the designed fact tables include the electromagnetic signal occupancy table at the second level, the electromagnetic signal occupancy table at the minute level, the electromagnetic signal occupancy table at the half-hour level, the electromagnetic signal occupancy table at the daily level, and the electromagnetic signal occupancy table at the monthly level. The dimension tables include attribute fields that describe the data in the fact tables, including time dimension, station dimension, and frequency dimension. The star schema relationship between the fact tables and dimension tables for spectrum resource occupancy analysis is shown in Figure 5.

Figure 5. Star thematic model for spectrum resource occupancy analysis



MULTIDIMENSIONAL ANALYSIS OF ELECTROMAGNETIC ENVIRONMENTAL DATA

Electromagnetic environment data mainly includes time-domain, frequency-domain, and spatial-domain information. Based on sorting and establishing correlations from the data perspective, various themes and models were designed. To better meet business needs, the original data needs to be analyzed from different dimensions to produce mining results. Different multidimensional analysis methods were designed, considering the characteristics of the themes and data attributes.

Multimodal Intelligent Frequency Allocation

By utilizing long-term past perception data and real-time perception data, the frequency usage within the coverage area is surveyed to identify available frequencies that can be allocated. In scenarios where no available frequencies are found, frequencies with minimal interference and weak environmental impact can be chosen. This mainly includes analysis of future available

frequencies, analysis of already used frequencies in the region, and the historical field strength changes of each frequency. Based on the designated equipment, time period, frequency band, occupancy threshold, and field strength threshold, the signal occupancy, signal appearance frequency, and recommended vacant frequency bands are calculated. As shown in Figure 6, between April 2, 2018, and June 4, 2018, the frequency occupancy and signal occurrence times within the 350–470 MHz frequency band are calculated, the space-time list is calculated, and the intermodulation interference diagram is drawn to obtain the available frequency band.

Figure 6. Available frequency analysis interface

Analysis of Electromagnetic Signal Composition

Real-Time Signal Composition Analysis

For real-time tasks, combined with the signal feature database, the signal results and summary analysis of each equipment are carried out, identifying the signal types, quantities, etc. It supports analysis of the signal composition results for all or individual equipment in real-time tasks.

Historical Signal Composition Analysis

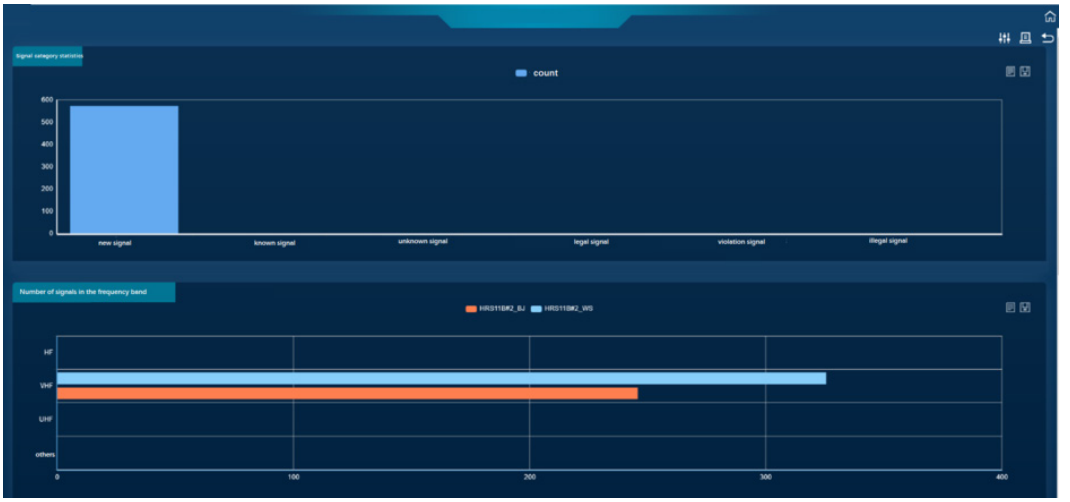
Based on massive historical data, combined with the signal feature database, the signal results and summary analysis of each equipment are conducted. For the monitoring equipment, next is to analyze all tasks within the specified past time period to obtain the historical signal composition results.

