Research on Power Grid Enterprises' Development and Application of Data Mid-End Under Current Trends

Xiao Zhan, Zhifeng Wei, Jie Xu

Abstract: Currently, China's power grid enterprises have considerable strength in data application – information and technology systems have fully covered every business sector, strongly supporting various operations. Considering current trends of information-based development and the practical achievements of Data Mid-End (DME), this paper aims to help Chinese power grid enterprises improve their data services by presenting research on the design of DME from the perspective of primary, compatible, and critical technologies. Moreover, the research provides reference for DME's key technology components by constructing the DME model with Operational Data Layer (source), Integrated Data Layer (sharing), and Application Data Layer (analysis) and ultimately offers theoretical evidence for DME's development and application by presenting typical application scenarios.
Keywords: Power Grid Enterprise, Data Application, Data Mid-End, Big Data Technology
CLC Number: TP3 **Document Code**: A

1 Introduction

In 2003, the U.S. Military initiated "Middle-End" strategy while in war with Iraq. By breaking the barrier between different military sectors and giving more power to the frontier soldiers, the U.S. Military improved every personnel's capability to effectively analyze information and make strategic decisions, and thus increased military power as a whole [1].

In recent years, market competition has only become fiercer due to the increasing popularity of the Internet. Accordingly, Alibaba and other leading corporations have proposed the DME strategy to quickly respond to consumer needs, seize rising opportunities in the market, and strengthen comparative strength [2-4]. Under competitive pressure from the changing consumer market, more companies have begun developing DME in an effort to effectively allocate resources and rapidly improve operational model [5-6]. Meanwhile, big data, a critical technology in DME's development, has been widely adopted in numerous business sectors to improve data quality and carry out scientific operations. For instance, Zhiliang Zhang et al has used big data to analyze data verification. Based on the Hadoop platform that enables parallel processing of data verification, he has proposed a method to manage index data storage that could be applied to DME's storing module.

With the continuous development of network and information technology, power grid companies have become more information based. The amount of enterprise data has drastically increased as the number of information systems increases [9-10]. To solve power grid companies' problems in data applications, these firms should develop DME according to the development needs of ubiquitous Internet of Things while referencing the successful implementations of previous projects in the Internet industry. Once power grid enterprises begin to build and integrate

DME to refine corporational operation systems in management and technology, they will effectively break the data barriers between different system sectors and process enterprise-wide generalized data systematically into corresponding DME with servitization, building a flexible and powerful data sharing service readily available for front-end data analysis.

Considering the strategic goals of power grid enterprises' development of ubiquitous Internet of Things and the development of DME in China and abroad, this paper presents research on power grid enterprises' development and management of DME under current trends. The research provides strategies and solutions for DME's development, effectively improving information systems' ability to quickly respond to needs and thus empowering front-end business operation applications. Enterprises will see an improvement in quality and stronger cross-specialty development. Surrounding the enumeration of key technology, algorithm, and service abilities, this paper provides typical application scenarios as examples for theoretical reference of power grid enterprises' DME's development.

2 Current State of Research

2. 1 Acceleration of Ubiquitous Electric IoT's Development

State Grid has started a program specialized in developing its Mid-End to execute the enterprise's strategy of "Three Modes Two Networks", solve problems related to the development of its information systems, and effectively support the development of ubiquitous electric Internet of Things. State Grid has upgraded resources, systems, and data to "Enterprise Scale" from a managerial perspective, encouraging the renovation of enterprises' organizational structure. From a technological perspective, it processes enterprise-wide operations and data into corresponding operation mid-end and data mid-end with servitization, forming a flexible and powerful service sharing ability, readily available for front-end application development or data analysis. This development satisfies the new requirement of "developing full awareness of state, high-speed processing of information, application of fast and flexible intelligent service system" proposed by ubiquitous electric IoT strategy, and it transforms ubiquitous electric IoT's traditional reliance on investment to dependence on innovative technology and creative models of development. In this way, the continuous growth fuels the updating and innovations centered around customers and allows the recycling of resources between different operations, fostering the exponential growth of businesses and cross-border business innovations.

2. 2 Brief History of Data Applications Development

In digitized society, database technology, an essential technological measure in scientific research and policy-making management, is the core component of information management system, office work automation system, policy-making support system, and other information systems. Ever since its birth less than half a century ago, database technology has formed sound theoretical basis, mature commercial products, and expansive fields of application, attracting more and more researchers. The advent and development of database has brought a significant revolution to the computer information management. Over the last few decades, China and the international community have developed tens of thousands of databases, establishing database as the infrastructure for daily office work, production, and life of enterprises, business units, and even individuals. At the same time, following the expansion and in-depth exploration of applications, the number and scale of databases have both increased, and the research of databases has correspondingly become more profound. Since the establishment of A.M. Turing Award in 1996, four awards have been given to works related to database (C. W. Bachmanl, 1973; E. F. Codd, 1983; J. Gray, 1998; Michael Stonebraker, 2014), underlining the fact that database is a field full of innovative energy.

