Data Middle Platform Construction:
the Strategy and Practice of National Bureau of Statistics of China

Zhang Chun-zhen[1]

[1] Data Management Center of NBS, Beijing, China, zhangcz@stats.gov.cn

Abstract:

Currently in the context of Big Data, as the traditional statistical information systems are designed with software (application) as the center, statistical data faces serious islandization problem. To meet this challenge, taking the opportunity of the Statistical Cloud construction, the National Bureau of Statistics of China adopts the concept of "data middle platform" to make data resource planning, and aims to build a comprehensive data capability platform that includes data collection and exchange, data sharing and integrating, data organizing and processing, data modelling and analysing, data management and governance, data service and application. The statistical data middle platform provides the basic capability for data application support, enables data to form a benign closed loop between data platform and business system, and in the end enables internal and social-oriented servitization of statistical data. As a new exploration and attempt, the statistical data middle platform will not only solve the long-standing data island problem of NBS, but also provides a basic guarantee for greater use of data potential, helping official statistics to transform from statistical analysis to predictive analysis, from single-domain to cross-domain, from passive analysis to active analysis, and from non-real-time to real-time analysis.

Keywords:

Data Island; Data Middle Platform; Data Servitization; Statistical Cloud; Data Ecology

1. Introduction:

Data island problem has a long history. The root cause of data island is that traditional information systems are designed with software application as the center. The problem arises because, organizationally, data belongs to different departments and operates independently; physically, data is stored independently, maintained independently, and isolated from each other; and in the end, logically, data is loosely connected or unconnected and cannot be integrated. With the advent of the big data era, traditional industries are deeply integrated with new technologies such as cloud computing, Internet of Things (IoT) and Artificial Intelligence (AI), and multi-source and multi-type big data is happening all the time. In this context, data supply is unbalanced with respect to the demand and output of data, and the problem of data island is becoming more prominent.

In IT industry, two main directions to solve the data island problem corresponding to its different causes are adopted. The first one is a top-down and business-driven direction, which starts from the upper level of the application system, considers the data dependence among the application systems, and develops interface services on demand, so as to achieve the mutual use and integration of the data among the respective systems; The second one is a bottom-up and data-driven direction, which gives the priority to considering global data characteristics, formulating data specifications, defining data standards, unifying the cognition of data among different departments, and thus realizing integration between upper-layer applications and wider external services of data with the integration of lower-layer data.

In practice, Enterprise Service Bus (ESB) is a typical example for the top-down direction, which adapts various heterogeneous systems through the service bus. In contrast, the practice of the bottom-up direction is relatively more adopted in IT industry as it takes a global perspective to conduct data governance in the true sense and correspondingly has a better effect. For instance, the results of Harbor Research in the United States believe that if the traditional application-centric data organization method is changed to an information-centric distributed structure, the problem of data island will no longer exist; The data sharing mechanism based on blockchain technology, as well as the "data middle platform" concept popular in the industry in recent years, are essentially bottom-up data island solutions; in addition, for data island elimination and data interoperability requirements across sectors, industries, and even entire domains, a wider range of internationally unified data specifications and standards should be established (both grammatical, such as SDMX, and semantic, such as Lined Open Data) to finally realize the vision of the Web of Data.

Currently, the National Bureau of Statistics of China (NBS) is implementing the construction of Statistical Cloud, a systematic project that will fundamentally change the IT operation mode as well as the statistical business of NBS. Taking this opportunity and constructing an important part of Statistical Cloud, the NBS adopts the concept of "data middle platform" to make data resource planning, and aims to build a comprehensive data capability platform. At present, the statistical data middle platform has already finished its general design. As a new exploration and attempt, the statistical data middle platform will not only solve the long-standing data island problem of NBS, but also provides a basic guarantee for greater use of data potential. In this study, we will introduce the methodology for middle platform and investigate the architecture and implementation method of the statistical data middle platform in application.

2. Methodology:

2.1 Middle platform

Alibaba[1] has positioned the middle platform as a support platform that provides agile response to front-end applications in the form of reuse capabilities. Wang of Thoughtworks[2] defined the middle platform as an "enterprise-level capability reuse platform", and Chen et al. (2019) described middle platform as an enterprise-level shared service platform. Combining various opinions, this paper argues that the so-called "middle platform" should have the following characteristics:

First, middle platform solves the problem of reinventing the wheel, and its core value is reuse. When developing new front-end softwares, people can directly use the finished components provided by the middle platform, which greatly improves the efficiency of software development and avoids repeated construction.

Second, middle platform is born for the front platform. On the one hand, the front platform is the driving force for the generation of the middle platform, and on the other hand, it is also the foundation of the precipitation-type construction approach for the middle platform

Third, the middle platform capability is decoupled natively. In the process of generating the middle platform, the cohesion of the middle platform capability is improved by comprehensively considering the business characteristics of the system to determine the division boundary and granularity.