Regional Signal Composition Analysis

The main tasks include analyzing signal categories within a specified region, frequency band distribution, time-domain signal distribution, and the correlation between signals and equipment. By using frequency band monitoring data over a certain period, the system analyzes signal categories, associated frequency bands, the number of signal occurrences, the frequency and count of signal appearances, the number of signals within different frequency bands, and the relationship between signals and equipment.

During the queried time period, the number of new signal occurrences was 582, while the occurrences of other types of signals (known signals, unidentified signals, legal signals, illegal signals, etc.) were 0. The number of very high frequency signal occurrences was 245, and the number of distinct very high frequency signals was 336. The signal category analysis interface is shown in Figure 7.

Figure 7. Signal category analysis interface



Analysis of Electromagnetic Signal Composition

Large Signal Statistics

Next is to analyze radiation source signals with large amplitudes and their impact on the surrounding electromagnetic environment. According to the designated time period, frequency band, threshold, and time granularity (minute, half-hour, day, month), calculate the occurrence of signals and the signal appearance within individual time granularity. For example, the start time is set to October 8, 2019, at 13:57:07, and the end time is set to October 15, 2019, at 13:57:07. The frequency range is specified with a start frequency of 80 MHz and an end frequency of 108 MHz, while the threshold value is set to 0. The time granularity is configured to half an hour, and the results are presented in Figure 8.

Figure 8. Large signal detail interface

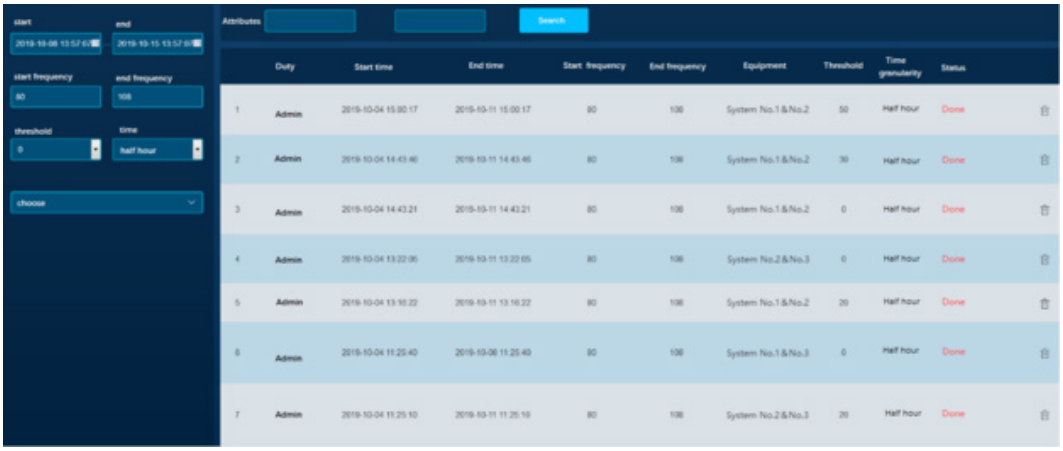


Figure 9 presents the detailed information of significant signals. Within the queried time period, the number of signal occurrences in the frequency bands 94 MHz, 92 MHz, 82.95 MHz, 82.975 MHz, 83 MHz, and 83.025 MHz are 233,936, 233,535, 58,418, 58,470, 58,412, and 58,428, respectively.

