Statistical Journal of the IAOS 36 (2020) 979–986
DOI 10.3233/SJI-200754
IOS Press

Data middle platform construction: The strategy and practice of National Bureau of Statistics of China

Chunzhen Zhang[a] and Lei Hou[b,*]
[a]*Data Management Center, National Bureau of Statistics of China, Beijing 100826, China*
[b]*Institute of World Economics and Politics, Chinese Academy of Social Sciences, Beijing 100732, China*

Abstract. To address the data 'islandization' issue in the statistical field and to take advantage of the opportunity of the Statistical Cloud construction, the National Bureau of Statistics of China (NBS) started adopting the concept of a "data middle platform" for data resource planning. With it, NBS aims to build a comprehensive data capability platform that includes data collection and exchange; data sharing and integration; data organizing and processing; data modeling and analyses; data management and governance; and data service and application. The statistical data middle platform provides the basic capability for data application support. It also enables data to form a closed loop between the data middle platform and the business system, and eventually realizes the 'servitization' of statistical data that meets internal and societal requirements. As a new innovative development, the statistical data middle platform will not only solve the long-standing data island problem of NBS but will also provide a basic guarantee for greater use of the data potential, and thus will help 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.
The paper was prepared under the kind mentorship of Ronald Jansen, Assistant Director and Chief of Data Innovation at the UN Statistics Division in New York.

Keywords: Data islandization, data middle platform, data servitization, statistical cloud, data ecology, Asia-Pacific statistics week

1. Introduction

The information technology (IT) is widely used in all walks of life and has become an important engine of economic growth and social development. Through software applications, the information technology digitalizes business, transforms business into data, and then excavates value in the data, which in turn drives business to expand in width and depth. With the explosive construction of various IT systems, problems have emerged in the accumulation and effective utilization of data. Data islandisation is one of the most typical problems and it eventually causes other problems.

1.1. Data islandization

Data islandization refers to such a status that data are distributed separately like islands and meanwhile they become more and more loosely connected with each other. A data island, somewhat similar to an information silo [1], is a situation wherein only one group in an organization can access a set or source of data. Consequently, the value of the data is greatly weakened. In the current context of big data, the problem of data islandization is becoming more prominent.

The data island problem originates from the technical characteristics of traditional information systems. Generally, traditional information systems are designed around a specific software application [2], which results in that (1) organizationally, data belong to different organizations or entities that operate independently;

*Corresponding author: Lei Hou, Institute of World Economics and Politics, Chinese Academy of Social Sciences, Beijing 100732, China. Tel.: +86 10 85195354; E-mail: houlei@cass.org.cn.

(2) physically, data are stored and maintained independently and isolated from each other; (3) logically, available data are only loosely or even not connected. With the advent of the big data era, multi-source and multi-type big data are being produced all the time. Thus, the data scale has reached an unprecedented level in terms of quantity and type. Though meanwhile, the value mining from the data is still insufficient due to the loose connection between the data, even with the help of advanced computing technologies, such as cloud computing, Internet of Things (IoT), Artificial Intelligence (AI), and the likes. Under this background, the supply of data is unbalanced with respect to the demand and the output of data, and the contradiction between increasing data supply and loose data connection is magnified.

1.2. Approaches to solve data islandization

In the IT industry, two main approaches are adopted to solve the data island problem. The first approach is top-down and business-driven. This approach starts from the upper layer of the application system. It considers the data dependence among application systems and develops interface services on demand to achieve the mutual use and integration of data among respective systems. The second approach is bottom-up and data-driven. This approach gives priority to considering global characteristics of the data. It formulates data specifications, defines data standards, unifies the cognition of data among different organizations, and thus realizes integration between upper-layer applications and wider external services of data with the integration of lower-layer data.