In 1990 Bill Inmon proposed the concept of database whose main functionality is to utilize its quintessential document storage structure to systematically analyze and sort the massive amount of On-Line Transaction Processing (OLTP) data generated through information systems, benefiting various analysis approaches such as On-Line Analytic Processing (OLAP) and Data Mining while supporting the development of Decision Support System (DSS) and Executive Information System (EIS). In this way, database can help decision makers pinpoint valuable information from vast amount of data and thus quickly respond to change in trends and develop Business Intelligence (BI).

Since 2009, "Big Data" has become popular in the internet information technology industry. The early development of Big Data attributed to its applications in internet businesses – internet data increased by 50% annually, doubling every two years. Global internet enterprises have all realized the advent of the Big Data era as big data has become an integral part of companies.

In May 2011, McKinsey Global Institute published "Big Data: The Next Frontier for Innovation, Competition, and Productivity". Since the release of the report, Big Data has become the leading concept in computer science.

In March 2012, the U.S. government published "Big Data Research and Development Initiative", announcing an investment of $200 billion to start the program.

In July 2012, China published "National Strategic 12th Five Year Plan for the Development of Emerging Industries" whose proposed key technologies are closely related to Big Data. In March 2013, Chinese Society for Electrical Engineering published "White Paper for the Development of Chinese Electric Big Data", underlining the drastic increase in the value resulting from the application of Big Data technology in the electric power industry.

In 2015, Alibaba proposed Data Mid-End strategy, enabling the decoupling of data layers and enhancing public data abilities. The strategy has three main components: data model, data service, and data development. Data model enables the cross-domain aggregation of data and the consolidation of knowledge. Data service enables the packaging and development data, quickly and flexibly meeting the needs of high-level applications. Data development tools satisfies the demands of personalized data and applications.

2. 3 China and the International Community's Research on the Design of Middle-End

In the International Community, Google publicized a series of mass data processing technologies that are implemented in-house – Distributed File System (DFS) based on redundant storage, the parallel processing framework MapReduce used for index computing search, efficient data storage model BigTable, etc. These modules make up Hadoop that employs a distributed system structure. In 2012, the measurement for global data processing has increased from terabytes to petabytes, and the open-source Hadoop environment has become commercialized. Through investment of data technology research and complete transformation to data technology firms, leading Internet companies have, in the commercial market of business expansion and data's exponential increase, trailblazed Big Data technology, accumulated valuable experience, and implemented Big Data, processing great potential for growth. Many international giants, such as Microsoft, Google, IBM, etc., have fueled the research on Big Data technology, quickly verifying the research results through business scenarios and cloud-sharing their implementations through IaaS and PaaS. Currently, the international community has explored Big Data and its application and has yet to propose the DME concept.

In China, Alibaba was the first to put forward the DME strategy, enabling the decoupling of data layers and enhancing public data abilities. Through the establishment of DME, Alibaba accomplished the sharing of business platforms such as TaoBao, TMall, 1688, JuHuaSuan, ELEME, Ant Financial, etc. With DME, Alibaba has also consolidated its data analysis capability to assist its business and store consulting service to promote sales. Alibaba Index has become the foundation of data analysis for media application, industry development, and government policy making. The cooperation between Alibaba's Data Mid-End and Business Operation Mid-End is the major force behind the exponential growth of Alibaba's businesses. For instance, its subcompany DiDi currently has 450 million users, 21 million car owners, and provides cab service for more than 400 cities. Originally, DiDi utilized vertical structure that had poor sharing abilities, redundant development, and low development quality. DiDi employed Mid-End strategy at the end of 2015, developing a sizeable Mid-End for daily transportation that contains a flexible three-layer service structure, achieving servitization and configurability. The operation end is configured through spontaneous platform dynamics and uses DME to support the rapid development of new business operations.

By meeting the needs of businesses that grow exponentially, Chinese Internet enterprises have accumulated theoretical and practical experience that serves as references for future endeavors, but the operation model of Internet companies is quite different from that of the power industry, so one must take into account the reality of power business operations in developing the structural design of DME and analyzing its real-world applications.