Figure 9. Large signal statistical interface

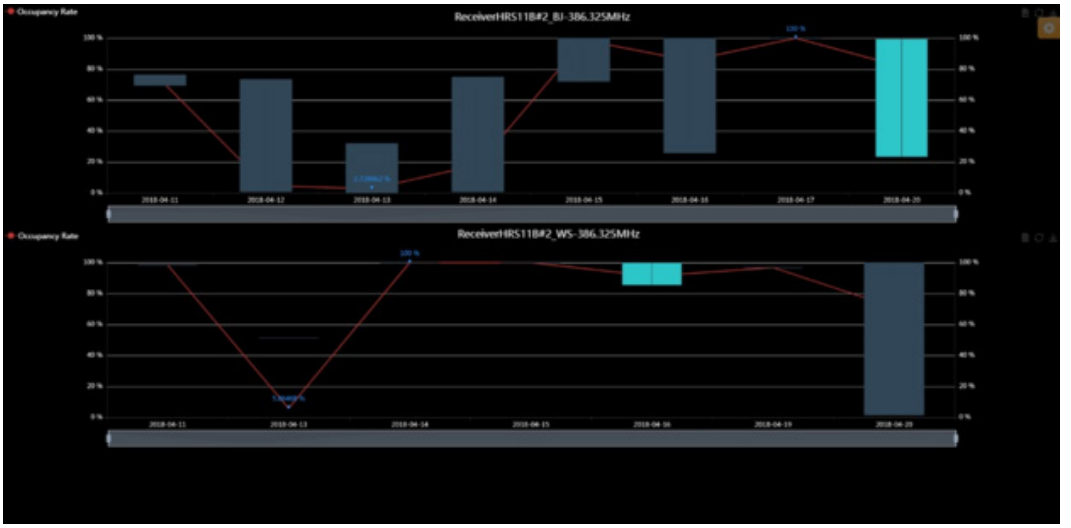
Signal Occupancy Statistics

By analyzing the signal occupancy in the time domain, the effective activity patterns of the signals can be derived. According to the designated equipment, time period, frequency band, threshold, and time granularity (minute, half-hour, day, month), the frequency band and time occupancy of each signal are calculated. By setting a threshold, signals that do not meet the criteria can be filtered or their occupancy reduced.

Figure 10 shows the occupancy changes of receivers HRS11B#2_BJ-386.325MHz and HRS11B#2_WS-386.325MHz. In the upper chart, the receiver HRS11B#2_BJ-386.325MHz

shows that the signal occupancy rate exceeded 70% on April 11, 2018, and then dropped sharply, reaching 2.728962% on April 13, 2018, nearly 0%. Starting from April 14, the signal occupancy rate rapidly increased, reaching nearly 100% on April 15, and remained at 100% on April 17 with minor fluctuations. In the lower chart, the receiver HRS11B#2_WS-386.325MHz shows that the signal occupancy rate was 100% on April 11, 2018, then dropped to 5.86468% on April 13. On April 14 and 15, the occupancy rate remained at 100%, and from April 16 to April 19, it remained relatively stable. On April 20, the occupancy rate decreased slightly, falling below 80%.

Figure 10. Signal occupancy interface

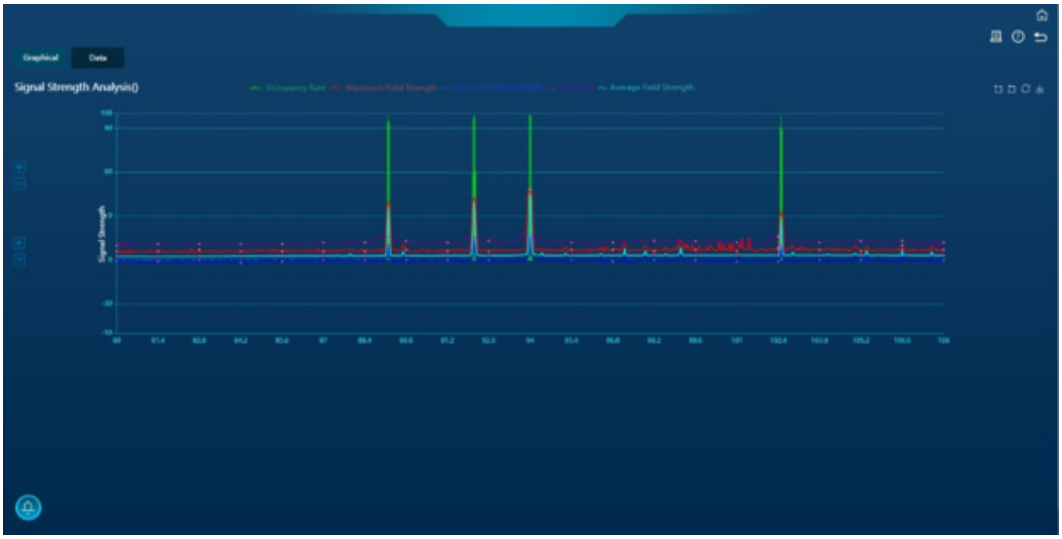


Signal Strength Analysis

Next is to perform various signal strength analyses on multiple monitoring task data at hourly, daily, freely defined daily time intervals, and monthly intervals. Using frequency as the x-axis and occupancy and field strength as the y-axes, the analysis captures the signal strength variations over specific time periods (continuous time, hours, days, freely defined daily periods, and months) within the channel, resulting in statistics, such as the median, maximum, and minimum field strength.

