Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 10, No. 1, 2026
www.scholink.org/ojs/index.php/asir

Original Paper

A Survey of Heterogeneous Data Aggregation, Integration and Sharing in Electric Power

Xinhong Zhang[1*], Jun Zhao[1], Yong He[1], Gui Peng[1] & Yue Xiang[1]

[1*] Emergency Management Institute of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region 83000, China

Received: December 18, 2025 Accepted: January 17, 2026 Online Published: January 20, 2026

doi:10.22158/asir.v10n1p30 URL: http://dx.doi.org/10.22158/asir.v10n1p30

Abstract

Heterogeneous data fusion technology for electric power big data is a crucial support for the construction and development of smart grids. The rapid development of smart grids has resulted in a significant increase in the amount of data within power systems, which is characterized by being multi-source, heterogeneous, and large-scale. This data includes real-time operational information, equipment status monitoring data, and market transaction records from various devices and systems. It is essential for enhancing grid operational efficiency, ensuring the security and stability of the grid, and optimizing resource allocation. By integrating multi-source heterogeneous data, electric power big data fusion technology enables data integration, sharing, and comprehensive analysis. This provides core technical support for the intelligent management of power systems. Currently, this technology has made significant progress in data cleaning, transformation, and integration. It is widely applied in areas such as grid dispatching, load forecasting, and equipment status monitoring.

Keywords

Big data of power, Heterogeneous fusion, Smart grid, Data cleaning, Data integration, Data sharing

1. Introduction

With the accelerated deployment of smart grids and the maturation of big-data technologies, power-system big data has become a key driver of digitalization in the electricity sector. It plays a crucial role in the sector's intelligent transformation. Modern grids generate massive, high-velocity, and highly diverse data streams by deploying sensors, smart meters, and advanced measurement infrastructures (Chen et al., 2019). These data span the full power-system chain—generation, transmission, substations, distribution, consumption and dispatch that form a representative power big-data ecosystem.

As smart grids evolve, heterogeneous data fusion has emerged as a central research topic in power informatization. Improvements in monitoring devices and communications have led to exponential growth in data volume and structural complexity. Meanwhile, inconsistent standards and strong heterogeneity across business systems have resulted in “information silos” which substantially limit the practical value of power big data (Zhang, Huang, & Bompard, 2018). To address this issue, researchers have explored heterogeneous data-integration and fusion methods to enable effective aggregation of multi-source power data. Conventional fusion techniques have made preliminary progress in areas such as image processing, but major challenges remain. These challenges include data-quality issues, high algorithmic complexity, and insufficient real-time processing capability (Hongjun, Wei, Ou, et al., 2020). As show in Figure 1, power big data exhibits the classic “4V” characteristics—volume, variety, velocity, and value. They encapsulate rich domain knowledge and operational semantics that are specific to power systems. Its applications not only include real-time grid situational awareness, fault detection, and load forecasting, but also support market operation decisions, renewable energy integration, and optimal dispatch (Hossein Akhavan-Hejazi & Hamed Mohsenian-Rad, 2018). Consequently, the industry faces pressing problems in efficiently acquiring, storing, processing, and sharing heterogeneous power data, as well as systematically extracting actionable value.

This paper presents a structured review of heterogeneous fusion technologies for power big data. It examines current technical bottlenecks and challenges, and explores multi-source integration strategies focused on efficient fusion methodologies (Alma’aitah, Wafa’ Za’al, et al., 2024). This study compares representative approaches and summarizes their advantages and limitations. It aims to provide theoretical insights and practical guidance for integrating complex power-system data and improving grid intelligence, resource allocation, and supply reliability.

2. Overview of Power Big Data

2.1 Data Sources and Characteristics

Power big data primarily originates from the large-scale deployment of sensor networks, smart metering devices, and advanced metering infrastructure (AMI). It also comes from various power information management systems in smart grids. Together, these sources form a multi-dimensional and multi-layer data acquisition framework, including SCADA, EMS, and WAMS systems. The collected data comprehensively cover key stages of the power system, such as generation, transmission, transformation, distribution, consumption and dispatching (I Made Putrama & Péter Martinek, 2024).

Power big data exhibits typical “4V” characteristics: massive volume, high variety, high velocity, and relatively low value density. These features result in strong heterogeneity, which is reflected in the diversity of data sources, complex data structures, and strict real-time processing requirements. Data from different sources often differ significantly in format, accuracy, and sampling frequency. This includes real-time measurements, historical records, textual data, and multimedia data.

To ensure reliable integration and utilization, data standardization, cleansing, and preprocessing are essential steps in power big data management. Due to the stringent real-time requirements of power system operations, data platforms must quickly respond to grid dynamics and provide accurate information to support decision-making. Challenges related to data quality, security, and privacy protection remain significant concerns in multi-source data integration. This underscores the importance of developing efficient and reliable fusion technologies.

