# A Survey of Data Governance Research Based on Industrial Internet

**Rong Li*, Weiye Meng, Xuemin Wang and Zhixia Guo**

China Telecom Corporation Limited Beijing Research Institute, Beijing, China

*Corresponding author: lirong6@chinatelecom.cn

**Keywords:** Data governance, Industrial Internet, Data quality, Data specification, Data security

**Abstract:** In the environment of massive and multivariate data provided by Industrial Internet, data governance highlights the practical significance and application value. Data governance includes the evaluation and supervision mechanism in the whole process from data acquisition to processing to final application, aiming at improving data quality, realizing data sharing and finally forming valuable data assets. Firstly, this paper introduces the necessity of data governance and the research problems encountered in data governance under the Industrial Internet environment. Secondly, from different perspectives such as data specification, data quality assessment, metadata management and data security, the existing data governance schemes in industry and academia are described in detail. Finally, the development trend of data governance technology in the field of Industrial Internet is summarized, and the prospects are given.

## 1. Introduction

With the rise of the cloud computing, the strong data processing capabilities help traditional manufacturing technologies come together with industrial knowledge, emerging the concept of Industrial Internet. Industrial Internet can connect the interconnection of equipment with user and data, reflecting equipment status in real-time, initiating control instructions and optimizing operation strategies, solving problems faced by the manufacturing industry in data usage.

Industrial Internet have become an important development trend. Under this background, how to manage the industrial data is a practical and challenging problem. The production data generated by equipments will form a large and complicated dataset after collection and aggregation. If there is no effective management method, these data are difficult to be utilized.

Data governance is a concept that is not completely defined in the industry. Its core is to evaluate, guide and supervise the whole process from data collection to use. The main goal of data governance is to improve the quality of data and make it valuable1. It is generally believed that data governance includes dimensions like data standard specifications, data quality management, metadata management, data security protection and many other different dimensions2.

At the China Industrial Internet Summit in 2019, many experts pointed out that data governance is facing some difficulties. Wang Chen and others pointed out that due to the acquisition accuracy of the sensor, the quality of the collected production data is often unsatisfactory; the existence of external interference, such as the gradual interference of the production environment, will cause data loss during transmission process and affects data quality3.

After data collection, it also faces data governance challenges. Wang Chen and others believe that traditional data integration aimed at data with similarity, but data integration solutions for discrete manufacturing needs to consider dynamic changes. As shown in Figure 1, timing changes is a category that traditional data integration method has not considered, which is a difficult point of data governance3.

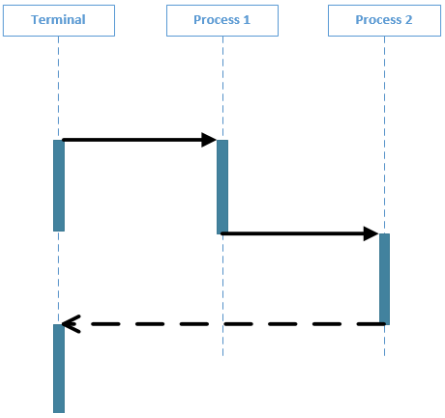


Figure 1. Example of time series change of data with process in the Industrial Internet.

Another data governance dimension that cannot be ignored is data security. Huang Wei and others pointed out that industrial enterprises used to be in a closed environment, data security risks mainly from hardware. However, as the Industrial Internet promotes data on the cloud, data security becomes very complex, including cloud security, protocol security, etc. If each part is solved in isolation, it cannot really meet the needs of the enterprises3.

This paper takes the above-mentioned problems of data governance as the starting point, expounding the current industry research status, giving some solutions from the industry and academia, summarizing the hot issues which is the latest concern.

## 2. Data Governance Sub-Dimension Content And Purpose

In the Industrial Internet, data have a multi-step process from generation to use. The process of generating data from production equipment, collecting through sensors, transmitting it to the data platform is called data collection phase. The data processing phase refers to the process of perform certain processing logic through various computing frameworks, carrying out data analysis. Data used by applications and displayed for users called data service phase.