Figure 11 shows the variation of signal strength at different frequencies. The chart displays several distinct spikes at specific frequency locations (such as 89.3 MHz, 92.2 MHz, 94 MHz, 102.4 MHz, etc.). Apart from these spikes, the rest of the curve is relatively flat, representing background noise or other weaker signals. The frequency bands where the spikes occur are occupied by strong signals, while no significant strong signals are present at other frequencies, reflecting an uneven utilization of the frequency bands.

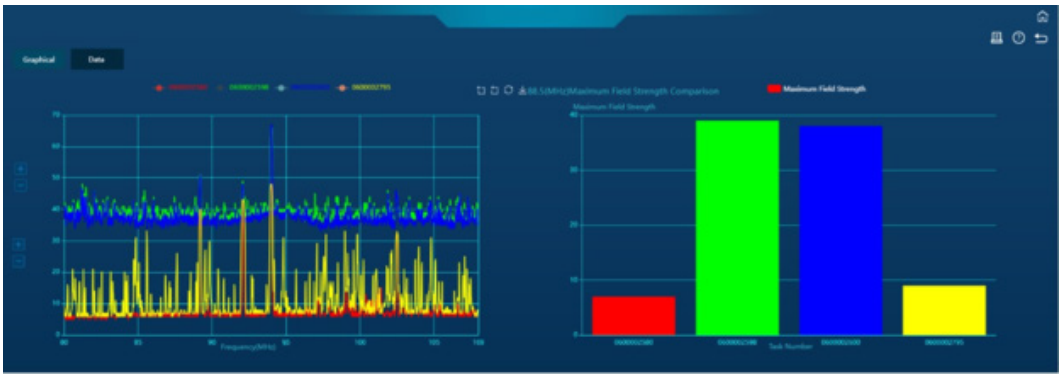
Figure 11. Continuous time graphical display



Signal Strength Comparison

As shown in Figure 12, the frequency occupancy of different monitoring tasks or monitoring stations in the same time period and the frequency occupancy of monitoring tasks and monitoring stations across time periods are compared. The left side shows the signal strength spectrum and the right side displays the maximum field strength comparison, used to analyze the signal strength of different task numbers within specific frequency ranges. In the signal strength spectrum, the red signal (task number 060002580) has the lowest strength, almost close to the background noise. The yellow signal (task number 060002795) exhibits the largest variation and has peaks in multiple frequency bands, especially around 94 MHz, where the signal strength reaches its maximum. The blue signal (task number 060002600) and the green signal (060002598) are generally more stable, with signal strength fluctuating around 40. In the maximum field strength comparison chart, the maximum field strength results are consistent with the trends observed in the left spectrum. The green and blue signals are relatively stable and have higher strength, while the yellow and red signals have weaker field strengths.

Figure 12. Data interface for signal strength comparison



CONCLUSION

This paper addresses the challenges of managing and analyzing large-scale electromagnetic environment data by proposing a comprehensive framework for an electromagnetic environment data warehouse. A thematic structure is constructed by designing the subject areas and dimensions of the electromagnetic environment, providing a theoretical foundation for multidimensional data analysis. A data warehouse architecture that integrates relational and non-relational databases is proposed. The architecture systematically incorporates data sources, extraction, cleansing, and standardization processes, ensuring fast data storage and efficient processing capabilities. Leveraging big data storage technologies, a flexible and efficient storage structure is designed to address the challenges posed by the large volume and diverse types of electromagnetic environment data, overcoming the limitations of traditional databases under high-load and heterogeneous data scenarios. By utilizing the OLAP capabilities of the data warehouse, a multidimensional analysis of electromagnetic signal data is conducted, demonstrating the practical effectiveness of the proposed framework in supporting large-scale data storage and analysis.

