

# Construction of E-government Data Sharing Framework Based on Big Data Technology

FangQin<sup>1</sup>, Linli<sup>1\*</sup>, WeijiaZeng<sup>1</sup>, TaoWang<sup>1</sup>, Can Wang<sup>1</sup>, LinlinYu<sup>1</sup>

<sup>1,\*</sup>Dalian University of Science and Technology

**Abstract.** With the continuous development of Internet technology, more and more fields are beginning to use Internet technology to improve their business construction. Moreover, in terms of e-government construction, introducing the Internet technology into e-government work at all levels is also an important content, and the Party Central Committee has proposed a series of major decisions such as network power and "Internet +" to accelerate the process of e-government informatization. Meanwhile, big data virtualization technology and access control technology are used to break through e-government data sharing barriers to realize virtual integration and trusted sharing across e-government data and support multi-departmental, cross-level, and cross-network data security interactive processing for e-government business collaboration. Besides, the multi-source circulation and on-demand access of cross-departmental e-government data are realized to break data barriers, form data transmission links, and create a model for exchanging and sharing e-government data. Through the exchange and sharing of e-government data, the barrier-free flow of information can be achieved and the efficiency of e-government business handling can be improved as well.

## 1 Introduction

In recent years, the government has continuously optimized its service model in terms of government services. In the context of the continuous development of information technologies such as the Internet, big data, cloud computing, and the Internet of Things, integrating the "Internet +" thinking has been a new way of thinking for the government. Moreover, through the development of a new model of "Internet + government services", it can provide the public with more convenient and efficient government services [1].

The concept of "Internet + government services" proposed by China mainly shows the development and transformation of government functions, which extends the services it provides to the Internet field. Nowadays, the provincial "Internet + government service" platform established by various provinces serves as the unified government service entrance of each province, which provides great convenience for the public and enterprises to receive services [2]. Meanwhile, by gathering government resources, a unified sharing platform is established to and integrate and reconstruct government services, and "Internet + government services" further optimizes and adjusts the internal organizational structure of the government, which continuously improves the government's management level [3].

Taking the perspectives mentioned above as the research ideas, according to the research and case analysis,

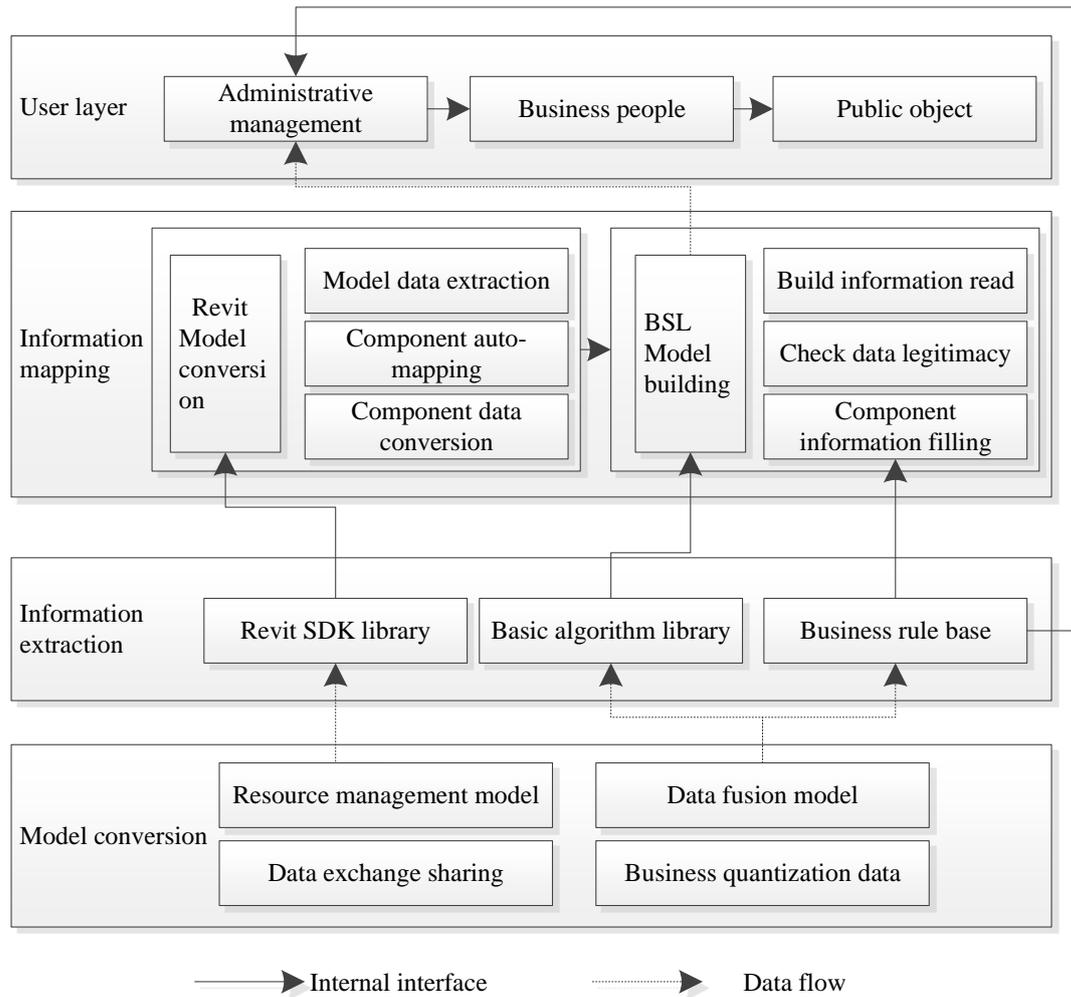
drawing on some advanced domestic experience, and combining institutional mechanisms and data sharing and integration, the obstacles and constraints in the implementation of the "Internet + government service" model in Fuzhou are explored in the paper. Meanwhile, considering the overall idea of building a national integrated online service platform and the existing experience of other local governments, basic countermeasures are proposed in the paper as well.