As mentioned earlier, data governance includes different sub-dimensions. In the Industrial Internet, the data governance framework based on these sub-dimensions is shown in Figure 2.

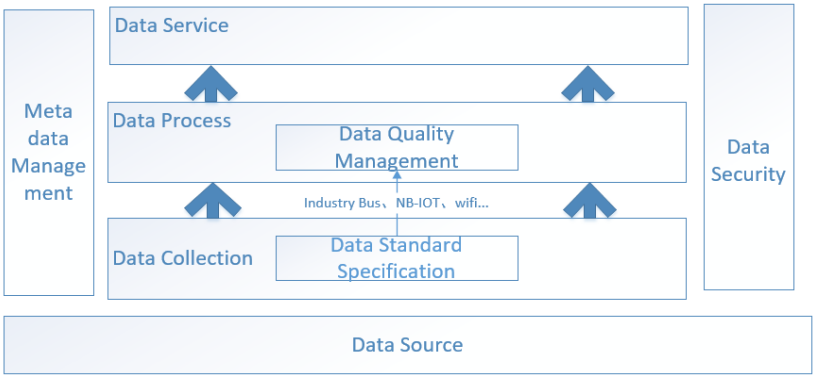


Figure 2. Data governance framework and main sub-dimensions under Industrial Internet.

As we can see from Figure 2, data governance in the data collection phase is data standard specifications mainly, the purpose is to uniformly the format and accuracy of the collected data. This is because the data acquisition accuracy of various device sensors is different, and the collected data comes from different production equipment, thus data often presents a variety of different structures. Standardizing the precision and structure of data can help developers to achieve on-demand processing of data more effectively.Data governance in the data processing stage mainly focuses on data quality management. Data quality management can evaluate the credibility of collected data and avoid the flow of low-quality data to subsequent data services.

Metadata management and data security protection cover the above three data phases. By implementing metadata management, data users can quickly grasp data state, obtain key information such as data types, data distribution, data relationships, and then build a data catalog for user. Data security protection provides multiple levels of access control rights for data users, ensuring data interconnection and privacy when data clouding.

According to above analysis, it can be seen that data governance under the Industrial Internet is a complex activity that concluds multiple sub-dimensions and achieves different goals. This paper will introduce the data standard specification, data quality management, metadata management, and data security in detail.

## 3. Data Standard Specification During Collection Phase

In big data technology, there is a common problem affecting data usage, it called multi-source heterogeneous data. It refers to data which from different sources and the data formats including structured data, unstructured data, semi-structured data, or the structured data is inconsistent. Multi-source heterogeneous data may cause significant problems for data analysis, and make data traceability difficult when data generated. In the Industrial Internet, data from different equipments collected by sensors is a typical multi-source heterogeneous data case. In order to solve this problem, it is usually necessary to introduce a data standard specification.

The data standard specification refers to the construction of a common data standard that conforms to industry standards and actual use conditions, so that data items of multi-source data can be unified and integrated. Appropriate data standard specifications can be established through a rule processing engine that includes multiple data transformation logic or a standard code base mapping.

in the Industrial Internet, enterprises common to follow the data standard specification before data enter into the processing platform, eliminating the problem of multi-source heterogeneity. At present, the industry community often adopt a method of formulat-ing a unified acquisition and transmission protocol for terminal data to clarify the data structure and data precision. Some Chinese large-scale enterprises working with their equipment providers to develop a unified data collection and transmission protocol, so that the industrial data generated by devices is formatted when data collected and transmitted.

Another solution is to develop sensor-based data-compatible middleware and provide data conversion capabilities through middleware. The main design ideas are shown in Figure 3. Zhang Jianxiong and others advocate a method to use industrial Ethernet, industrial bus, NB-IoT and other wired or wireless communication tech-nologies accessing production equipments to collect industrial data4; then use protocol analysis and conversion technologies like ModBus, CAN, Profinet to realize data format unification; then use HTTP, MQTT, etc. to transfer the converted collected data to the cloud platform.