In future work, we aim to explore how artificial intelligence (AI) technologies can further enhance the data preparation and management process within the proposed framework. This includes investigating automated methods for data acquisition and ingestion from heterogeneous sources, where AI can optimize extraction and filtering processes. We will also examine the use of AI techniques for intelligent data cleansing, anomaly detection, and standardization to ensure data quality and consistency. Furthermore, we aim to study AI-driven approaches to streamline schema mapping and the integration of diverse data structures, enabling more effective thematic model construction. Lastly, leveraging AI for advanced analytics, including predictive modeling, pattern recognition, and knowledge discovery, will be a key focus to maximize the value of the electromagnetic environment data warehouse and meet the growing demands of big data in this domain.

COMPETING INTERESTS

The authors of this publication declares that there are no competing interests.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the authors of the article.

CORRESPONDING AUTHOR

Correspondence should be addressed to Zhenxing Li; lizx@crirp.ac.cn

PROCESSING DATES

This manuscript was initially received for consideration for the journal on 10/23/2024, revisions were received for the manuscript following the double-anonymized peer review on 03/30/2025, the manuscript was formally accepted on 12/12/2024, and the manuscript was finalized for publication on 03/31/2025

REFERENCES

Chen, J., & Zhao, G. (2020). Application of Hadoop-based CTK clustering algorithm in radio monitoring spectrum data analysis. *China Radio*, (0*8)*, 31-33+41. Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_Ge7TeR3LZAk1MMrOmiSWbEkU4lVDdwyCKqWgabxs3gc5Lv_Cn8MElQ0Rn8s3x7DtUPsrJJ86EVebUWgDDR-6e1w63Ru1VU4MNmCMW2xcsjkmlHI8oHKkedQoltAmrH7KSbrVhigQSR2N2pdIBsGBbWYfmEqvjYIXpV_FMfqySQh0PKiQfo2RGoCgmkavC25cgU=&uniplatform=NZKPT&language=CHS

Cheng, J., Xu, Z., Jia, X., Zhou, C., & Yu, D. (2019). Methods and systems for monitoring the space electromagnetic environment based on shared databases. *Journal of Terahertz Science and Electronic Information Technology*, *17*(3), 435–439. DOI: 10.11805/TKYDA201903.0435

Ding, F. (2015). Research on the construction mode of provincial radio monitoring cloud platform. *China Radio*, (*12)*, 25–27+33. Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=qKY3Y0De9e7THrTW7dRMUP9mEGEeHk0N5uM1lOUlT-6Q9f1AADbDHJMIXHlx5b98JW9v1tngVZgS3VcSeGM3cX8SEuUidK635mGmQ066lLhMhycZITNSHSfsSpcwJg7S0ZpNYQpzQUiH67zIZrwWY_aT8l0vc9vp5wxXrQsijq2Y0n0qohKYVRDdOF-SecDu&uniplatform=NZKPT&language=CHS

Group, A. (2017). *The road to big data: Alibaba's big data practice*. Publishing House of Electronics Industry.

Ha, F., & Jia, N. (2015). Study on a distributed storage system for massive radio monitoring data. *Journal of Chengdu Technology University*, *18*(2), 25–28. DOI: 10.13542/j.cnki.51-1747/tn.2015.02.008

He, H., Li, G.-Y., Kuang, S.-Y., Jiang, G., & Wang, G.-H. (2022). Design of electromagnetic reconnaissance big data storage based on HBase. *Aerospace Electronic Warfare, 38*(3), 22–26+34. DOI: 10.3969/j.issn.1673-2421.2022.03.006

Hu, Q. (2017). *Research on transmission, storage and scheduling technology of mass spectrum monitoring data* [Master's thesis]. (Master's thesis). Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GeZov2wCYDe1IROb3VmoK8fbx2h_NPyzP7OmcbdBuRAA7CizLPr2X3iM76LN7AZgzuEHDuApTwzFMDZYxV_RYQjpkonrcMWGOUo5-CiejKVP36sZORTnjwpQmOb6r8uDH8hHv733WrZTp0AjVFFCxOwjHkqByJvTu-6lyL-0asgJfTM0KV6H1NuisplW1ZxVB0=&uniplatform=NZKPT&language=CHS

Imran, S., Mahmood, T., Khan, A., Qamar, A. M., Siddiqui, A. J., Ahmed, I., & Rehman, N. S. (2021). NODW framework for data warehousing-a NoSQL big data perspective. TechRxiv. *January 16, 2024.* DOI: 10.22541/au.170537198.88138048/v1

Inmon, W. H. (2005). *Building the data warehouse* (4th ed.). Wiley.