2.2 Processing and Analysis Technologies

In smart grids, large-scale sensing devices continuously collect operational data and transmit them to centralized data warehouses for further analysis and decision support.

For data storage, distributed architectures are widely adopted to accommodate the rapid growth and massive scale of power big data. A typical big-data platform architecture based on Apache open-source technologies is illustrated. This framework relies on the Hadoop Distributed File System (HDFS) for scalable storage. It uses MapReduce as the core distributed computing engine, enabling the efficient management of petabyte- and even zettabyte-level datasets (Kyuseok Shim, 2012).

During data processing, distributed computing models such as MapReduce and Spark significantly improve computational efficiency by decomposing tasks and executing them in parallel.

From an analytical perspective, the core of this architecture is built on data mining and intelligent analysis techniques. These include association rule mining, clustering analysis, and pattern recognition. The techniques have been extensively applied to power load forecasting, fault diagnosis and system condition monitoring that accelerate the development of intelligent grid operations. Furthermore, machine learning models, such as neural networks and support vector machines, have shown strong capabilities in capturing complex nonlinear relationships. They are also effective in extracting deep feature representations from power data.

In summary, the processing and analysis framework of power big data integrates data acquisition, distributed storage, parallel computing, and intelligent analytics. This forms a complete technical pipeline that supports efficient utilization and value extraction in modern power systems (Michał Kunicki, Sebastian Borucki, Dariusz Zmarzły, & Jerzy Frymus, 2020).

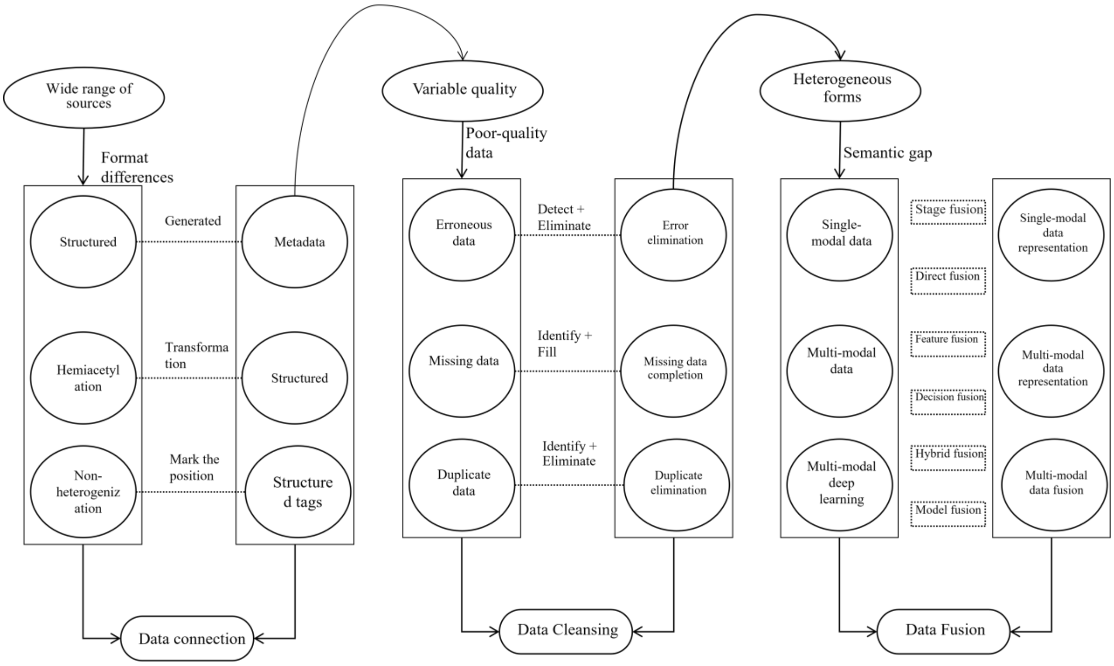

Figure 1. Overall Architecture of the Power Big Data Platform

3. Technical Principles and Methods of Heterogeneous Fusion for Power Big Data

3.1 Data Acquisition and Preprocessing

With the rapid advancement of grid intelligence and informatization, the large-scale deployment of sensors and high-speed data acquisition devices has increased data diversity. At the same time, data volumes are growing rapidly. The power big data has become a prominent trend in modern power systems. In heterogeneous fusion of power big data, data acquisition serves aim to extract key information from multiple heterogeneous sources (Chu, Dong, Chen, Yu, & Huang, 2020). This process involves collecting operational data related to power supply and demand, grid topology, load profiles, and energy storage conditions. At the data access layer, the system must support multiple communication protocols and interface standards. These include WebService-based standardized interfaces, power-specific communication protocols (IEC 61850) and various proprietary file formats. At the data-type level, the platform needs to handle diverse data modalities, such as real-time monitoring data, historical operational records, equipment text logs, inspection multimedia data, and load time-series data. These data include structured data, semi-structured data, and unstructured data.