2. 4 Power Grid Enterprises' Groundwork

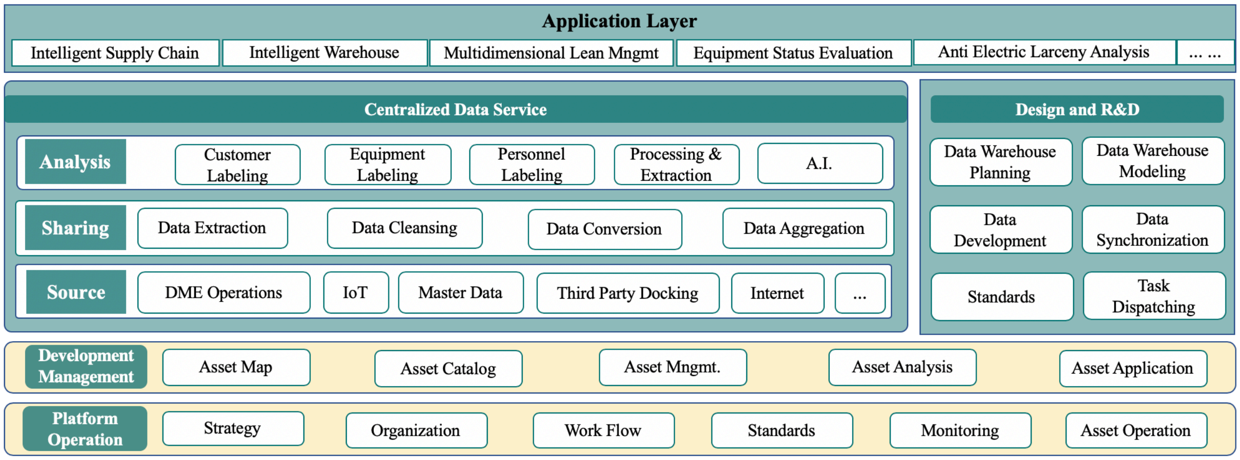
From an early stage, State Grid has accumulated a wealth of experience in data aggregation, analysis, and application through the development of data center, data resource management tools, master data, data sharing, business operation integration, and centralized data analysis domain. State Grid has seen success in establishing the workflow and standards of data management as well as managerial strategies. For data management workflow, State Grid has established management systems and procedures in data indicator definition, data storage, data sharing, and data quality. In terms of data management standards, State Grid has implemented "Data Center Management Methods". As for managerial strategies, State Grid has researched and developed data resource management tools to administer the information flow and quality of structural data resource's standing book. The results and experience from early work has established a solid foundation for State Grid's data asset management.

2016 marked the initiation of State Grid's Centralized Data Center after three years of comprehensive development. The Centralized Data Center is able to achieve "the unification of model standards, the clarity and transparency of data, and the flexibility and intelligence of analysis." As the convergence of data of every operation, every class, and every time period, the data center also provides complete data resources, efficient analysis computing capability, and unified operating environment, accomplishing the development goal of "managing enterprises with data, driving businesses with information." In 2018, State Grid begun the development of holistic and comprehensive demonstration, effectively verifying the holistic data model. In the aspect of data processing, State Grid's Shanghai Power and Jiangsu Power have accomplished the improvement of business operations' collaboration in the project domain and in the streamlining of sales, operation inspection, and allocation, exploring the development of data processing based on the Centralized Data Center as well as the problems and needs of horizontal business collaboration. As for data management, ShaanXi Power leads holistic model design. Considering the core business operations and based on the needs of cross-specialty horizontal business collaboration and generalized application demonstration, it has finished designing the core business logic and physical data model for the following domains: personnel, finance, supply, project, asset, power grid, customer, market, security, and the general domain. In data analysis, State Grid has imported 45 business systems containing 820,000 structured tabular data, achieving the data cleansing and conversion of 50 source business operation systems, 3495 source tables, 63366 source fields, 2772 standard codes. The Centralized Data Center contains five main components: data importation and conversion, data storage and computing, data analysis service, data management, and master data management. The development of Centralized Data Center lays a solid foundation for future development of Ubiquitous Electric IoT's Data Mid-End.

3 Design of Data Mid-End

3. 1 Structural Design of Data Mid-End

The overall structure of Data Mid-End consists of Operational Data Layer (source), Integrated Data Layer (sharing), and Application Data Layer (analysis), Data Asset Management, Design and R&D that all support centralized data service, intelligent supply chain, intelligent warehouse, multidimensional lean management, and other data applications.



Graph 1: Data Mid-End Structural Design

Data Mid-End focused on improving capabilities of data asset management, data importation and integration, data sharing and analysis, and support of primary modules. According to the principle of "Visible Data, Mature Modules, Standardized System", State Grid would release a list of mature modules, and province-wide (city-wide) companies can choose modules to develop their own Data Mid-End based on Centralized Data Center. Data Mid-End and Centralized Data Center would develop simultaneously; after demanded core data and services are transferred to Data Mid-End, companies would merge utility-maximizing modules into Data Mid-End.