Lü, M. (2019). *Research on fast storage and simulation technology for electromagnetic big data* [Master's thesis]. Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_Gdgo-m0N5F5THIKz87weevjDANyoby4spiP-nUKLHfOG_Uv5mF3gu6pJR9VahWkfDcIvyY8TucgxPilnJNA8HxZLkcWUi3xJD-L9RGcjBlDN4pXEFA4CZJ7zaLJ5bbaSZGRLabBXIRVE7MEpc6kxjtfS-e0Pdl6mEQuNGHc2QgpcLyk5a5r1nHK_SSKPqiDACk=&uniplatform=NZKPT&language=CHS

Pan, C. S., Cai, R., Shi, H. F., Shi, J. F., & Wang, Y. Y. (2023). Spectrum intelligent sensing method based on cooperative learning. *Telecommunication Technology*, *63*(12), 1839–1846. DOI: 10.20079/j.issn.1001-893x.220721005

Tian, B., Zhu, Y.-L., Zhang, Y.-C., Hu, J.-T., & Zhang, C.-B. (2017). Research on construction and application of radio monitoring data warehouse. *Computer Science*, *44*(6A). https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GdMC0VhABzLJ8MydQmyjPKnlQ1qccmhe-S8YoDapYp5f-3sU-5sE05Mpm2fUY0YOR0uy3JWW9_fYWHqEIWXz1Wf-0l2SCBIayc5ZYDg1LcY-Txnl4njTyRxdGXd8R-PJoL4k7LqVUkLa_DvJV835BHHq50vML0-_cgJ3dCRyjLwMDq3AhtFgIkDtYdUBYM53-s=&uniplatform=NZKPT&language=CHS

Wang, R. (2023). Design of massive spectrum monitoring data storage based on MongoDB. *Software*, *44*(11), 67–70. DOI: 10.3969/j.issn.1003-6970.2023.11.016

Wang, Y. (2020). Discussion on the development of electronic countermeasure technology in complex electromagnetic environment. *Telecom Power Technology*, *37*(4), 265–266. DOI: 10.1016/j.powtec.2019.12.043

Yang, Q., Zhong, X.-C., & Liang, J.-Y. (2020). Discussion on radio monitoring data about the North Sea of Guangxi. *China Radio, 3*, 58–60. Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GfBEjjREQuyQHEEdkvJnuAwRUSIChgsjTXYSS7F0T30LLObAXrtn3befNz49Oif-7VPodwBdbED5iO1s2emTy9Qyy0QgBDjMI4biY-OML6z-e-qz7dln7zRJsMcF5rmwVYciRelI2sPWZJJ63W_CWcHmFWJOwF2mKQ7-wyCwB28XmPMrpUSlVR9esg61AuF2vE=&uniplatform=NZKPT&language=CHS

Yang, X. (2023). *Research on the storage of large-scale satellite data and data service system development of Zhangheng-1 satellite* [Master's thesis]. Retrieved from https://link.cnki.net/doi/10.27899/d.cnki.gfzkj.2023.000006

Zhang, W. (2020). Design and application analysis of a radio monitoring data warehouse. *Computer Knowledge and Technology*, *16*(7), 10–11. DOI: 10.14004/j.cnki.ckt.2020.0740

Zhang, Y. T. (2024). A spectrum sensing-based demand response management method for smart grids. *Power Big Data*, *27*(3), 1–8. DOI: 10.19317/j.cnki.1008-083x.2024.03.001

Zheng, Z. (2017). *The design and implementation of analysis system of radio monitoring and management based on data warehouse* [Master's thesis].Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GfarKY29xIF0BGa3uJcW5FoX3UkO2SnCmMTvlRiiqxuS4ZhhXKWwE-G_ZRuy7rHbiN1kmDBazBkt8p1OcdIoj4TAHy801-af-9qDlrPNcVf5gJUxTu6ERP-dhBM9ZBC378PpkjXvIAjP6iL7x5nzLjKFYSrMUXGgxPhJEkYo0oWmFxO9ybnenZSTrLj2GU8sEM=&uniplatform=NZKPT&language=CHS