## 2 Overall Module Introduction and Module Division

### 2.1 Overall Introduction of the Module

In China, e-government functions cover all aspects of the administrative field, which uses administrative means to maintain and guarantee the social harmony and stability[4]. Moreover, the entire system is used for sharing and exchanging multi-business administrative data, realizing cross-business administrative data circulation, solving the problem of data islands, and greatly improving the efficiency of administrative business processing. Additionally, the overall architecture is divided into five parts: user layer, application system layer, application support layer, information resource layer and basic support. Each part complements with each other to form an organic whole together. Figure 1 is the overall architecture diagram [5].

\* Corresponding author: 734380844@qq.com



**Figure 1.** The overall architecture of the e-government system

(1) User layer. The user objects mainly include urban administrative leaders, administrative objects such as relevant staff engaged in administrative work in the urban area, and the general public.

(2) Application system layer. The application system layer is quickly built based on the application support layer, including the operation supervision and evaluation system, the collaborative scheduling system, the administrative service business processing system and the administrative service analysis and judgment system [6].

1) Operation monitoring and evaluation system. During the operation of the system, the operation of the system is uniformly supervised and evaluated for effectiveness.

2) Cooperative scheduling system. It is mainly oriented to management objects and internal personnel engaged in administrative work, which captures service needs from multiple channels such as physical service halls, online service halls, community terminals, and mobile terminals. Meanwhile, through the collaborative scheduling platform, the matters, quality, time limit and other elements handled by the department are respectively pushed to the business processing platform, and the business platform carries out unified task reception, acceptance, processing and feedback of the processing

results [7].

3) Administrative service handling system. It mainly includes new subsystems, expansion and upgrade and data docking of existing system functions, .

4) Administrative service analysis and judgment system. The data visualization technology is used to display administrative data. Meanwhile, statistics, analysis and judgment can be made to make the data results more intuitive [8].

(3) Application support layer. The application support platform is the infrastructure of the overall project, including data display components, data analysis components and application support components. Through the cooperation among different components, the execution efficiency of the system can be improved and a more complete system will be constructed.

(4) Information resource layer. The information resource layer is the administrative service information resource center. The main function is to collect administrative data from the administrative business system, administrative departments or the Internet. Moreover, according to the set standards, the collected administrative data is aggregated and integrated to form a basic personnel model and a basic organization model. Meanwhile, more business information models will be

built on this basis. In addition, through a series of methods such as standardized resource management, resource collection management, data fusion management, etc., a complete administrative information resource center will be formed.

(5) Basic support layer. The basic support layer is the infrastructure platform that supports the overall project. The basic support layer includes the network and the host/storage, and the network is composed of the Internet, the administrative extranet and the administrative intranet. The host/storage is the administrative cloud platform of a certain city [9].

## 2.2 Demand Analysis and Module Division

This system needs to build administrative data barrier-free exchange and sharing. The whole system is divided into three sub-modules, namely administrative data management sub-module, administrative data on-demand access sub-module, and administrative data sharing and exchange sub-module. The requirements of each part are shown in Table 1.

**Table 1** Demand analysis table

Module name	Module function
Administrative data management sub-module	Manage data sources, basic data, metadata and catalogs
Administrative data access sub-modules on demand	Realize data service management and administrative data access with designated permissions
Administrative data sharing and exchange sub-module	Circulate data within a certain range to realize data sharing

The detailed requirements of each part are as follows:

(1) Integrated management of multi-source administrative data. In order to solve the problems of inconsistent data storage and difficulty in sharing and exchange in the development of administrative processes, it is necessary to manage multi-source administrative data. The functions that should be implemented are as follows. Collect administrative data from different sources, then record and manage the basic information, characteristic information and connection information of the data. Moreover, carry out the logic package and class design of basic data management, metadata management and catalog management to further explain the administrative data structure [10].