From dispersed collection tools, Operational Data Layer conducts unified convergence and servitization on various data ports such as terminal collection, operation system, geographic information, exterior data, etc. Operational Data Layer enables the unified management and monitoring of data collection service, providing service management capabilities for the efficient collection of Ubiquitous Electric IoT's data. Recently State Grid has improved the standardized management of terminal data collection service based on the development of IoT management center.

Integrated Data Layer provides a unified platform for data sharing, data exchange, data integration, data trading, and other data activities through standardized management. This layer enables the standardization and unified monitoring and administration of various data sharing and porting services, supporting data connectivity of applications in power grid resource center, providing centralized service and management capabilities for Ubiquitous Electric IoT's fast and secure sharing of data.

Application Data Layer integrates and processes data based on business operation demands and unified data standards, developing public data layer, data extraction layer, and service analysis layer. This layer develops data labeling and data analysis lab, providing an exploration- and experimentation-focused environment and quickly responding to demands while consolidating generalized data. It also contributes to the rapid development of enterprises' valuable data analysis applications by forming enterprise-wide analysis system, supporting multidimensional lean management and its statistical reports, and digitized audit report, etc.

Focusing on data standards and data resource service capabilities, Data Asset Management Layer develops master data standardization, data model, etc., supporting the development of multidimensional lean management. Utilizing data graphing technologies, this layer presents straightforward, clear, and easily understandable enterprise scaled data maps, including business operation data map, data link map, data model map, etc., forging data resources maps that allow the traceability and transparency of ledger.

As for Design and R&D, DME's structure consists of Platform Layer, Storage & Computing Layer, Development Tools Layer, Centralized Service Layer, Data Importing & Processing Tools Layer, and Data Asset Management Center Layer from bottom to top. Explanations and corresponding technologies of each layer is shown below.

LAYERS	COMPONENTS	TECHNOLOGY
Platform	* Based on the development of State Grid Cloud platform, this layer supports the stable operation of DME using resources offered by State Grid Cloud such as computation, storage, network, security, data, and middleware.	
Storage & Computation	Modules include: Structural Data, Nonstructural Data, Data Computation (Batch Computing, Real-Time Computing, Memory Computing)	NoSQL Database, Distributed Database, Analysis Database, Big Data Platform, etc.
Development Tools	Audit Report, Data Mining, Data Exploration, Algorithm Management	MapReduce, Hive, Spark, Flink, Impala, Tez, Akka, Storm, S4
Centralized Service	Service Packaging, Service Registration, Service Routing, Configuration Center, Permission Control, Service Choreography, Service Monitoring, Service Container, etc. These components support the packaging and releasing of data services, management of	ZeroMQ, Dubbo, Socket (Mina, Netty), Echarts, Tableau, TreeSoft

	service catalog, sharing of model algorithms, and subscription of digital products, etc.	
Data Importing & Processing Tools	Data Replication, ETL, Data Aggregation, Data Exchange, Data Cleansing / Conversion, Data Integration / Correlation, and other importing modules.	Kafka, ActiveMq, PKafka, ZeroMQ, Dubbo, Socket (Mina, Netty), FTP / SFTP, RestFul, Web Service
Data Asset Management Center	Management modules include: Asset Catalog, Data Security, Master Data, Data Sharing, Index, Labeling, Metadata, Data Quality, etc.	Web Service

3. 2 DME Deployment Strategy

Intranet can be divided into two parts, headquarter level and province-wide (city-wide) company level, and the two levels will perform data exchange. Extranet would solely be based on the headquarter level. Headquarter Data Mid-End is stationed at SanDi Data Center and willy be deployed at Beijing Data Center in early stages. Intranet Data would be synchronized to Extranet Data Mid-End after data masking and decryption according to State Grid's security standards. Likewise, Extranet Data would be synchronized to Intranet Data Mid-End, and the Extranet applications that require services from Intranet Data Mid-End would develop data service proxy within the Extranet.

3. 3 Description of Core Technologies

3. 3. 1 Technologies

1. Kafka system. Kafka is a dispersed publishing – subscription information system. Its consistent log service equipped with redundant backup is dispersed and divisible, largely utilized in processing active data streaming. Kafka lowers the complexity of coding and system networking – various subsystems no longer negotiate for ports. Rather, they can be seen as plugs in sockets, and Kafka would handle high-throughput data.
2. HBase database. HBase is an open-source version of BigTable (written in Java). HBase is Apache Hadoop's database based on HDFS. Designed to provide reliable, high-performing, column storing, scalable, and multi-version NoSQL dispersed data storage system, enabling the real-time, stochastic reading and writing of large quantities of data. Hadoop's stochastic storing database applications include HBase, Cassandra, CouchDB, Dynamo, and MongoDB.
3. Spark computing engine. Spark is a fast and versatile dispersed computing system, offering

high-level API such as Java, Scala, Python, and R. It also supports high-level tools such as Spark QL for processing structural data, MLib for handling machine learning, GraphX for graph computation, and Spark Streaming for processing data streaming, indicating that Spark offers flexible Big Data processing power that enriches interfaces.