Fourth, the capabilities of the middle platform are stable and will not change frequently, which is the basis for the middle platform to play the one-to-many service sharing ability, and embodies the value of the middle platform.

Fifth, the capabilities of the middle platform are evolving and have their own life cycles. With the continuous access of new services, shared services have continuously adapted to various business processes in their self-evolution, and have truly become valuable IT assets for enterprises.

According to different "capability" categories, the middle platform can be logically divided into technical middle platform, business middle platform, and data middle platform etc.[3]

2.2 Data middle platform

[1] https://www.alibabagroup.com/

[2] https://www.thoughtworks.com/cn/

[3] Generally, the classification of the middle platform is based on logical division, and the boundary is not strict. With the continuous expansion of the middle platform capacity category, the classifications of the middle platform have also enlarged, including for example, algorithm middle platform, organization middle platform, mobile middle platform etc.

The data middle platform focuses on the reuse of data-related capabilities, and meanwhile as an important branch of the middle platform, it naturally inherits the characteristics of the middle platform described above. In this paper, we define the data middle platform as an enterprise-level comprehensive data capability platform, which includes data collection and exchange, data sharing and integrating, data organizing and processing, data modelling and analysing, data management and governance, data service and application, fundamentally breaks through the technical barriers of data production, storage, analysis, service, and circulation, and is a global bottom-up solution for the data island problem in enterprises.

It is a process of two-way selection and adaptation for enterprises to apply to data middle platform. As an emerging concept, the data middle platform has no unified methodology for its construction mode. However, after the rapid development and continuous exploration in recent years, IT industry has initially formed some universal common cognition, which has become the guiding ideology of data middle platform construction.

Cognition 1: The basic premise for an enterprise to build a data middle platform is that it has a certain scale of informatizable business, the stock or the expected incremental business is diversified, and the business is coupled.

Cognition 2: Data middle platform is a strategic choice for enterprise development. The data middle platform is not a short-term behaviour, but for the overall planning and long-term development of the enterprise.

Cognition 3: Data middle platform is a necessary condition for enterprise innovation.

3. Result:

The problem of data island in the statistical field has brought many annoyances to business users, mainly manifested in the following aspects: on one hand, horizontally, an enterprise have too many information systems with redundant coupling between each other, the data is distributed and stored in different types and versions of databases according to the statistics discipline, survey type, survey year and other dimensions, and the specifications among the data are not uniform. Therefore, data sharing and analysis are only driven by a single specific business, it is difficult to carry out value mining on global data, and the value of data assets cannot be reflected in terms of scale and effect; on the other hand, vertically, within the information system, application and data are tightly coupled, thus data readability and availability are heavily dependent on the business system, and data cannot be autonomous. As a result, faced with the need to change software for new businesses, decision-making is often difficult, which delays the delivery and so it is incapable of responding quickly to statistical surveys.

In response to the above problems, NBS has proposed a complete set of solutions based on the concept of data middle platform. The core points are as follows:

3.1 Construction basis

Corresponding to the above common cognitions, data middle platform construction consists of the following bases: Firstly, regardless of classified as various types, statistical application softwares share many similarities when analysing them according to the "statistical data production process": the first is that they are all based on unified metadata, regime design, report design, user and authority management, etc.; the second is that they use the same source roster, survey objects, and the same norms and standards for sampling; the third is that the processes of data collection, examination, summary, and acceptance in data processing are mostly similar; the fourth is that data processing and publishing, archiving management, and further in-depth analysis also follow consistent management standards. All these similarities run through the main line of business dataization, and provide a foundation for breaking through business and data barriers to solve data island, which satisfies the prerequisites for the construction of the data middle platform in Cognition 1 above.

Secondly, it is based on the development of Statistical Cloud. In the 13th Five-Year Plan for Statistics, it is clearly proposed to establish a statistical cloud. Statistical Cloud uses the achievements of the Internet, big data, cloud computing, artificial intelligence, and spatial geographic information technology to promote the in-depth integration of modern information technology and statistical business, transform statistical production methods, improve government statistical capabilities and credibility, and comprehensively promote the modernization of the national statistical system and statistical capabilities. As a strategic project of national statistics, Statistical Cloud provides a good opportunity for overall planning and governance of statistical data resources, which satisfies the above Cognition2.

In addition, Statistical Cloud sets up the "three onto-cloud" goals, that is, to promote statistical business onto the cloud and realize centralized and unified management of statistical business application systems; to promote data onto the cloud to achieve centralized and unified management of statistical data resources; to promote management onto the cloud to achieve unified management of information resources, centralized business scheduling, and provide users with a unified service interface. Statistical Cloud aims to create a “cloud statistics” to deeply change the statistical production method and innovate the way statistical data serves government decision-making and social governance, which satisfies Cognition 3 above.