Real-time operational parameters, such as power, voltage, current, load fluctuations, and energy storage status, are continuously generated from these sources. In addition, equipment condition data play a vital role in health monitoring and fault diagnosis. Indicators such as current, voltage, temperature, and vibration directly reflect equipment operating states and potential abnormal conditions (Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, & Joel Saltz, 2013).

Furthermore, non-IoT data sources, including inspection records and maintenance reports, also need to be integrated. These data are typically uploaded through periodic backups and provide essential support for grid safety management and asset maintenance. Log data generated by enterprise systems, such as OA and ERP platforms, record critical processes related to dispatching, operations, and fault handling. These logs provide valuable insights for safety supervision and incident analysis.

In the era of big data, information acquisition and preprocessing form the foundation of intelligent decision-making. Data cleaning and preprocessing are particularly important, including duplicate removal, missing-value imputation, outlier detection, and data format transformation. These procedures aim to reduce noise, improve data consistency, and standardize the representation for subsequent analysis. This ensures data accuracy and reliability, providing a solid foundation for effective heterogeneous fusion and advanced analytics (Xiao, Wu, Li, Liu, Zhou, Deng, Yang, Hou, Liu, & Mao, 2019).

Table 1. Power Heterogeneous Data Collection

Input Data Type	Description
System Operation Data	Includes energy flow, operational status monitoring data, covering key links such as power generation, grid infrastructure, load demand, and energy storage.
Equipment Condition Data	Includes parameters such as current, voltage, temperature, and vibration characteristics, which are used for equipment health monitoring and fault diagnosis.
Non-IoT API Transmission Data	Includes power inspection data and laboratory experimental data, which are uploaded periodically through scheduled transmission mechanisms.
Log Data	Generated from internal management systems such as OA and ERP, recording business process logs related to dispatching, operation, and fault handling, which are critical for power grid security management and incident investigation.

3.2 Data Aggregation, Storage, and Management

Heterogeneous power big data aggregation is a critical component of informatization in the power industry. Its primary objective is to integrate heterogeneous data from multiple business systems in order for enhancing overall data value and application effectiveness. Power big data originates from a wide range of sources. Core operational data are collected from power production and management systems such as Supervisory Control and Data Acquisition (SCADA), Energy Management Systems (EMS), and Wide-Area Measurement Systems (WAMS), which continuously monitor grid operating conditions, equipment parameters, and load profiles (Zhao & Wang, 2019).

In addition, equipment monitoring data play a vital role in system operation and maintenance. These data include condition information from key assets such as generators, transformers, and transmission lines. This includes measurements like temperature, vibration, and oil levels, which are acquired in real-time through distributed sensor networks and transmitted to centralized data centers. External environmental data, such as meteorological information and geographic information system (GIS) data, also constitute important data sources, providing essential support for load forecasting, fault early warning, and operational planning. During data aggregation, the Extract–Transform–Load (ETL) process is widely adopted to integrate data from heterogeneous sources, resolve format and semantic inconsistencies, and construct a unified data view (Dong Xin Luna, Gabrilovich Evgeniy, Heitz Geremy, Horn Wilko, Murphy Kevin, Sun Shaohua & Zhang Wei, 2014).

Big data storage technologies serve as fundamental infrastructure for the digital transformation of the power industry, enabling reliable storage, efficient management, and high-performance access to massive datasets. In heterogeneous fusion scenarios, the main categories of stored power data include system operation data, equipment condition data, non-IoT API transmission data, and log data generated by enterprise information systems.

To achieve scalable storage and high availability, distributed file systems such as HDFS and distributed storage platforms such as Ceph are commonly deployed. Column-oriented databases, including HBase and Cassandra, are suitable for large-scale structured data storage and analytical workloads, offering improved query performance. Graph databases such as Neo4j are applied to represent and manage complex topological relationships in power networks, including connectivity among grid components. Furthermore, time-series databases such as InfluxDB are specifically designed for storing and managing high-frequency monitoring data generated by grid sensors and measurement devices (Dong Xin Luna, Gabrilovich Evgeniy, Heitz Geremy, Horn Wilko, Murphy Kevin, Sun Shaohua & Zhang Wei, 2014).

From a data management perspective, the power industry emphasizes data standardization and unified interface design. ETL tools such as Talend and Informatica are widely used for data extraction, transformation, cleansing, and preprocessing to ensure data consistency and quality. Enterprise-level data warehouses, including Teradata and Greenplum, are commonly adopted to support integrated data storage and analytical processing. Meanwhile, access control mechanisms and encryption technologies are implemented to enhance data security and privacy protection. In addition, data lake architectures built on Hadoop-based platforms enable the storage of raw-format data and provide flexible support for multi-type heterogeneous data analysis. For data governance, platforms such as Apache Atlas are employed to manage metadata, monitor data quality, and enforce data security and compliance policies.