(2) Control administrative data access as needed. In order to ensure the security of administrative data sharing and exchange, it is necessary to control and manage data query and access. In terms of data services, it manages data access services, data resource permissions, and ETL services. Moreover, task division and role design are used to further control the acquisition of administrative data [11].

(3) Virtualized data exchange and sharing. With the continuous improvement and development of administrative work, the current administrative data storage formats in China are diverse, and the amount of data is very large. In order to complete the sharing and exchange of administrative data, it is necessary to use data virtualization technology. Therefore, data virtualization technology is used to design a new administrative data sharing and exchange architecture, which aims to realize information resource management, multi-source data collection, dispatch command for basic information sharing and exchange, source data integration management and control, basic information model formation, and dispatch and command for basic information sharing and exchange[12].

## 3 Virtual Database Design

### 3.1 Virtualized Database

In the administrative bureau's business system, due to the different responsibilities of each department and the different administrative business involved, the data storage methods are not consistent, which makes it difficult to share administrative data and cannot realize the barrier-free flow of data. However, after combing the administrative business as a whole, it is found that many types of administrative data are interrelated and can perform unified storage access management. For example, the basic information of administrative users needs to be added in many basic administrative services. If the information is operated one by one, it is bound to be a very tedious task. Therefore, a unified virtualized database can be established to eliminate the problems caused by heterogeneous data from multiple sources, and to better realize the sharing of administrative data.

Data virtualization can realize the unified management of data. When querying and retrieving data, there is no need to know the details of the data, but only need to use the data virtualization technology to integrate the administrative data in a unified manner. Then, on the basis of not knowing the specific storage location of the data, the data results can be queried with the help of metadata. In addition, by establishing a virtualized database, users' operations can be greatly simplified. Meanwhile, effective barrier-free query of administrative data can be realized, and the efficiency of administrative work can be improved.

### 3.2 Virtual Database Construction

The focus of the construction of a virtualized database is divided into three processes. The first one is to obtain

administrative data from different businesses, and the second one is to set a unified standard for data and assimilate. The third one is to provide external access interfaces and realize data query on the basis of shielding internal details. Among them, configuring the wrapper and establishing the mapping relationship are two important contents.

(1) Configure the wrapper. In the access process of data virtualization, there is no need to know the data access operation at each step. The same effect can be achieved by directly using the wrapper. What is more, the wrapper can store the necessary data access information and encapsulate the relevant information of the database. When performing data access, according to the address information data provided by the virtualization system, access the related services, and the services contained in the services are the required administrative data information.

(2) Establish a mapping relationship. A virtual view is established in the virtual database, and the multi-source database can be connected with the virtual data through the virtual view. Meanwhile, the required administrative data in the multi-source database can be correspondingly found through the field information in the virtual view. In addition, through the virtual view, the user's query operation can be more simple and convenient.

### **3.3 Sub-module of Administrative Data On-demand Access**

#### *3.3.1 Problem Analysis*

Due to the extremely high confidentiality of administrative data, administrative data will involve a large number of personal privacy, personal information and other information that needs to be kept secret. Therefore, attention needs to be paid to data security when accessing administrative data, and data should be obtained on demand according to the level of data visitors and the type of administrative business.

#### *3.3.2 Data Service Management*

The data service management system provides data access to the basic database corresponding to the shared database in the form of service, realizing the unified management, maintenance and control of data access to the shared basic database and providing a unified technical way for all business systems to access data resources.

(1) Data access service. Data access service is to offer batch data service to meet business requirements, providing basic database-based service generation, service

management, service access and service monitoring functions.

(2) Data resource authority management. Data resource authority management provides data access service authorization, authorization recovery and authority viewing functions.