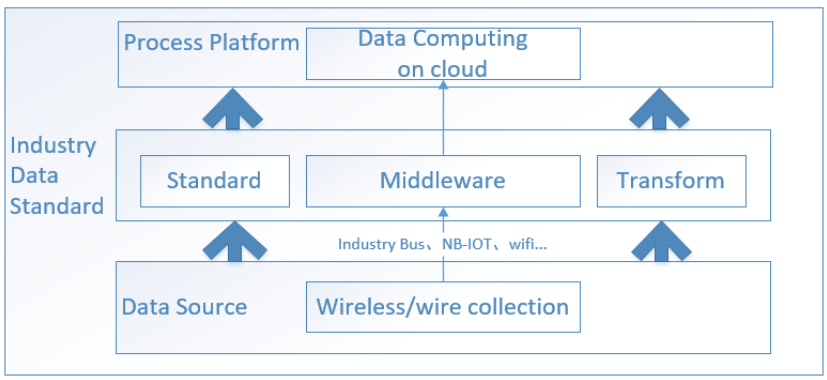


Figure 3. Data format unification method using data conversion middleware.

With the rise of edge computing technology, data standard specification based on edge data processing has become a new hotspot solution. This solution is based on high-performance

computing, real-time operating system and edge analysis algorithm, processing data in the network close to the production equipments, converting data to a unified format.

## 4. Data Quality Management In The Data Processing Phase

Data quality management is a key part of the data governance framework. In the context of Industrial Internet, data quality management can ensure the reliability of data collected by sensors, so that the result of data processing can reflect the real commands of users and produce practical application value.

At present, the research work on data quality management can be divided into the following categories: data quality evaluation model, data quality evaluation method, data quality rules model, data quality evaluation algorithm5. in the Industrial Internet, the most common method is data quality evaluation model. Data evaluation model can be judged from different indicators according to user needs. The general data quality evaluation indicators are based on 15 dimensions, including accuracy, completeness, consistency, credibility, timeliness, etc.5.

The rule base is also an important mode. It can determine the evaluation index of data quality according to the business situation. For example, the ratio of the null value to the outlier value of the collected data can be calculated to judge the accuracy of the collected data. For the data collected periodically, starting from the integrity indicator, monitor the total amount of data collected by the current batch, comparing the increase and decrease extent of the average value of the data with the previous N batches, to confirm whether the data transmission fails, resulting in a decrease in data quality.

There is no standard widely recognized by the industry currently. Therefore, the scheme of establishing a data quality evaluation model in combination with the rule base is more easily accepted by the industry in production. Liu Fang proposed a evaluation scheme based on the rule base5, combining various rules in the rule base, setting different weights according to the importance of data quality evaluation indicators in specific needs, adopting a combination method of simple ratio and weighted average, calculating the evaluation results and determine the level of data quality. It reflecting the proportion of abnormal data (empty data / dirty data) successfully. An example of this method is shown in Table 1.

Table 1. Example of data quaility assessment combined with rule base and indicator weights

| Rule ID | Rule Name | Rule Type | Weight | Mutable |
|---|---|---|---|---|
| 1 | Increasing or decreasing extent less than 3% | completeness | 0.25 | Yes |
| 2 | empty fields less than 2% | accuracy | 0.2 | Yes |
| 3 | abnormal fields less than 5% | accuracy | 0.1 | Yes |
| 4 | data collected in an hour | timeliness | 0.05 | Yes |
| 5 | whether data have same structure | consistency | 0.25 | Yes |

The above method of combining the rule base with the evaluation model has the advantage of being able to adapt well in different manufacturing industries. For different requirements of industrial data collection, users can flexibly configure rules and weight values to achieve personalized data quality management, which has the significance of actual promotion. The disadvantage is that the operation and maintenance of the data quality management after the implementation is more complicated, and it is necessary to continuously maintain the rule base or modify the index weight according to the business adjustment.

## 5. Metadata Management For All Phases Of Data

Metadata management is the highlight of the data governance framework. In the traditional definition, metadata is a type of data that describes data, mainly describing the data structure and establishment method of the data system storage. Taking an example, the table name, owner, physical storage location, primary key, and fields structure are all metadata. Under big data technology, metadata can help data system managers and developers quickly find the data which

they care about. Metadata in big data systems can be divided into two categories according to their purposes: technical metadata and business metadata6. Technical metadata is data that stores technical details about big data systems and is used to develop and manage data used by data storage systems7. Business metadata describes data in big data systems from a business perspective, which provides the semantic layer between user and actual system enables the business person to quickly understand the meaning of the data in the database8.