The enterprise service bus [3] is a typical example for the top-down approach as it adapts various heterogeneous systems through the service bus. On the other hand, the practice of the bottom-up approach is relatively more common in the IT industry as it takes a global perspective to conduct data governance in the true sense and correspondingly, leads to a better result in the quality of data integration. For instance, a research conducted by Harbor Research in the United States argued that if the traditional application-centric data organization method is changed into an information-centric distributed structure, the problem of data islands will no longer exist [4]. The data sharing mechanism based on block-chain technology [5], as well as the "data middle platform" concept popular in IT industry in recent years, is essentially a bottom-up solution to data islands problem.

Another example for the bottom-up approach is from the World Wide Web Consortium (W3C). W3C is committed to building a Semantic Web (also known as the Web of Data), aiming to make Internet data machine-readable. For the strategy of strengthening data interoperability across sectors, industries, and even the entire Internet, W3C advocates that a wider range of internationally unified data specifications and standards should be established both in grammatical aspect, such as Statistical Data and Metadata eXchange (SDMX, https://sdmx.org/), and in semantic aspect, such as Linked Open Data [6]. These measures also help to fundamentally solve the problem of data islands, and finally realize the vision of the Web of Data [7].

Currently, NBS is building a Statistical Cloud which is a systematic project that will substantially change the IT operation mode as well as the statistical business of NBS. As an important part of the Statistical Cloud construction, NBS adopts a concept of "data middle platform" for data resource planning, and aims to build a comprehensive data capability platform. In this study, we examine the strategy and practice of a data middle platform construction taking the NBS project as an example. We first propose three common facts of a data middle platform construction, and based on these facts, explore the application of the data middle platform to the statistical field. Specifically, we attempt to design the overall functional architecture of the statistical data middle platform, analyze its core capabilities, and further investigate its technical route and construction method.

2. Historical background and methodology

'Middle platform' is not a new term. Many traditional industries are practicing a middle platform model, of which the banking industry is a most typical example of a middle platform application. In a bank, the front-end platform (office) directly faces customers and the back-end platform (office) mainly deals with transaction processing and business support. Meanwhile, the middle platform (office) utilizes resources from both the front-end and the back-end platforms to provide professional management and guidance such as risk management, product development, marketing channel management, human resource management, strategic planning, and risk control.

In comparison, the middle platform has become popular in IT industry only in recent years, and there have been until now few studies on this topic in either industry or academia. In reality, the prototype of a middle platform in IT industry originates from Supercell a

Finnish mobile game software development enterprise. Supercell adopts a "least-powerful CEO" idea for enterprise management and innovates its business based on a "Super platform" plus "small Cell team" (consisting of only 5–7 people) organization structure [8]. Supercell separates the relatively stable public and common parts of the game development process (such as infrastructure, game materials, algorithm engines, development tools, etc.) from the business, and organizes them in a supporting platform for integration. After such an improvement, the development teams in the front-end platform can develop new games based on the supporting platform, which dramatically shortens the development cycle by re-using existing components, and thus makes it possible to quickly respond to the ever-changing user tastes and occupy the market.

2.1. Middle platform

Inspired by Supercell's organizational structure and development model, Alibaba (https://www.alibaba.com/) took the lead in proposing the concept of a middle platform in the IT industry and raised it to a strategic level. Alibaba has positioned the middle platform as a support platform providing an agile response to front-end applications in the form of a set of reusable capabilities [9]. Jian Wang of Thoughtworks defined the middle platform as an "enterprise-level capability reuse platform" [10], and Chen et al. described a middle platform as an enterprise-level shared service platform [11].

Referring to various opinions, this paper argues that a so-called "middle platform" should have the following five characteristics.