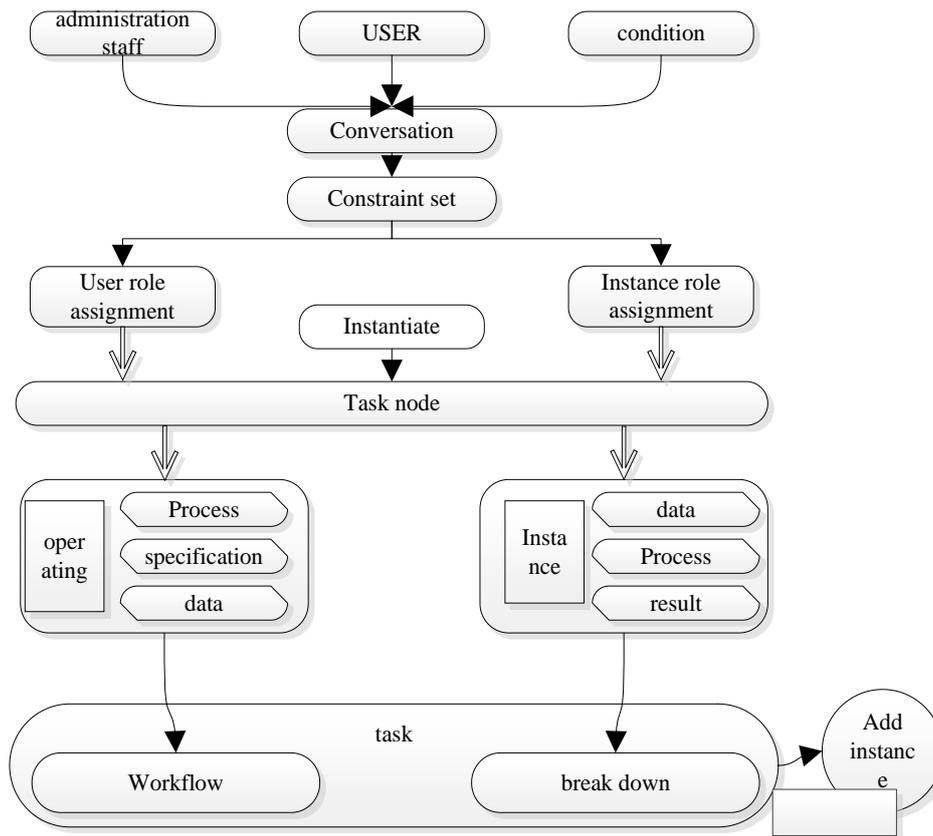
(3) ETL service management. ETL service management mainly includes conversion file management, task management, conversion monitoring and other functions, which provides data resource integration services through the ETL service management system, realizing the data exchange not only between the basic database and each business system database, but between the basic database and the subject database, and ensuring the integrity and consistency of the data. Moreover, the ETL service management system provides a graphical way to arrange data exchange rules, including data extraction definition, conversion rule definition, and data loading definition functions. Additionally, the ETL service management system realizes data exchange, process monitoring and result reporting.

#### *3.3.3 Multi-node Access Control Technology Model Based on T-RBAC*

When the business department of the administrative bureau conducts data query and access, in order to ensure the security of data information, it needs to be carried out through specific access control technology. Access control technology can prevent access to data without authorization of resources. Moreover, access control includes three elements, namely subject, object and control strategy. For administrative data access control, the subject is administrative staff, the object is administrative data, and the control strategy is multi-node access control technology based on T-RBAC.

T-RBAC-based multi-node access control technology first decomposes the user's access request into different workflows, and then assigns them to different tasks. Meanwhile, a different administrative execution role is set for each task, and the role executes query instructions by accessing the task node. Among them, the T-RBAC model is extended according to the RBAC-based access control model. In addition, the traditional RBAC model only has a three-tier structure of users, roles, and permissions, while T-RBAC adds tasks on the basis of the RBAC, expanding the three-tier structure to four layers, which assigns permissions to roles through tasks, instead of directly assigning to roles, so that dynamic allocation can be realized.

The technical model of multi-node access control based on T-RBAC is shown in Figure 2.

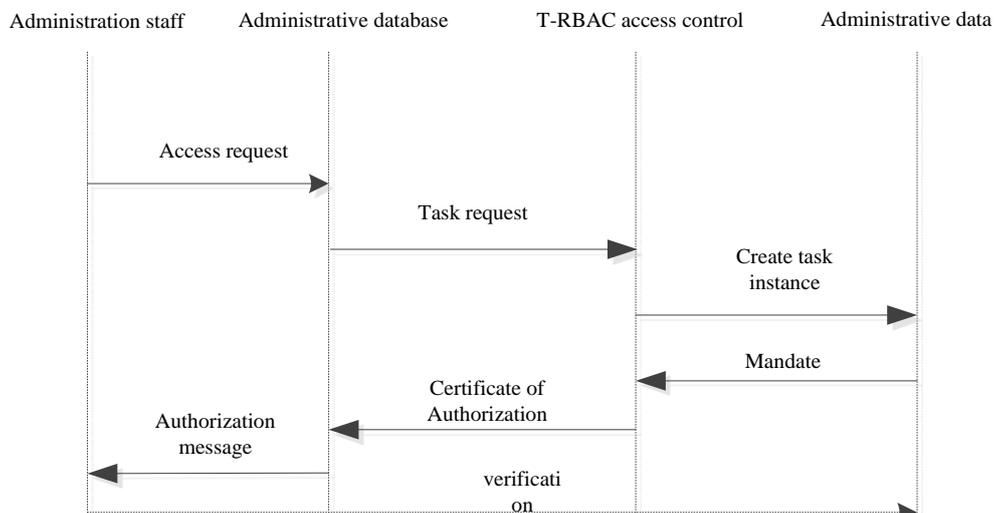


**Figure 2.** Technology model based on T-RBAC multi-node access control

The T-RBAC model defines every visit to an object as a workflow, and then divides the workflow into tasks to complete the execution of tasks by executing task instances. Meanwhile, instance permission assignment is used to assign access permissions to task instances, and user role assignments are applied to assign roles to users. Then instance role assignments are applied to execute task instances and permissions under corresponding instance roles to complete access control of administrative data. Among them, roles can be divided into four categories, namely information owner, highest visitor, internal visitor, and external shared visitor. The task security level can be divided into four security levels, namely A, B, C, and D.