3.2 Overall architecture

The statistical middle platform implements the idea of "large middle platform, small front end", and is located in the middle layer of the cloud application system. The statistical data middle platform adopts the industry-leading shared architecture design and horizontal layered construction ideas to build a unified cloud-based and service-oriented basic data platform. The overall functional architecture is shown in Figure 1.

3.3 Core capabilities

The statistical data middle platform has six core capabilities as follows:

(1) Collection and exchange. Data is the carrier of information and the destination of all business datalization. As for the function of the data middle platform, first of all, it should have a powerful collection and exchange capability before helping complete the original accumulation of data and reach the big data scale in aspects of both data source and data content. Here we divide the collection and exchange into two different methods: collection is oriented to the data of the enterprise itself that reflects business characteristics; exchange is oriented to the data of external sources. The collected and exchanged data exists in the middle platform in the form of operational data store (ODS).

(2) Aggregation and integration. The original data that enters the middle platform after collecting and exchanging process may come from different types of databases such as Oracle, MySql, SQL server, MangoDb, etc., and show different storage types such as structured, semi-structured, and unstructured. The data is isolated and scattered, and has no uniform standard, thus it is not yet available. The data middle platform should establish unified metadata and master data standards, and perform operations such as extraction, mapping, transformation, and verification on the original data, and ultimately form a unified library with unified standard and logical concentration to achieve basic data availability.

(3) Organization and processing. The organization and processing of data is a more fine-grained process oriented to themes and special topics. After organization and processing, the data forms corresponding thematic bases and special-topic bases, and provides data support for modelling and analysis in the form of data marts. After organized and processed, the data achieves the easy-to-use value and has realized assetization.