1. The middle platform should solve the problem of "reinventing the wheel", and its core value is therefore anchored on its ability to be reused. When developing new front-end software, people can directly use the finished components provided by the middle platform, which will then greatly improve the efficiency of software development and avoid repeated construction.
2. The middle platform should be created for the front-end platform, that is, the natural mission of middle platform is to serve the front-end problem. On one hand, the front-end platform is the driving force for the generation of the middle platform; and on the other hand, the front-end platform also provides the foundation of the precipitation-style construction approach for the middle platform.
3. The capabilities of a middle platform should be natively decoupled as well as highly cohesive. The generation of a middle platform is a process of decomposing various software systems into software components, which serve as autonomous capabilities. Thus, the components in the middle platform are natively decoupled and independent from each other. Meanwhile, the components are divided according to the business characteristics. The business characteristics confine the boundary between capabilities, and correspondingly, determine the granularity of capabilities. Hence, the various capabilities are highly cohesive as well.
4. The capabilities of a middle platform should be relatively stable and not change frequently. Only with this stability can they serve as a strong basis for the middle platform and make it possible for the middle platform to act as a one-to-many service sharing entity.
5. The capabilities of a middle platform should be evolving and having 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 their capabilities, middle platforms can be logically divided into various types such as a technical middle platform, a business middle platform, and a data middle platform etc. Generally, the classification of the middle platform is based on logical division, and there are no strict boundaries. With the continuous expansion of their capabilities, the classifications of middle platforms have been extended, including for example, algorithm middle platform, organization middle platform, and mobile middle platform as well.

2.2. Data middle platform

The data middle platform focuses on the reuse of data-related capabilities. As an important branch of the middle platform, the data middle platform naturally inherits the characteristics of the middle platform described in preceding subsection. In this paper, we define a data middle platform as an enterprise-level platform of comprehensive data capability, which includes data collection and exchange, data sharing and integrating, data organizing and processing, data modeling and analyzing, data management and governance, data service and application. The data middle platform can fundamentally break down 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.

In practice, the application of a data middle platform solution is based on at least two considerations. On one hand, a data middle platform solution is not suitable for all enterprises. On the other hand, not all enterprises can meet the prerequisites and have the necessity to construct a data middle platform. An enterprise can maximize the benefit from a data middle platform only after it has completely adapted to and fully made use of the data middle platform. As for the construction of a data middle platform, the IT industry has not yet reached a unified methodology. However, after the rapid development and continuous exploration in recent years, the IT industry has initially formed some universal common facts, which we summarize as the guiding ideology of a data middle platform construction.

Fact 1: Basic premises for an enterprise to build a data middle platform is that (1) it has a certain scale of 'informatizable' (or digitalized) business; (2) the stocks or the expected incremental businesses are diversified, and (3) there exist reusable contents between different businesses, i.e., the businesses are coupled with each other.

Fact 2: The construction of a data middle platform is a strategic choice for enterprise development. The data middle platform is not a short-term solution, but is an element of the overall planning and long-term development of the enterprise.

Fact 3: The construction of a data middle platform is a necessary condition for enterprise innovation. By using the reusable components of a data middle platform, it is as simple as building blocks for an enterprise to develop a new software, because it no longer needs to build many things from scratch. Therefore, enterprises can quickly materialize their innovative ideas into specific softwares. As the trial and error cost of this mode of software development is relatively lower than the traditional ones, enterprises can repeat the iterative upgrade for multiple rounds, to finally realize the innovation.

3. Results

The problem of data islands in the statistical field has brought many annoyances to statistical business users, mainly manifested in two aspects. Horizontally, a statistical business sector has too many information systems with redundant overlap between each other, the data are distributed and stored in different types and versions of databases according to various statistical sections, survey types, survey years and other dimensions, and the specifications among the data are not uniform. As a result, data sharing and analysis are only driven by a single specific business. Therefore, 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. Vertically, within the information system, the software application and the data are tightly coupled. Hence, data readability and availability are heavily dependent on the business system, and data cannot be autonomous. Faced with the need to change software for new businesses, decision-making is often difficult, which will thus delay delivery and result in an incapability to respond quickly to statistical surveys. In response to the above problems, NBS proposed a complete set of solutions based on the concept of a data middle platform, which includes the following core contents.

3.1. Construction basis

Corresponding to the above common facts, the construction of the statistical data middle platform consists of the following bases:

Firstly, though classified into various types, statistical application softwares share many similarities when being analyzed according to the "statistical data production process". First, they are all based on unified metadata, method design, report design, user and authority management, and the likes. Second, they use the same source rosters, survey objects, and the same norms and standards for sampling. Third, the processes of data collection, evaluation, summary, and verification are mostly similar. Fourth, 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 'datalization' (that is, transforming business into data object, and driving business innovation and development with data), and provide a foundation for breaking down business and data barriers to eliminate data islands, which satisfies the prerequisites for the construction of the data middle platform in Fact 1.