4. RESTful is a design style and development approach for Internet applications. Based on HTTP, RESTful can be defined with XML or JSON. RESTful is compatible with operation enabling interface for mobile Internet manufacturers, empowering third party OTT's adding, changing, and deleting mobile Internet resources.
5. Tableau is a data analysis software that is easy to use. By importing data and manipulating data, Tableau is able to analyze data and generate tables and graph presenting information based on the given data. Companies can transfer large amount of data onto Tableau's digital canvas, and Tableau will convert the information into tables and graphs in an instant. Tableau's friendly user interface allows companies to have a thorough understanding of the business operation decisions they have made.

3. 3. 2 Algorithms

3. 3. 2. 1 Text Segmentation Algorithm

Suppose a sentence *M* has *m* number of segmentation choices:

$$\{A_{11}A_{12}\ldots A_{1n_1}, A_{21}A_{22}\ldots A_{2n_2}, \ldots, A_{m1}A_{m2}\ldots A_{mn_m}\}$$

where n_i indicates the number of words for the *i*th segmentation choice. If we have selected the best segmentation method that happens to be the *r*th choice, then it must be certain that this method has the largest probability distribution, which indicates:

$$r = \arg\max_i P\left(A_{i1}, A_{i2}, \ldots, A_{in_i}\right).$$

To simplify calculations, we could use the Markov Assumption. Suppose the probability of the appearance of a unit solely relies on the previous unit:

$$P\left(A_{ij} \middle| A_{i1}, A_{i2}, \ldots, A_{i(j-1)}\right) = P\left(A_{ij} \middle| A_{i(j-1)}\right)$$

After joint distribution,

$$P\left(A_{i1}, A_{i2}, \ldots, A_{in_i}\right) = P(A_{i1})P(A_{i2}|A_{i1})P(A_{i3}|A_{i2})\ldots P\left(A_{in_i} \middle| A_{i(n_i-1)}\right)$$

Through the software's standard language library, we can approximate the two-variable conditional probability between any two units. For instance, for any two words a_1, a_2, their conditional probability distribution can be expressed as:

$$P(a_2|a_1) = \frac{P(a_1, a_2)}{P(a_1)} \approx \frac{freq(a_1, a_2)}{freq(a_1)}$$

$$P(a_1|a_2) = \frac{P(a_2, a_1)}{P(a_2)} \approx \frac{freq(a_1, a_2)}{freq(a_2)}$$

where $freq(a_1, a_2)$ expresses the frequency that the two words have appeared back to back in the given language library, and $freq(a_1)$、$freq(a_2)$ respectively expresses the frequency of a_1, a_2 in the language library. Therefore, the best segmentation method can be found by looking for the highest probability in the joint probability distribution for every segmentation method considering the probability distribution of each word based on the language library.

The most common method is Jieba Chinese text segmentation. Given the careful selection of language library, Jieba segmentation efficiently classifies vocabularies through data structure's trie (prefix tree). It also solves the problem of misinterpreting combination of words utilizing Directed Acyclic Graph (DAG). What's more, for vocabularies not included in the language library, Jieba segmentation divides the word into beginning, middle, end, and stand-alone words using the Viterbi algorithm and the HMM model that constructs Chinese words. This segmentation method is applicable for text mining capabilities under data analysis and data mining, providing keyword libraries by extracting and producing keywords from power grid facility resources.

3. 3. 2. 2 The Entropy Weight Method for Quantity of Information

The Entropy Weight Method is an algorithm that determines measures of dispersion of a certain parameter. The more dispersed it is, the more effect the parameter has on general evaluations. The most dispersed parameter would be seen as most critical. Specific steps of computation are shown below:

1. Given *n* samples and *m* parameters, r_{ij} is the value of the *j*th parameter of the *i*th sample, ($i = 1,2,...,n; j = 1,2,...,m$).
2. Normalization of parameters: for positive parameters,
$$r_{ij}^{'} = \frac{r_{ij} - \min\{r_{ij},...,r_{nj}\}}{\max\{r_{1j},...,r_{nj}\} - \min\{r_{1j},...,r_{nj}\}}$$
For negative parameters,
$$r_{ij}^{'} = \frac{\max\{r_{1j},...,r_{nj}\} - r_{ij}}{\max\{r_{1j},...,r_{nj}\} - \min\{r_{ij},...,r_{nj}\}}$$
where $r_{ij}^{'}$ is the value of the *j*th parameter of the *i*th sample. To simplify our calculations, normalized data would still be written as r_{ij} and $r_{ij} = \left|r_{ij}^{'}\right|$.
3. Calculate the corresponding weight of the *i*th sample under the *j*th parameter:
$$p_{ij} = r_{ij} / \left(\sum_{i=1}^{n} r_{ij}\right), i = 1,...,n, j = 1,...,m$$
4. Calculate the entropy of the *j*th parameter:
$$e_j = -k\sum_{i=1}^{n} p_{ij} In(p_{ij}), k = 1/In(n), e_j \geq 0$$
5. Calculate the redundancy of information entropy:
$$d_j = 1 - e_j$$
6. Calculate the corresponding weight of each parameter:

$$\omega_j = \frac{d_j}{\sum_{j=1}^{m} d_j}$$

which would be the parameter's importance score.

This algorithm is applicable for classification prediction capabilities of data analysis and data mining such as customer labeling service under power grid customer service operation.

3. 3. 2. 3 k-Nearest Neighbors Algorithm

Given a training data set and a new instance, this algorithm finds *k* data points that are closest to the new instance in the training data set. The new instance would be put into the class that contains the majority of the *k* data points.

Input: training data set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ where $x_i \in \chi \subseteq R^n$ is the new instance's eigenvector and $y_i \in \gamma = \{c_1, c_2, \ldots, c_K\}$ is the new instance's category, $i = 1,2,\ldots,N$.

Output: the *y* class where instance *x* belongs.

1. According to distance measurements, we find *k* data points that are closest to the new instance *x* in training set *T*. The region containing the k data points is $N_k(x)$.
2. Determine *x*'s category according to classifying strategies:

$$y = \arg\max_{c_j} \sum_{x_i \in N_k(x)} I(y_i = c_j), i = 1,2,\ldots,N\,;\, j = 1,2,\ldots,K$$

In the above equation, *I* is a binary indicator function, which means $I = 1$ when $y_i = c_j$, otherwise $I = 0$.

This algorithm is applicable for classification prediction capabilities of data analysis and data mining such as research on power-usage-sensitive customers, Artificial Intelligence algorithms, etc.