Table 2. Storage and Management of Power Heterogeneous Data

Data Storage Technology	Description	Data Management Tool	Description
Distributed Storage	Uses distributed file systems such as Hadoop HDFS and Ceph to achieve reliable storage and efficient access for massive-scale data.	Data Warehouse	Tools such as Teradata and Greenplum are used to build enterprise-level data warehouses for data integration, storage, and analytical processing.
Column-Oriented	Technologies such as HBase and Cassandra are suitable for storing	Data Lake	Built on big data platforms such as Hadoop to store raw-format data and

Storage	and analyzing large-scale structured data, improving query performance.		support multi-type data analysis and processing.
Graph Database	Systems such as Neo4j are applied to store and manage complex relational data in power grids, including connectivity among equipment.	ETL Tools	Tools such as Talend and Informatica are used for data extraction, transformation, and loading, enabling efficient data integration.
Time-Series Database	Databases such as InfluxDB are specifically designed to store and manage time-series data, such as power grid monitoring data.	Data Governance Tools	Platforms such as Apache Atlas are used for metadata management, data quality monitoring, and data security control.

3.3 Data Fusion and Sharing

Data integration aggregates information from multiple heterogeneous sources and performs cross-domain correlation analysis, thereby enriching the information content available for power system applications. The objective of heterogeneous data integration is to unify different data formats and standards, enabling a more comprehensive system-level perspective to support intelligent decision-making in power grids. Integration and sharing technologies play a central role in this process by facilitating efficient data aggregation, high-speed processing, and secure data exchange. Through algorithm-driven fusion mechanisms, data from diverse sources can be effectively combined. This approach addresses the challenges of large data volume, high velocity, and multi-type heterogeneity in power systems, while ensuring data accuracy, consistency, and reliability.

3.3.1 Data Fusion Techniques

Kalman Filter-Based Data Fusion: In power systems, the diversity of data acquisition sources and the heterogeneity of data formats lead to highly heterogeneous measurement environments. These data include multi-source, multi-temporal, and multi-precision measurements obtained from different sensors, such as voltage, current, and power factor signals. By integrating such heterogeneous observations, Kalman filtering provides more accurate and robust state estimation results, thereby supporting grid monitoring, control, and operational optimization.

The Kalman filtering algorithm adopts a recursive estimation framework, in which the core computational procedure consists of two key stages: prediction and update. In heterogeneous power big data fusion, these stages are implemented through a series of mathematical operations to dynamically integrate multi-source measurements and minimize estimation errors, enabling optimal state estimation under noisy and uncertain measurement conditions.

(1) Prediction Step: Based on the optimal state estimate at the previous time step and the system noise, the prior state at the next time step is predicted. In power system applications, the state prediction and the corresponding error covariance prediction can be expressed as

$$\hat{x}_{k|k-1} = F_k\hat{x}_{k-1|k-1} + B_k u_k$$

where $\hat{x}_{k|k-1}$ denotes the state estimate at time step k predicted from the estimated state at time step $\hat{x}_{k|k-1}$. F_k is the state transition matrix that describes the evolution of the system state from time step $k-1$ to k; $\hat{x}_{k-1|k-1}$ represents the optimal state estimate at time step $k-1$; B_k is the control input matrix that characterizes the influence of control inputs on the system state; and uk denotes the control input vector at time step k.

(2) Prediction Error Covariance Matrix: The prediction error covariance matrix is used to characterize the uncertainty associated with the predicted system state and can be expressed as:

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k$$

Where $P_{k|k-1}$ denotes the prediction error covariance matrix at time step k, $P_{k-1|k-1}$ represents the estimation error covariance matrix at time step $k-1$, and Q_k is the process noise covariance matrix, which quantifies the uncertainty introduced by system process noise.

(3) Update Step: The update step refines the system state estimate by incorporating new measurement information and the predicted state. In power system applications, this process can be expressed as:

$$K_k = P_{k|k-1}H_k^T\left(H_k P_{k|k-1} H_k^T + R_k\right)^{-1}$$
$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H_k\hat{x}_{k|k-1})$$
$$P_{k|k} = (I - K_k H_k)P_{k|k-1}$$

Where K_k denotes the Kalman gain that balances the contribution of the predicted state and the measurement information; H_k is the observation matrix that maps the system state to the measurement space; zk represents the measurement vector at time step k, $\hat{x}_{k|k}$ is the updated optimal state estimate at time step k; R_k is the measurement noise covariance matrix that characterizes the uncertainty of observation noise; $P_{k|k}$ denotes the updated error covariance matrix at time step k; and I is the identity matrix.

(4) Mathematical Principles and Computational Process: The Kalman filtering algorithm is founded on Bayesian inference and Gaussian distribution assumptions, and recursively estimates the system state by minimizing the variance of prediction errors. During the prediction stage, the algorithm estimates the next system state and its associated uncertainty based on the system dynamic model. During the update stage, newly acquired measurement data are incorporated to correct the predicted state, and the Kalman gain is employed to achieve optimal state estimation. Notably, this recursive framework does not require storing historical data sequences. It only relies on the current state estimate and error covariance matrix. This ensures high computational efficiency and real-time performance, making it well-suited for large-scale power system applications.