Secondly, the construction of the statistical data middle platform is based on the development of the Statistical Cloud. In the Outline of the 13th Five-Year Plan for Statistical Information Construction in China, it is clearly proposed to establish a statistical

cloud. The 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 the national statistical system as a whole, the Statistical Cloud provides a good opportunity for overall planning and governance of statistical data resources, which satisfies the prerequisites for the construction of the data middle platform in Fact 2.

Thirdly, the Statistical Cloud sets up a "three onto-clouds" goal, i.e. to promote statistical business onto the cloud for a centralized and unified management of statistical business application systems; to promote data onto the cloud to achieve a centralized and unified management of statistical data resources; to promote management onto the cloud to achieve a unified management of information resources, centralized business scheduling, and provide users with a unified service interface. The Statistical Cloud aims to create "cloud statistics" and thus to deeply change the statistical production method and innovate the way statistical data serve government decision-making and social governance, which satisfy the prerequisites for the construction of the data middle platform in Fact 3.

3.2. Overall functional architecture

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

3.3. Core capabilities

According to the function orientation and design intention, as show in Fig. 1, we define the following six core capabilities for the statistical data middle platform. The definitions and their specific meanings are as follows:

1. Collection and exchange. Data is the carrier of information and the destination of all business 'datalization'. As for 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 the aspects of both data source and data content. Here collection and exchange are two different methods: Collection is oriented to the data within NBS that reflects business characteristics; In comparison, exchange is oriented to the data outside NBS, that is, the data from external sources. The collected and exchanged data exist in the data middle platform in the form of operational data store (ODS) [12].
2. Aggregation and integration. The original data that enter the data middle platform after collecting and exchanging 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. Data are isolated and scattered, and have no uniform standard, thus they are not yet available. The data middle platform should establish unified metadata and business data standards, 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 middle platform forms corresponding thematic bases and special topic bases, and provides support for data modeling and analysis in the form of data marts. After being organized and processed, the data achieve the easy-to-use value and realize 'assetization' (data is transformed from common digits to asset).
4. Management and governance. Platform-level management and governance capabilities are the foundations for the external empowerment of a data middle platform. The management and governance capabilities that a 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, service manage-