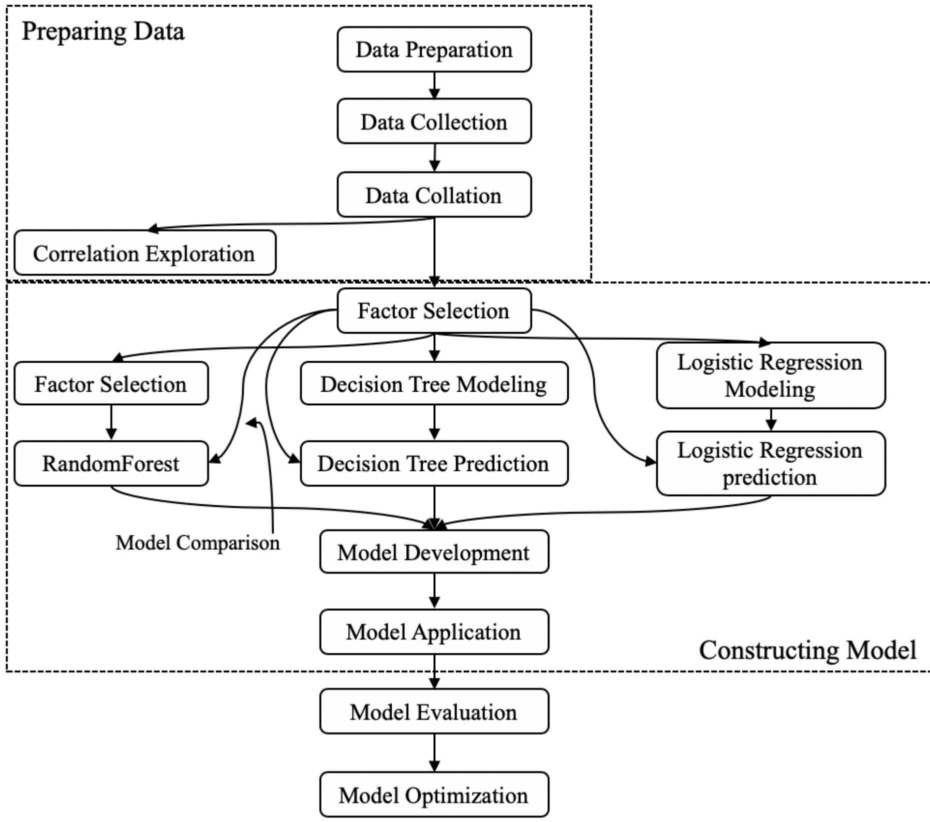
4 Typical Application Scenarios

4. 1 Application of Customer Labeling Service in Data Mid-End

4. 1. 1 Description of Scenario

A case in point is the online customer service application scenario. Combined with data from online customer service channels, DME analyzes customers' basic attributes, hidden attributes, and preferences utilizing the entropy weight method, main component analysis, decision tree, logistic regression. DME constructs the analysis model for online customer service personnel, refines customer labels, and enriches user profiles through which DME can accumulate knowledge concerning online customers. Online customer service would be refined at the same time in order to improve customer service perception and user experience.

Through data analysis and data mining of user characteristics on online customer service channels, DME constructs a preference model and generates online customer channel preference labels.



Graph 2: Online Customer Service Customer Labeling Workflow

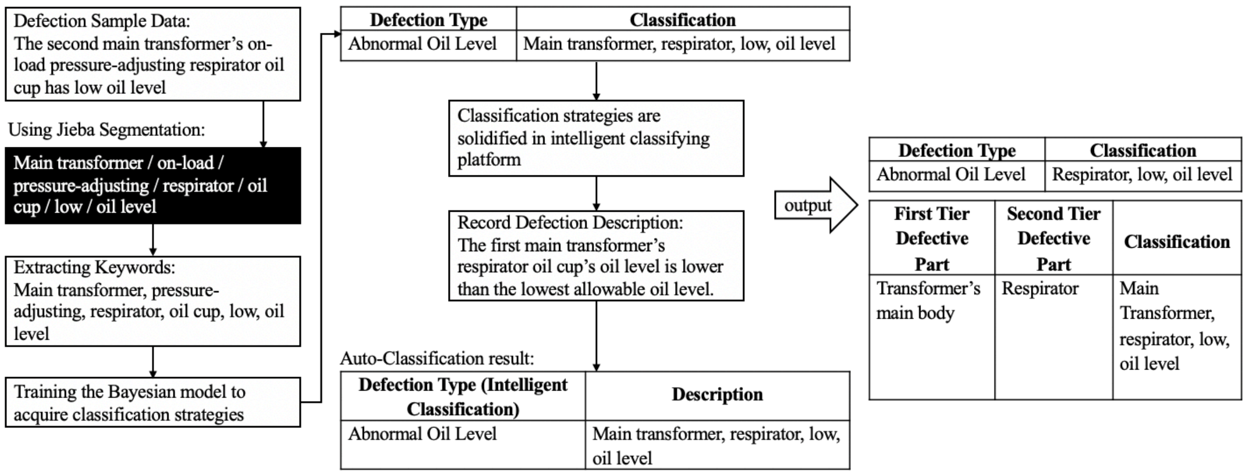
4. 1. 3 Core Technology Application

Based on business operation analysis, DME uses various Big Data service engines offered by conceptual data. In this way, DME can employ comparison and classification model, information Entropy Weight Method, k-Nearest Neighbor Algorithm, logistical regression, etc. to develop, evaluate, and refine models.