A is the highest level.

When an administrator requests data, an access request will be sent to the administrative database, and the data owner will send a task request to T-RBAC access control and create a task instance. Then the authorization certificate will be returned to the owner, and the administrator's data request operation further obtains an authorization message. At this time, the administrative staff can directly send the request for administrative data. The object will complete the verification, and the verification result will be returned to realize the access control of the administrative data, which is shown in Figure 3.



**Figure 3.** T-RBAC access authorization process

According to experiments, when administrators want to obtain information about the untrustworthy information of community corrections personnel, the role is granted to internal visitors. In order to ensure the information security of community corrections personnel, when administrative personnel request access to data, the system initiates the corresponding task. The task security level is C, and the administrative personnel need to obtain authorization information through T-RBAC before querying information from the relevant database. However, when other unrelated personnel want the untrustworthy information of community corrections personnel, the role is granted to external visitors and cannot access the C tasks, so the relevant information cannot be queried.

## 4 Administrative Service Metadata Model

### 4.1 Metadata in Administrative Data Virtualization

The administrative data virtualization system itself does

not store the data source, but stores the metadata information corresponding to the physical data information, and the data location can be located through the metadata information, so that the multi-source data integration and further data operations can be realized without changing the physical location.

In a data virtualization system, metadata is the core part of the system. The administrative data in the traditional data system is stored in the designated physical space, while the data virtualization system does not store the real administrative data, but stores the metadata information corresponding to the real data. Through the metadata, the multi-source heterogeneous data information can be effectively managed in a unified manner, thereby ignoring the data differences caused by the storage process.

Metadata has five types of attributes, namely identification and definition, data collection and usage guidelines, source and reference text, relationship and management. The metadata structure is defined through these five attributes. The administrative metadata structure is shown in Table 2.

**Table 2.** Administrative metadata structure table

Attribute type	Property item	Example (case information)
Identification and definition	Chinese name	Case name
	English name	
	Abbreviation	Case code
	Synonymous name	
	Metadata type	Data element
	Registration status	Draft
	Object class	Case
	concept	Case information
	Data category	coding
	type of data	Text
	content	Case content
	length	Data length

There are two ways to obtain data in the data virtualization system. One is to cache, and the other is to directly locate the storage location of the corresponding administrative data through metadata information. Then, ODBC/JDBC, JSON, API and other interfaces are used to obtain data information. Through the metadata information in data virtualization, physical space can be greatly saved, and data sharing operations can be completed without physical integration. What is more, when performing query requests with complex levels and a large amount of tasks, metadata query can be used to find the target data more quickly and accurately, which can optimize query effect.

### 4.2 K-means clustering Model Based on Metadata

Since the ISON language is easy to read and write in various data platforms or databases, the JSON language is used to encapsulate the data source information, and then the clustering model is applied to divide the ISON documents into different clusters to achieve metadata clustering. Through hierarchical cluster query, the time required to query administrative data is shortened.

A metadata-based K-means clustering model is used in the paper. Since the JSON language is easy to read and write in various data platforms or databases, the JSON language is adopted to encapsulate the data source information, and then the clustering model is used to divide the JSON document into different clusters to realize

metadata clustering. Therefore, when querying administrative data, the first level of clustering information is first found, and then further search in it, which shortens the data query path. For example, when querying the administrative information of the defendant  $w$ , by clustering the  $id$  metadata information of  $w$ , the relevant administrative metadata information of each department of  $w$  can be directly queried. There is no need to query each administrative department, which improves the efficiency of data virtualization query to a certain extent.

(1) Standardize the JSON document format. First, set the JSON document format for the metadata. Each document is divided into 5 parts.  $id$  represents the metadata type, and  $source$  refers to the metadata source department.  $position$  indicates the source data location of the administrative data, and  $reinformation$  represents metadata-related information. The size of all parameters refers to any integer from 1 to 10. For example, for a piece of administrative office information,  $id$  is JG SFJG,  $source$  is administrative office, and  $position$  is people's mediation.  $reformation$  is administrative office-related information, and other JSON document data corresponds to the above-defined pattern, providing a standard data format for later metadata clustering.

(2) Determine the initial value of  $k$  to confirm the starting position of each center point. For large-scale data queries, the selection of the  $k$  value is more important. How to select the  $k$  value will have a greater impact on noise and outliers. Therefore, the Silhouette Coefficient is defined to determine the initial  $k$  value:

$$s_i = \frac{b_i - a_i}{\max(a_{i,j})} \quad (1)$$

Supposing that all data sets are divided into  $k$  categories, which is the average distance from the currently selected data to other data in the cluster, and  $h$  is the minimum value of the average distance from the currently selected data to the data in other clusters, when  $\&$  is a positive number, it means that the clustering of the current object is correct, which is quite different from other cluster data. Meanwhile, when  $\&$  is a negative number, it means that the clustering effect is not good, and the current data is quite different from the data of its own cluster. The choice of the number of clustering centers affects the overall clustering effect. When different values of  $k$  are calculated as  $\blacksquare^\wedge$ , the appropriate initial value of  $k$  will be selected.