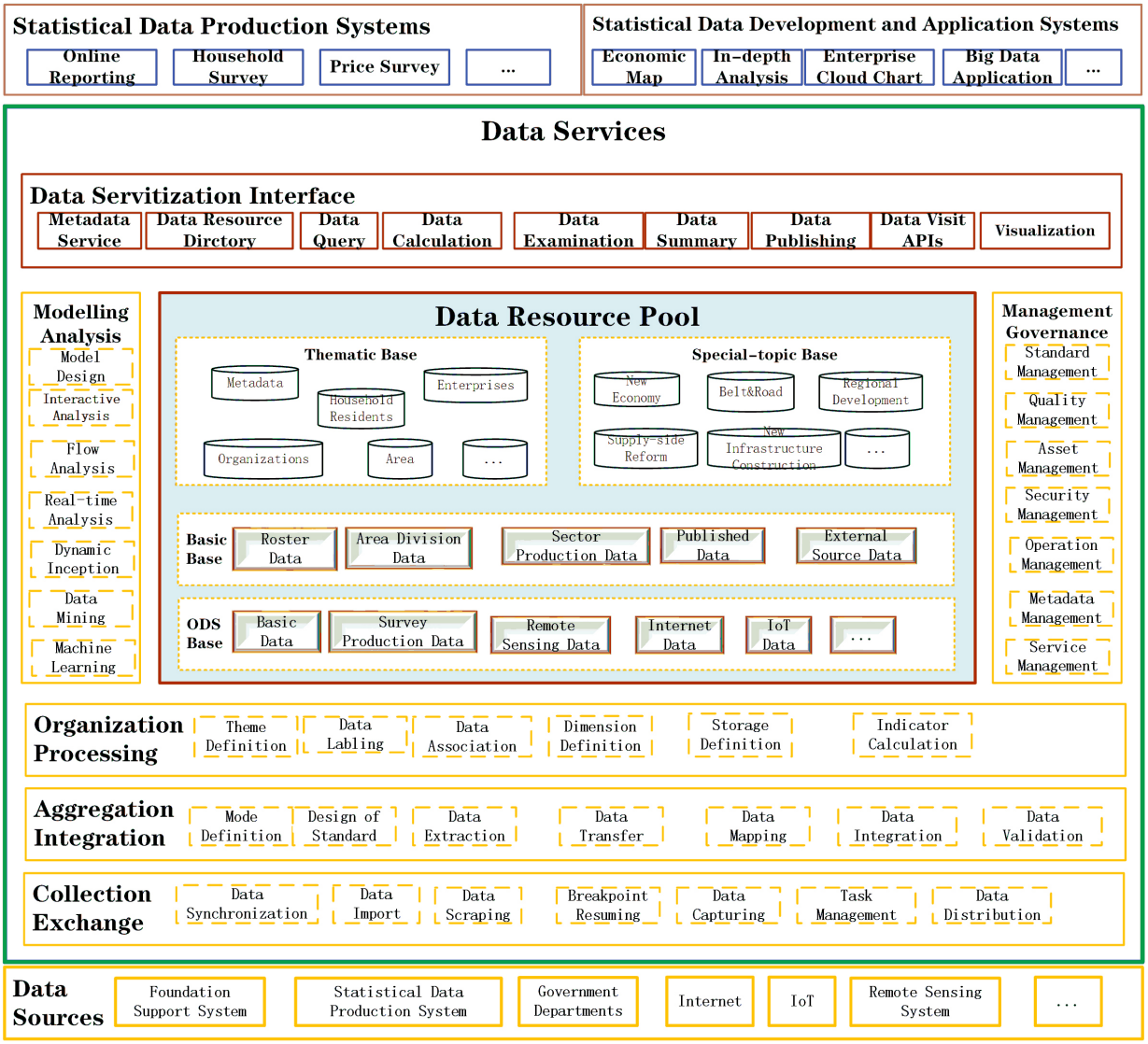

Fig. 1. Overall functional architecture of the statistical data middle platform.

ment, and the likes, which help build asset-based, service-oriented, and standardized data systems.

5. Modeling and analysis. Modeling and analysis are important carriers of the outputs of a data middle platform's capabilities, and they are also direct expressions of the value materialization of a data middle platform. Modeling and analysis include traditional OLAP-type data analysis, data statistics, and data mining. With the continuous addition of new modeling and analysis methods such as streaming analysis, real-time analysis, dynamic perception, machine learning, etc., the modeling and analyzing capabilities of the data middle platform are also continuously enhanced in depth and width.
6. Service and application. Based on its accumulated big data assets and powerful storage, calculation, and processing capabilities, a data middle platform provides accurate services to front-end applications in an intelligent and visual way. Such capabilities allow data to be used and flowed thereby providing an important means of external empowerment, and reflecting the ability of a data middle platform to feed the business system in a service mode. Data form closed loops between the data middle platform and the business system through the form of service and application,

which can promote the continuous iterative upgrades of the business system and the data middle platform itself.

3.4. Technical construction of the middle platform

The functional architecture of the data middle platform provides a static mechanism based on the reuse capability and external empowerment in the form of a service. To make this mechanism play well its role, the cloud business platform requires a suitable operation carrier. The Statistical Cloud adopts a "container + micro-services" technical route, and uses micro-services as the operation carrier for the platform. As the middle platform emphasizes the core reuse 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 micro-service 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 they are developed and maintained by different teams. Lightweight communication mechanisms (such as HTTP RESTful API, or RPC) are used between micro-services, which can be deployed independently and automatically, and can use different protocol stacks, languages, and storage. The micro-service embodies decentralization and natural distribution, and is of a self-closing loop service, which makes it a suitable technological framework for the implementation of the platform.