Model-Based Data Fusion: Model-based data fusion approaches use standardized data models to achieve interoperability across heterogeneous systems. For example, Wu introduced the Common Information Model (CIM) as a unified data representation framework, enabling heterogeneous system data to be mapped into the CIM structure for standardized integration and fusion. The CIM model defines typical object structures and relational semantics in power systems. By extending the CIM schema to accommodate application-specific requirements, data mapping and fusion across different systems can be effectively realized. For instance, by constructing standardized equipment condition models and performing semantic mapping and structural alignment with CIM, multi-source equipment status data can be efficiently integrated. Model-based fusion techniques effectively address heterogeneity challenges, improve data standardization and interoperability, and help eliminate "information silos" by promoting cross-system data sharing and exchange. The unified and structured representation further provides a solid foundation for subsequent data analytics and decision-making processes, significantly enhancing processing efficiency and result reliability.

Machine Learning-Based Data Fusion: Machine learning-driven fusion methods leverage advanced algorithms such as deep neural networks to extract and integrate features from multi-source heterogeneous data. For example, Ji proposed a machine learning-based data integration framework that employs deep learning techniques to automatically discover intrinsic data patterns and correlations, achieving high-accuracy fusion performance. Specifically, deep Boltzmann machines have been utilized to project heterogeneous data into a unified feature space for effective integration. Other studies have applied recurrent neural networks to construct temporal feature models for anomaly detection and state recognition tasks. These approaches are capable of capturing complex nonlinear relationships within power systems and adaptively learning discriminative feature representations, thereby significantly improving fusion accuracy and computational efficiency. Consequently, machine learning-based fusion techniques provide new technical pathways for power big data processing and accelerate the intelligent evolution of power system operations.

Rule-Based Data Fusion: Rule-based fusion methods integrate heterogeneous data by applying customized integration rules and domain knowledge constraints. For example, Liu Jia proposed a rule-driven data integration framework that performs data filtering, cleansing, and merging based on predefined integration standards tailored to specific application scenarios. These rules are typically designed according to data attributes and operational requirements to ensure targeted and accurate fusion. In addition, event-driven mechanisms have been introduced to control information exchange and data sharing, enabling dynamic and on-demand integration. Such approaches enhance fusion efficiency and precision through explicit logical constraints and flexible control strategies, providing adaptable solutions for complex power big data environments and improving the practical value of heterogeneous data integration.

3.3.2 Data Fusion Tools

Hadoop and Spark: Hadoop and Spark serve as core technological frameworks for big data processing and provide essential support for distributed storage architectures, parallel computing, and advanced analytics in the power industry. Both platforms are capable of handling large-scale datasets and fully satisfy the high-throughput and high-reliability requirements of power system data processing. Hadoop achieves efficient storage and access of massive data through its distributed file system and parallel computing mechanisms, while Spark significantly improves processing throughput by leveraging inmemory computing paradigms. The synergistic integration of Hadoop and Spark enables fast processing and large-scale integration of power big data, thereby establishing a solid foundation for heterogeneous data fusion and system-level analytics.

Deeplearning4j: Deeplearning4j is a distributed deep learning framework that can be seamlessly integrated with big data platforms such as Hadoop and Spark, enabling large-scale parallel training and efficient processing of massive datasets. By exploiting deep neural network architectures, Deeplearning4j can automatically extract discriminative features and discover latent patterns in power data, thereby facilitating high-accuracy data integration and intelligent analysis. Its strong performance in power data fusion and anomaly detection applications provides effective technical support for enhancing the intelligence level of modern power systems.

ETL Tools: ETL frameworks play a central role in data integration by performing data extraction, transformation, and loading operations. Based on predefined data flow and transformation rules, these frameworks collect data from heterogeneous sources, perform data cleansing and normalization, and load standardized data into integrated platforms, ensuring data accuracy, consistency, and completeness. In power big data integration scenarios, ETL tools significantly optimize processing pipelines, improve integration efficiency and quality, and support systematic data governance. Consequently, they establish a reliable foundation for subsequent analytics and decision-making processes and promote the continuous advancement of heterogeneous data fusion technologies in power systems.

Table 3. Heterogeneous Data Fusion Technologies and Tools for Power

Fusion Technology	Brief Description	Tool	Brief Description
Model-based fusion	Map heterogeneous system data into a unified CIM schema for standardization and integration.	Hadoop & Spark	Big-data platforms enabling large-scale storage and distributed/fast computing for fusion and analytics.
ML-based fusion	Use deep learning/neural networks for feature extraction and multi-source fusion.	Deeplearning4j	Spark-based deep learning framework for fusion and anomaly detection.
Rule-based fusion	Apply predefined rules to filter, cleanse, and merge multi-source data.	ETL tools	Tools (e.g., Talend, Informatica) for extract–transform–load pipelines supporting integration.