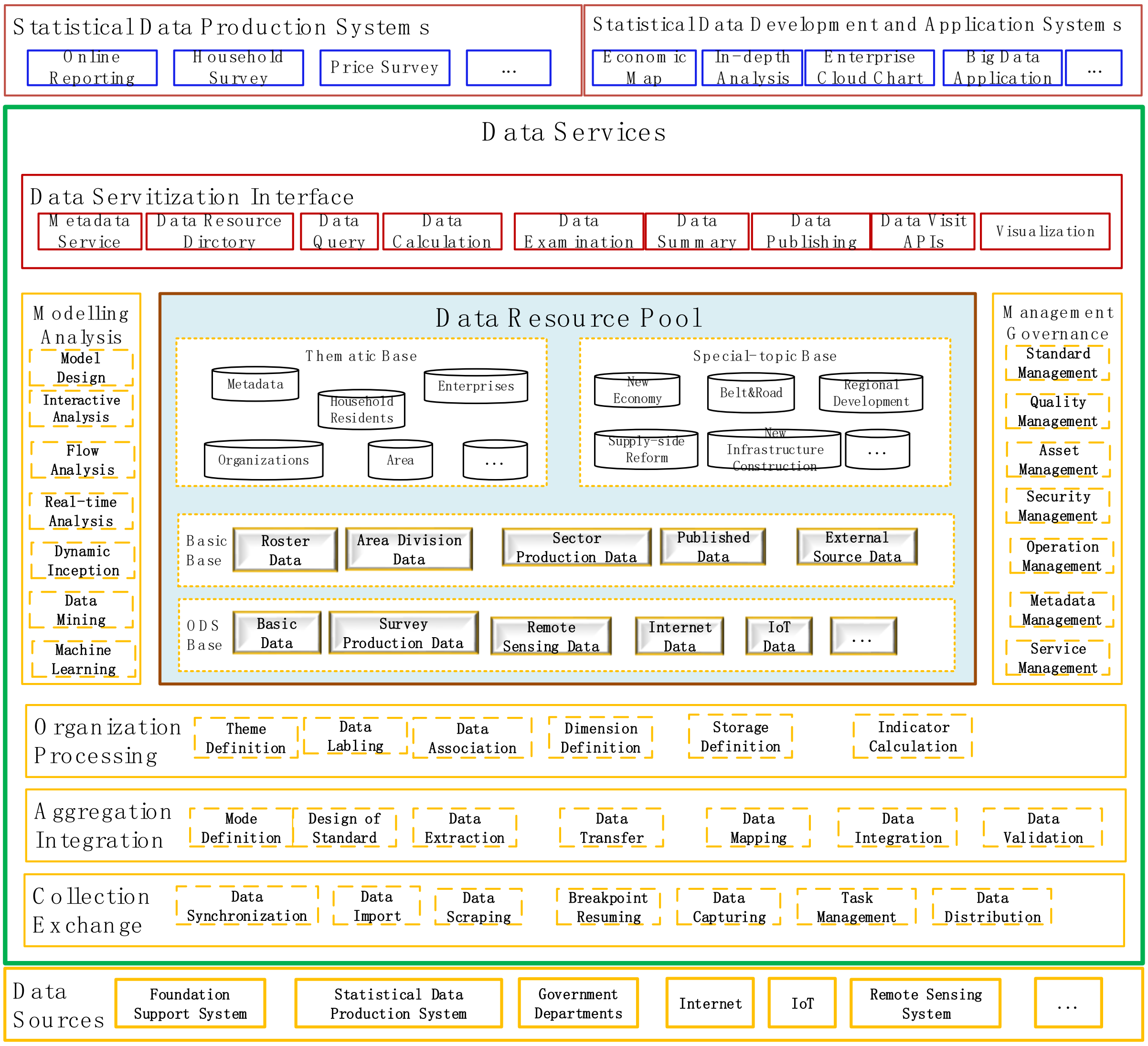

Figure 1. Overall architecture of the statistical data middle platform

(4) Management and governance. Platform-level management and governance capabilities are the foundation for external empowerment of data middle platform. The management and governance capabilities that the data middle platform provides specifically include metadata management, unified business dictionary management, data lifecycle management, data quality management, security management, platform operation and maintenance management, and service management, etc., which helps build asset-based, service-oriented, and standardized data systems.

(5) Modelling and analysis. Modelling and analysis is an important carrier of the output of the data middle platform's capacities, and it is also a direct expression of the value materialization of the data middle platform. Modelling and analysis includes traditional OLAP-type data analysis, data statistics, and data mining, and with the continuous addition of new modelling and analysis methods such as streaming analysis, real-time analysis, dynamic perception, machine learning, etc., the modelling and analysis capabilities of the data middle platform are also continuously enhanced in depth and breadth.

(6) Service and application. Based on its accumulated big data assets and powerful storage, calculation, fusion, and processing capabilities, the data middle platform provides accurate services for front-end applications in an intelligent and visual way, allowing data to be used and flowed, which provides an important means of external empowerment of the data middle platform, and reflects the ability of the data middle platform to feed the business system in a service mode. Data forms a closed loop between the data middle platform and the business system through the form of service and

application, which can promote the continuous iterative upgrade of the business system and the data middle platform itself.

3.4 Operation carrier

The architecture of data middle platform provides a static mechanism based on the reuse ability and external empowerment in the form of service. To make this mechanism land on the ground and play its role, the cloud business platform requires a suitable operation carrier.

The Statistical Cloud adopts the "container + microservices" technical route, and uses microservices as the operation carrier for the platform. As the middle platform emphasizes the core reuse basic capabilities, the basic capabilities should take the smallest service as a unit, with high cohesion and low coupling, and support rapid iteration and innovation of various scenarios on the business side through the arrangement and coordination of service units. The microservice architecture splits the monolithic application into multiple small services with high cohesion and low coupling according to the business field. Each small service runs in an independent process and is developed and maintained by different teams. Lightweight communication mechanisms (such as HTTP RESTful API, or RPC) are used between microservices, which can be deployed independently and automatically, and can use different protocol stacks, languages, and storage. Microservices embody decentralization and natural distribution. These characteristics of microservices and the idea of service self-closing loop that it advocates make it a suitable architecture for the implementation of the platform.

4. Discussion and Conclusion:

IDC predicts in its white paper that by 2022, more than 60% of global GDP will be digitally replaced. Data is increasingly valued and recognized by the whole society. Taking the opportunity of constructing Statistical Cloud, the NBS adopts the concept of "data middle platform" to make data resource planning and build a comprehensive data capability platform. As a new exploration and attempt, the data middle platform can fundamentally break through the technical barriers of data production, storage, analysis, service and circulation, and provide a global bottom-up solution for the data island problem in enterprises.

In addition, apart from solving the problem of the data island problem of NBS, it also has reference significance for data governance in other industries. By solving data islands, tolerating the entry of various data sources, and building statistical big data platforms, government statistics can use data servitization to empower data partners, and cultivate healthier and more dynamic statistical data ecology, so as to integrate into the Web of Data.

References:

1. Chen, X., et al. (2019). *Middle-Platform Strategy: Middle-Platform Construction and Digital Business*. China Machine Press: Beijing.

2. China Institute of Information and Communication (2019). *Data Infrastructure White Paper 2019*.

3. https://en.wikipedia.org/wiki/Semantic_Web

4. https: //www.chaoqi. net/ganhuo/2019/0405/192928.html

5. Ontotext (2019).*What are Linked Data and Linked Open Data?*