(3) Calculate the position of the remaining nodes from the center point of  $k$ , and assign the result to the cluster with the smallest distance. For each sample  $\&$ , mark it as the nearest category  $j$  from the category center  $a_j$ . The formula for the distance between the remaining nodes and the center point of  $k$  is as follows:

$$Diff_y = \operatorname{arsmin} \left\{ \sqrt{\sum_{i=1}^n (x_i - a_j)^2} \right\} \quad (2)$$

In the distance calculation, the data category of the current data is obtained from the source to calculate its position from the center point in the overall XML document.

(4) Calculate the average value of the cluster again and set a new center point.

(5) Repeat the third and fourth steps to calculate the minimized square error  $E$  until the criterion function converges. The least square error formula is as follows:

$$E = \sum_{i=1}^k \sum_{k=6} \sum ||x - H_i||^2 \quad (3)$$

The main content of this chapter is the construction of administrative database, which is designed and realized from three aspects: data source and acquisition, administrative database construction plan and administrative service metadata model. First of all, the administrative structure and the data situation explain why the administrative database is built. Secondly, the data structure of administrative data is discussed to carry out basic information model construction and business data model construction, including administrative service personnel basic model, administrative service organization basic model, administrative administrative business database, personnel credit business database, and institutional credit business database. Then, metadata clustering algorithm is used to improve the efficiency of administrative data query. Finally, carry out integrated management of administrative data through integration methods such as ETL, data services, and information channels.

## 5 Administrative Data Sharing Performance Test

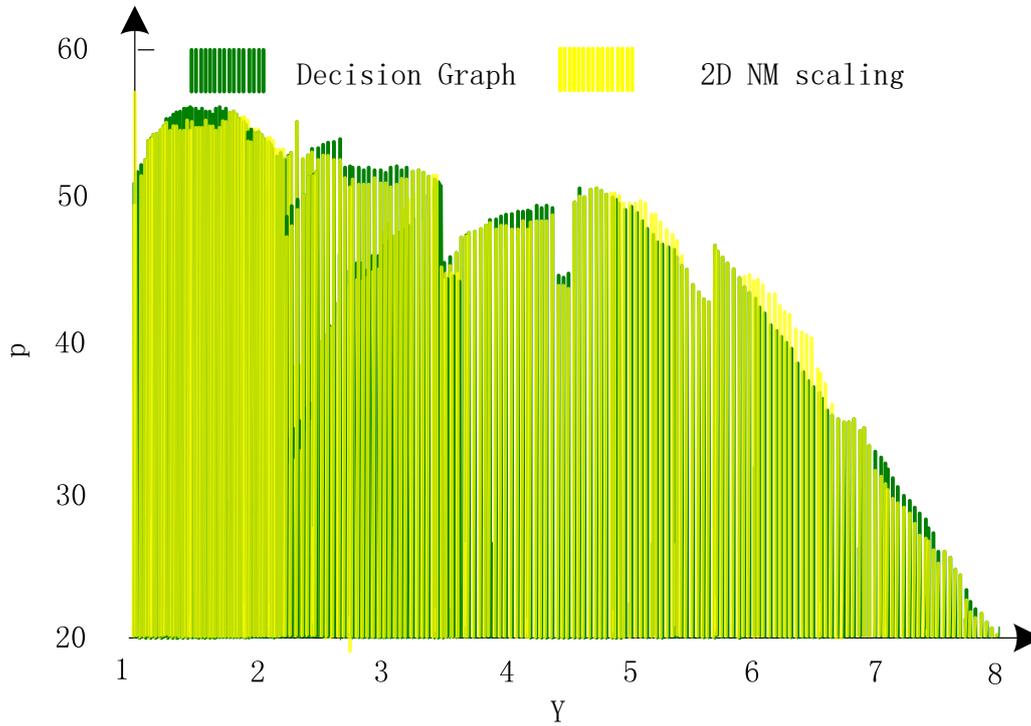
In order to build a unified administrative service website for the city, with the Beijing administrative service website as the hub, all levels of the city's administrative administration network, law popularization network, special business service website and various industry association website resources are connected the websites of administrative service agencies at all levels of the city, such as entity halls, legal aid centers, law firms, notary offices, and administrative appraisal offices, to build an administrative service platform integrating urban and rural areas and online and offline. Moreover, the system data proposed in the paper mainly includes basic information business volume, basic information database business volume, credit information business volume, theme database information business volume, exchange platform business volume and information service business volume. Among them, the basic information business volume is mainly obtained through research, and business data is generated as the business progresses. In addition, some of the accumulated resources can be entered into the basic information. The system is deployed in the government cloud, and the storage capacity is about 4.45Tb.