3.3.3 Data Sharing Technologies

SOA-Based Data Sharing: Service-oriented architecture (SOA) enables data sharing across heterogeneous power information systems by exposing standardized service interfaces, thereby improving inter-system data exchange and interoperability. For example, Wang developed an SOA-based data sharing platform that reduces coupling among systems while enhancing adaptability and scalability of the sharing mechanism. Standardization and compatibility of interfaces are central to effective SOA implementation, requiring a shared semantic understanding and consistent rules across participating systems. In the power sector, SOA can integrate business functions across dispatching, operation, and customer service departments, facilitating cross-department collaboration and data reuse. With ongoing technological advances and standardization efforts, SOA is expected to play an increasingly important role in power big data sharing.

Cloud Computing-Based Data Sharing: Cloud platforms support cloud-based storage, management, and distribution of power industry data by leveraging elastic computing and large-scale storage capabilities. Li reported that cloud-based infrastructures can improve data management efficiency and enable convenient data access and sharing through network connectivity. Nevertheless, data security and privacy protection remain major challenges for cloud-based sharing, and platform stability and reliability are also critical in power system applications. Power utilities can build centralized data centers on cloud platforms to manage business data and deliver data-driven services to users through cloud-based interfaces.

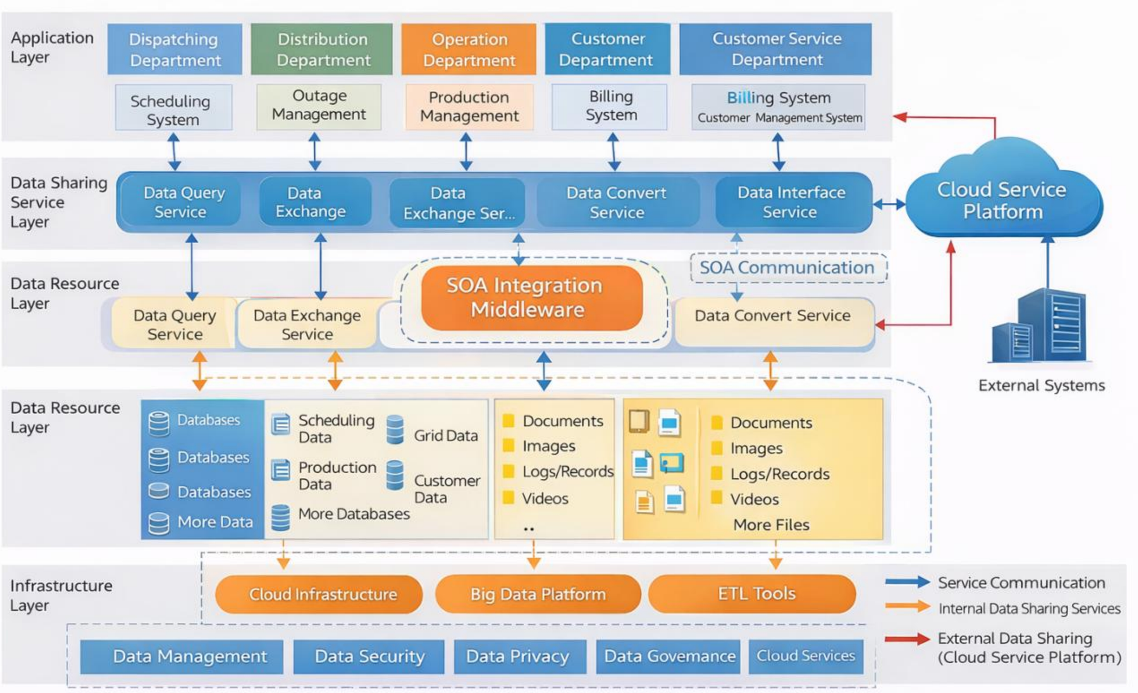

Figure 2. Data Sharing Integration Framework

3.3.4 Data Sharing Tools

RESTful APIs: RESTful APIs are widely adopted interface standards in service-oriented architectures and provide an effective mechanism for cross-system and cross-platform data sharing. By defining unified service interfaces, interoperability among heterogeneous systems can be significantly enhanced. Based on the HTTP protocol, RESTful APIs employ standardized operations such as GET and POST to access and manipulate resources. In power big data sharing applications, multiple API endpoints can be designed to support data querying, retrieval, and service invocation, enabling efficient data exchange among different systems. The standardized interface design simplifies sharing workflows, improves interaction efficiency, and promotes seamless data interoperability across platforms, thereby supporting the digital and intelligent transformation of power systems.

Cloud Storage Services: Cloud storage platforms, such as Alibaba Cloud and Amazon Web Services (AWS), provide scalable and reliable infrastructures for power industry data storage and access. These services are characterized by high scalability, availability, and security, making them well suited for large-scale power big data sharing. By adopting distributed storage architectures, power data are replicated and distributed across multiple nodes, ensuring high system availability and strong fault tolerance. In addition, cloud storage platforms offer flexible access interfaces and fine-grained access control mechanisms, which enhance data security and privacy protection while facilitating efficient data sharing and collaboration.