4. 2 Defective Equipment Management

4. 2. 1 Description of Scenario

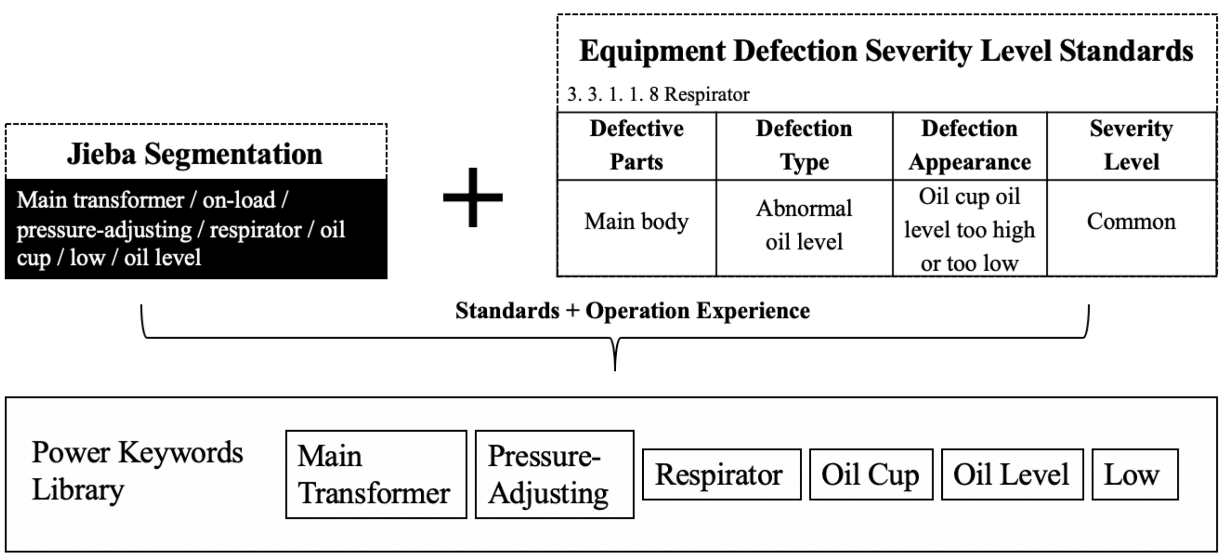
As one of the most important aspects of power equipment maintenance, defective management has been the center of attention for power enterprises management personnel as well as production and maintenance personnel. The fact that equipment sometimes lacks standardization in describing and resolving defections negatively affects defective management abilities. An accurate description of on-site defections can better assist in text segmentation and acquire a keyword library. Defective equipment management classifies defection data into three categories: defective parts, types of defection, and causes of defection. The defective parts category can be further divided into first tier, second tier, and third tier. The causes of defection category can be divided into first tier and second tier which could use algorithms to conduct iterated computation. The graph below presents the entire process of automated classification. The process uses types of defection as inputs for classification. For instance, the input could be “The second main transformer’s on-load pressure-adjusting respirator oil cup has low level of oil”, and the output would be “unusual oil level” (the actual input and output would be in Chinese as Jieba supports Chinese text segmentation).



Graph 3: Classification Process

4. 2. 2 Core Technology Application

Using Jieba text segmentation and TF-IDF algorithm that calculates the frequency of vocabularies, the classification process generates work order keywords, builds keyword library and uses the keyboard library as a basis to construct data samples for classifying strategies. Ultimately, the process employs machines learning and the Bayesian Model to train classifying strategies.



Graph 4: Keywords Library Construction

5 Conclusion

Referencing China's and the international community's research on Mid-End development and considering the current data application trends of power grid enterprises, this paper explores the design of power grid's Data Mid-End and mainly improves data asset management, data importing and integration, data sharing and analysis, and support of primary modules. The current research also describes the technologies and algorithms used in the development of Data Mid-End and presents typical application scenarios for Data Mid-End's service in an effort to provide theoretical reference and lay the groundwork for power grid enterprises' DME development.

Works Cited

[1] Xinyu Chen, Jiaying Luo, Tong Deng, et al. *Mid-End Strategy – Development of Mid-End and Digitized Businesses*. China Machine Press, 2019.

[2] Yuan Li. *The Advent of "Mid-End".* China Entrepreneur, 2019, 565(05):52-55.

[3] Xiaoyong Chen, Hongyu He, Li Huo. *Research on Design Scheme of Intelligent Operating Enterprise Mid-End*. [J] Digital Technology & Application, 2018, 36(11):120-121.

[4] Hong Lin, Xuemin Fang, Bao Yuan, Hong OuYang, et al. *Research and Design of Electric Power IoT's Multi-Channel Customer Service Mid-End Strategy.* [J] Distribution & Utilization, 2019, 06:39-45.

[5] Hongfu Zhu. *Develop Enterprises' Data Mid-End, Encourage Enterprises' Intelligent Operation.* [J] C-Enterprise Management, 2018.

[6] Gang Xu. *Research on Omni-Channel Scenario-Based Operation of Enterprises' Mid-End*. [J] Information & Communications, 2017(8).

[7] Yunhao Li, Jianxue Wang, Xiuli Wang. *Evaluation Methods for Power Grid System Loss based on Mixed Cluster Analysis*. [J] Automation of Electric power Systems, 2016(1): 60-65

[8] Rui Mai, Xie Zhou, Zhou Peng, et al. *Data Mining of Load Characteristics and Quota Correlation Features Considering Temperature Factors*. [J] Proceedings of the CSEE, 2015 (1).

[9] Zhiliang Zhang, Yehua Sun, Chengzhi Chen, et al. *Quality Verification Methods and System for Power Grid Data based on Hadoop.* [J]. Journal of Computer Research and Development, 2014 (S2): 134-144.

[10] Xiaoli Ju. *Electric Power Sales Analysis System based on Data Warehouse Technology*. [J] Digital Technology & Application, 2012 (3): 68-69.

[11] Yijun Zhu, Cheng Wang. *Data Warehouse and its Design in the Application of Electric Power Systems*. [J] ZheJiang Electrical Power, 2003, 22 (4): 146-148.