### 5.1 Metadata Clustering Model Efficiency Test

The K-means clustering model based on metadata is used to improve the efficiency of data query in the administrative data sharing system based on data virtualization. The model proposed in the paper reduces the scope of deep search through metadata clustering,

which can find the target query data in a shorter time. The following figure shows the clustering effect produced

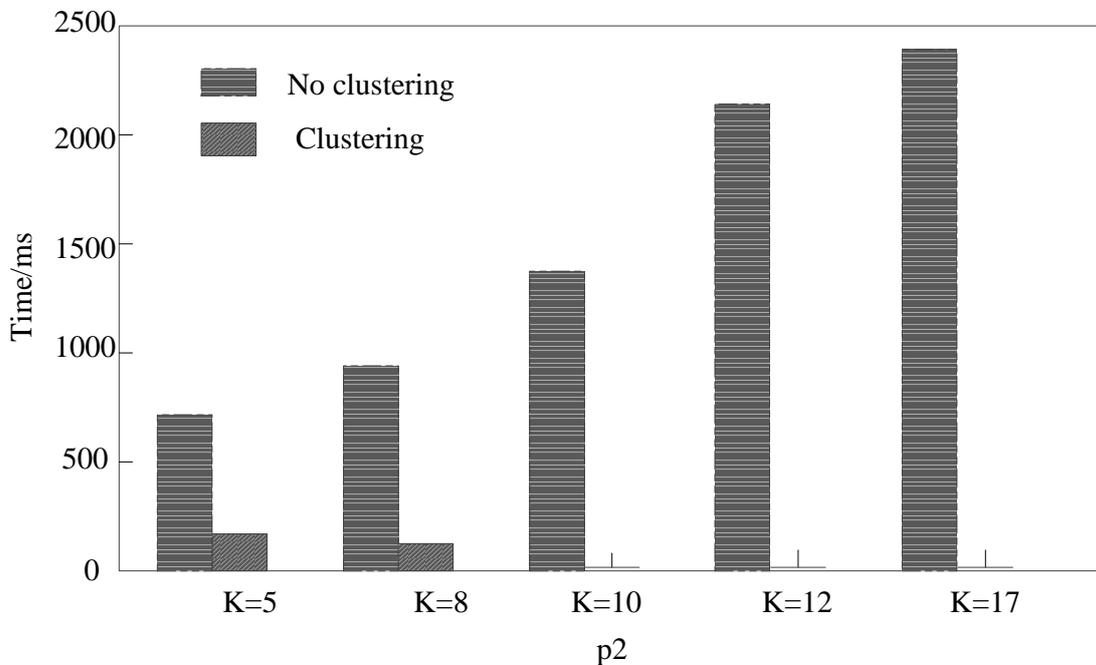
when choosing different  $k=6$  as shown in figure 4.



**Figure 4.** Clustering result graph

With the help of the clustering model of metadata, the heterogeneity of data sources is shielded to compare whether to add clustering method to the performance of

administrative data integration method. Figure 5 shows whether the clustering model time  $k$  is added under different data sets.



**Figure 5.** Comparison of administrative data query time

According to the time comparison in Figure 5, a clustering model is added to the metadata in the data virtualization layer, which shortens the data query time. Moreover, the integration method based on data

virtualization technology uses metadata to achieve efficient data access without occupying physical space, improving the performance and reliability of the overall administrative data integration.

## 5.2 Precision and Recall Test

As the business departments that need to query administrative data continue to increase, the accuracy of

query operations will be affected to a certain extent. Figure 6 shows the comparison test of precision rate between traditional data system and data system based on data virtual user.