Table 4. Heterogeneous Data Sharing Technologies and Tools for Power

Sharing Technology	Brief Description	Tool	Brief Description
SOA-based sharing	Standardized service interfaces for cross-system interoperability.	RESTful API	SOA-friendly interfaces enabling cross-platform sharing.
Cloud-based sharing	Store and distribute power data via cloud platforms.	Cloud storage	Services (e.g., Alibaba Cloud, AWS) for scalable sharing.

In heterogeneous power data environments, data integration techniques commonly perform multi-dimensional fusion across temporal, spatial, and semantic dimensions. For example, Gaussian models and hierarchical Bayesian methods can be adopted to improve integration accuracy and computational efficiency. In practice, the power sector often relies on standardized data models and interfaces (e.g., SG-CIM) to normalize heterogeneous data sources and establish a unified foundation for interoperability and data sharing. Moreover, ETL pipelines combined with distributed storage systems enable automated data cleansing, transformation, and loading, thereby ensuring consistent data quality across systems. Finally, shared data platforms and open APIs further enhance cross-department and cross-system data circulation, supporting the continuous development of smart grids.

4. Development Trends and Future Research Directions

4.1 Technology Development Trends

With continuous advances in smart grids and data analytics, heterogeneous integration of power data is exhibiting expanding application potential and accelerating convergence with emerging technologies. Looking ahead, power data integration will increasingly emphasize real-time capability, accuracy, and intelligent processing to support secure, reliable, and high-efficiency power system operations.

Recently, more artificial intelligence (AI) techniques have been introduced into power data fusion. For instance, distance-based clustering methods can detect abnormal attributes by measuring deviations from cluster centroids, while pattern-recognition-driven knowledge discovery can further identify anomalous behaviors in complex datasets. Deep learning has also been applied to load forecasting and renewable generation prediction, improving the accuracy and robustness of forecasting results. Moreover, AI-based data processing can enhance data quality by mitigating the effects of noise and outliers. In addition, distributed ledger technologies (DLT) show promise for enabling more transparent and secure data sharing and management. Due to their distributed storage and tamper-resistance properties, such technologies can strengthen security during data transmission and integration, providing technical safeguards for power system operation.

4.2 Future Research Directions

As smart grids and analytical technologies evolve, heterogeneous power data integration is becoming a key enabler of power system intelligence. Future studies are expected to focus on the following directions.

4.2.1 Development of Efficient Data Fusion Algorithms

Given the diversity, high dimensionality, and real-time requirements of power data, future work should develop more efficient and accurate fusion methods. In particular, leveraging deep learning and AI to automatically learn representative features and identify latent patterns is expected to be a major research focus for improving integration performance.

4.2.2 Development of Data Fusion Platforms and Toolchains

To streamline integration workflows and facilitate data aggregation, sharing, and advanced analytics, there is a strong need for integrated data fusion platforms and toolkits. Such platforms should be scalable, user-friendly, and robust, supporting heterogeneous data modalities and remaining compatible with diverse fusion paradigms, thereby providing a solid foundation for large-scale power data applications.

4.2.3 Establishment of Standards and Specifications

To ensure data quality, security, and interoperability, it is necessary to establish comprehensive standards and specifications for power big data. These standards should cover the full lifecycle of data acquisition, storage, processing, fusion, and application, providing unified technical guidance and strong guarantees for heterogeneous data fusion in power systems.

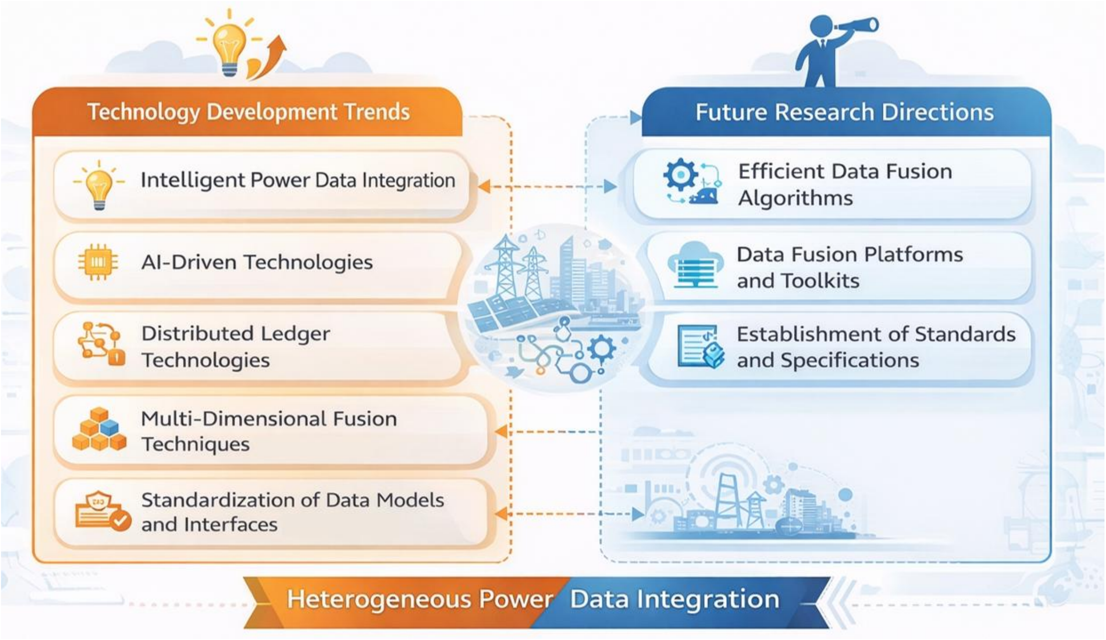
Technology Development Trends
Intelligent Power Data Integration
AI-Driven Technologies
Distributed Ledger Technologies
Multi-Dimensional Fusion Techniques
Standardization of Data Models and Interfaces
Future Research Directions
Efficient Data Fusion Algorithms
Data Fusion Platforms and Toolkits
Establishment of Standards and Specifications
Heterogeneous Power Data Integration