As for the stage of the data middle platform construction in practice, NBS has already completed the general design of the Statistical Cloud, in which the statistical data middle platform is included as one of the important architectural components. The further development of the statistical data middle platform is still going on. In this study, we conduct a round of proof-of-concept. Specifically, we construct a prototype for the Statistical Cloud, which includes all basic layers and elements of a typical Cloud such as IaaS, PaaS and SaaS. The proof-of-concept result shows good effectiveness and adaptability of the Statistical Cloud prototype to the statistical businesses. After completing proof-of-concept, NBS will further push forward the detailed design of the Statistical Cloud, and implement the development and deployment of both softwares and hardwares. Considering that the micro-service technical framework has advantages in supporting different development languages such as JAVA, C#, Python, Go, etc., the software stack will certainly not be constrained on the overall systematic level.

4. Discussion and conclusion

The International Data Corporation (IDC, https://www.idc.com/) predicts in its white paper that 60%+ of global GDP will be digitized by 2022, with growth in every industry driven by digitally enhanced offerings, operations, and relationships and almost $7 trillion in IT-related spending in 2019–2022 [13]. Data are increasingly valued and recognized by the whole society. Taking the opportunity of constructing a statistical cloud, the NBS implements its data resource planning in a global perspective with the concept of "data middle platform" and builds a comprehensive data capability platform that is oriented to value reuse and long-term development. As a new innovative approach, the data middle platform can fundamentally break down the technical barriers of data production, storage, analysis, service and circulation, and provide a bottom-up solution for solving the data island problem.

Apart from solving the data island problem of NBS, the data middle platform is also of great significance for improving data governance in other industries. By solving the data islands problem, tolerating the entry of various data sources, and building statistical big data platforms, the government statistics could realize the data 'servitization', that is, the data middle platform can provide data support to the end users in the form of service-oriented application programming interfaces (APIs). Then after data 'servitization', the government statistics could better empower its data partners, cultivate a more healthy and more dynamic statistical data ecology, and thus integrate into the Web of Data.

Acknowledgments

The authors would like to thank the editor Pieter Everaers and the anonymous referees for their constructive suggestions. We are also grateful to Gemma Van Halderen, Matthew Shearing, Eileen Capilit and other colleagues from the UNESCAP team for their helpful comments. Special thanks go to Ronald Jansen from UNSD for his kind mentorship. We also extend our gratitude to the participants at the Asia Pacific Statistics Week 2020 for their valuable discussions. Chunzhen Zhang gratefully acknowledges the colleagues at the National Bureau of Statistics of China for their supports and inspirations for this research.

References

[1] https://en.wikipedia.org/wiki/Information_silo.
[2] China Institute of Information and Communication. White Paper on Data Infrastructure in 2019.
[3] https://en.wikipedia.org/wiki/Enterprise_service_bus.
[4] https://harborresearch.com/tale-two-smart-cities/.
[5] https://www.dataversity.net/can-blockchain-eliminate-the-data-silo/.
[6] Ontotext. What are Linked Data and Linked Open Data? 2019.
[7] https://en.wikipedia.org/wiki/Semantic_Web.
[8] Valentine R. Success and sustainability at Supercell. https://www.gamesindustry.biz/articles/2019-01-15-success-and-sustainability-at-supercell.
[9] Zhong H. The transformation of enterprise IT architecture: Alibaba's strategic thinking and structure practice of middle platform. China Machine Press, Beijing; 2017.
[10] https://www.chaoqi.net/ganhuo/2019/0405/192928.html.
[11] Chen X, Luo J, Deng T, et al. Middle-Platform Strategy: Middle-Platform Construction and Digital Business. China Machine Press, Beijing. 2019.
[12] https://en.wikipedia.org/wiki/Operational_data_store.
[13] Gens F, Prete CD, Minton S, et al. IDC FutureScape: Worldwide IT Industry 2019 Predictions. https://www.idc.com/research/viewtoc.jsp?containerId=US44403818.