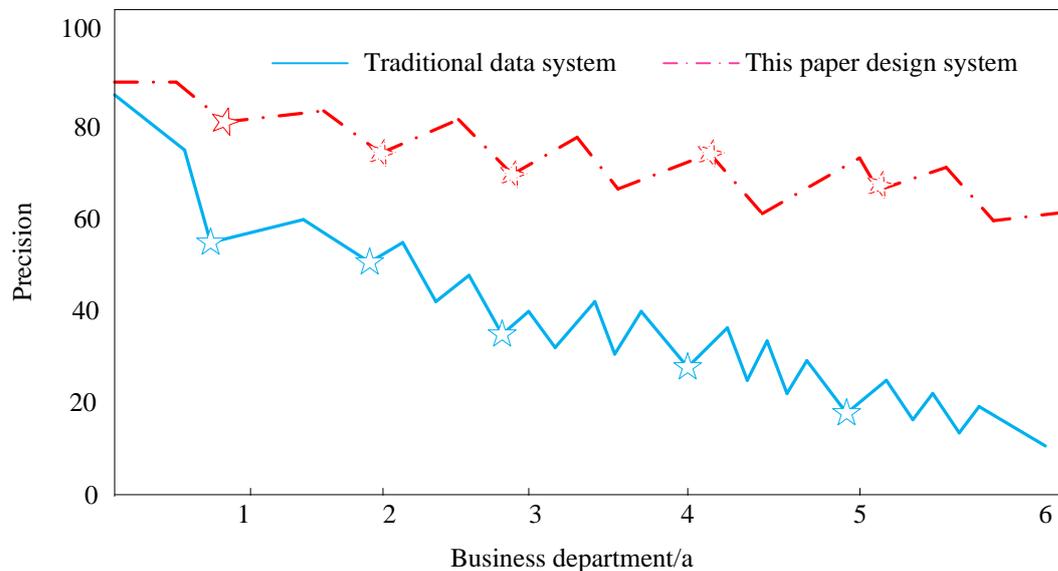


Figure 6. Accuracy comparison of the two systems

It can be seen from Figure 6 that when only one or two business department data queries are performed, the accuracy of the traditional data system and the data system based on data virtualization are almost the same. However, with the increase in the number of business departments involved, the accuracy of data systems based on data virtualization is significantly better than traditional data systems, which can achieve better query results.

## 6 Conclusion

Managing the data source based on the multi-source heterogeneous characteristics of the data source, metadata is first used in the paper to record the basic information of the data source, connection information, and characteristic information of the data source, and then the data catalog management is applied to realize the classified control of administrative data. What is more, on this basis, a virtual administrative database construction plan is designed and constructed, which is divided into basic information model construction and business database construction. Additionally, data integration technology and data virtualization are adopted for administrative data management. Meanwhile, the K-means clustering model based on metadata is used to improve the efficiency of administrative staff in querying data and shorten the access time. Finally, in order to ensure the security of administrative data, data service management and multi-node access control technology based on T-RBAC are used to achieve on-demand access by administrative personnel.

## Acknowledgement

Social Science Planning Fund Key Program in 2020, Liao

Ning Province (L20AJY001) : Research on the Countermeasures of high quality development of e-government driven by digital economy in Liaoning Province in the post epidemic Era

## References

1. Ning Jiang, Lixin Tao. A Knowledge-Based Approach to Reducing Complexity of Maintaining Semantic Constraints in Data Exchange and Integration[C]// The 5th IEEE International Conference on Cyber Security and Cloud Computing. IEEE, 2018
2. SAEED EBRAHIMI JAM, CAHIT ADAOGLU, KORHAN K. GOKMENOGLU. A Synergistic Forecasting Model for High-Frequency Foreign Exchange Data[J], Social Science Electronic Publishing
3. Kakulapati V, Pentapati V. Switch hierarchical provider (SHP) networks for data network virtualization[J]. 2018, 10(2):169-179.
4. WU Zhixue. Advances on virtualization technology of cloud computing[J]. Journal of Computer Applications[40]Khanmohammadi S, Adibeig N, Shanebandy S. An Improved Overlapping k- Means Clusterin
5. Shohei Hayashi, Shigenari Nakamura, Dilawaer Duolikun, et al. Evaluation of a Protocol to Prevent Illegal Information Flow Based on Maximal Roles in the RBAC Model[M]// Nutzwertanalysen in Marketing und Vertrieb. 2019.
6. Xinze Guo, Chuanchang Liu, Zhiyuan Su. A Novel OA System Access Control Method Based on Improved RBAC Model[C]. 2018.

7. Hayashi S, Nakamura S, Duolikun D, et al. Evaluation of a Protocol to Prevent Illegal Information Flow Based on Maximal Roles in the RBAC Model[M]// Nutzwertanalysen in Marketing und Vertrieb
8. Liao H.C, Gou X.J, Xu Z.S, Zeng X.J, F. Herrera. Hesitancy degree-based correlation measures for hesitant fuzzy linguistic term sets and their applications in multiple criteria decision making[J]. *Information Sciences*, 2020, 508: 275-292.
9. Liu H.B, Ma Y, Jiang L. Managing incomplete preferences and consistency improvement in hesitant fuzzy linguistic preference relations with applications in group decision making[J]. *Information Fusion*, 2019, 51: 19-29.
10. Li C.C, R.M. Rodríguez, L. Martínez, Y.C. Donga, F. Herrera. Personalized individual semantics based on consistency in hesitant linguistic group decision making with comparative linguistic expressions[J]. *Knowledge-Based Systems*, 2018, 145: 156-165.
11. Zhang R.C, Li Z.M, Liao H.C. Multiple-attribute decision-making method based on the correlation coefficient between dual hesitant fuzzy linguistic term sets[J]. *Knowledge-Based Systems*, 2018, 159: 186-192.
12. R.M. Rodríguez, Á. Labell, L. Martínez. An Overview on Fuzzy Modeling of Complex Linguistic Preferences in Decision Making[J]. *International Journal of Computational Intelligence Systems*, 2016, 9: 81-94.