Figure 3. Development Trends and Future Research Directions of Heterogeneous Power Data Integration

References

Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, & Joel Saltz. (2013). Hadoop GIS: a high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow*, *6*(11), 1009-1020. https://doi.org/10.14778/2536222.2536227

Alma'aitah, Wafa' Za'al, et al. (2024). Integration Approaches for Heterogeneous Big Data: A Survey. *Cybernetics and Information Technologies*, *24*(1), Sciendo, 3-20. https://doi.org/10.2478/cait-2024-0001

Chen, Z. et al. (2019). An Uniform Access Method of Heterogeneous Big Data with Power Grid Application. In J. Abawajy, K. K. Choo, R. Islam, Z. Xu, & M. Atiquzzaman (Eds.), *International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018*. ATCI 2018. Advances in Intelligent Systems and Computing, vol 842. Springer, Cham. https://doi.org/10.1007/978-3-319-98776-7_44

Chu, P., Dong, Z., Chen, Y., Yu, C., & Huang, Y. (2020). Research on Multi-source Data Fusion and Mining Based on Big Data. 2020 *International Conference on Virtual Reality and Intelligent Systems* (ICVRIS), Zhangjiajie, China, 2020, pp. 606-609. https://doi.org/10.1109/ICVRIS51417.2020.00149

Dong Xin Luna, Gabrilovich Evgeniy, Heitz Geremy, Horn Wilko, Murphy Kevin, Sun Shaohua, & Zhang Wei. (2014). From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, *7*(10). https://doi.org/10.14778/2732951.2732962

Hongjun, D., Wei, L., Ou, W, et al. (2020). A Power Big Data Processing Method for Heterogeneous Data Sources. *IOP Conference Series Earth and Environmental Science*, *512*(1), 012156. https://doi.org/10.1088/1755-1315/512/1/012156

Hossein Akhavan-Hejazi, & Hamed Mohsenian-Rad. (2018). Power systems big data analytics: An assessment of paradigm shift barriers and prospects. *Energy Reports*, *4*, 91-100. https://doi.org/10.1016/j.egyr.2017.11.002

I Made Putrama, Péter Martinek. (2024). Heterogeneous data integration: Challenges and opportunities. *Data in Brief*, *56*, 110853. https://doi.org/10.1016/j.dib.2024.110853

Kyuseok Shim. (2012). MapReduce algorithms for big data analysis. *Proc. VLDB Endow*, *5*(12), 2016-2017. https://doi.org/10.14778/2367502.2367563

Michał Kunicki, Sebastian Borucki, Dariusz Zmarzły, & Jerzy Frymus. (2020). Data acquisition system for on-line temperature monitoring in power transformers. *Measurement*, *161*, 107909. https://doi.org/10.1016/j.measurement.2020.107909

Xiao, Z., Wu, X., Li, P., Liu, Z., Zhou, Z., Deng, K., Yang, H., Hou, H., Liu, L., & Mao, W. (2019). Power communication network design considering global information fusion part one overall architecture. *Proc. Comput. Sci.*, *155*, 758-762. https://doi.org/10.1016/j.procs.2019.08.110

Zhang, Y., Huang, T., & Bompard, E. F. (2018). Big data analytics in smart grids: a review. *Energy Inform*, *1*, 8. https://doi.org/10.1186/s42162-018-0007-5

Zhao, S., & Wang, E. (2019). Fault diagnosis of circuit breaker energy storage mechanism based on current-vibration entropy weight characteristic and grey wolf optimization–support vector machine. *IEEE Access*, *7*, 86798-86809. https://doi.org/10.1109/ACCESS.2019.2924056