Metadata management usually includes the development of business vocabularies, definitions of data elements and entities, business rules and algorithms and data characteristics9. in the Industrial Internet, the main content of metadata management is same as above, and the purpose is to form a data asset catalog. The most basic metadata management is the collection, organization and maintenance of business metadata. In data governance, the application of technical metadata is an important part.

A popular current metadata management tool in the industry is Apache Atlas, a metadata management software under the Apache Foundation, which provides users with open metadata management to build a system's data catalog, classifing and managing these datasets. This software provide data users, data analysts, and data governance teams with collaboration capabilities around these datasets. In addition, Atlas manages metadata sharing, data grading, auditing, security, and data protection to integrate with Apache Ranger, whichi is another product under thr Apache.

## 6. Data Security Protection At All Stages Of Data

Industrial Internet promotes a trend of transmitting large amounts of industrial data on the cloud, which brings strong data usage capabilities and creates corresponding economic benefits. However, because the Industrial Internet has broken the closed and credible industrial environment, thus the Internet security threats can be entered into the industrial environment. Industrial control system's protocols are mostly in clear text, the industrial environ-ment uses common operating systems not updated in time, the network security awareness of employees not high, above conditions means that there are many exploits in the Industrial Internet10.

The most famous events about data security include the hacking of US water companies in 2017 cause a leakage of large user informations. The "magic cave" ransomware is spread around the world based on Windows security vulnerabilities also in 2017, directly threatening the network field. Therefore, researching the safety management of industrial data and strengthening the data protection for enterprises is a hot issue of data governance in the Industrial Internet.

China Electronics Technology Standardization Institute divides the protection system of industrial data security into three levels: accessment security, platform security, and application security11. As shown in Figure 4, industrial accessment security provides a security mechanism for collection, transmission and conversion industrial data; platform security provides the basis for security for industrial data storage and computing; application security provides strong security control for upper-layer applications and data access.

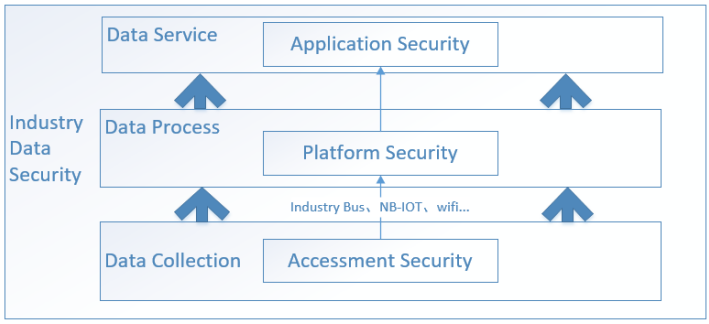


Figure 4. Industrial data security hierarchy and stage.

The most popular research direction at present is industrial application security. The application security should be considered in several aspects: support application access signature mechanism, to

ensure that only authorized applications can submit data access requests; support application data on-demand access, avoiding data access scope expansion; support application behavior real-time monitoring, intercepting the attack behaviors in real-time including abnormal access frequency and SQL statement illegality; establish a complete application process management mechanism, ensuring the approval of each application, avoiding malicious manipulation or misoperation from high-rights personnel; build a complete application testing environ-ment and test specifications to ensure that only applications that comply with security policies can be approved for execution.

In the industry, Siemens has established a strategic partnership with Identify3D (ID3D) for data security licensing and encryption. A solution for digital production process safety and data traceability is provided by ID3D and integrated into Siemens' product lifecycle management system12. By working with Identify3D, Siemens has implemented an end-to-end security solution that protects industrial data security throughout the process and demonstrated the integration of the software in September 2017. In this way, Siemens's customers are able to protect their data, produce products according to the defined parameters, and track the production components.

Research in other areas of data security is constantly evolving. Research in other areas of data security is constantly evolving. In terms of data accessment security, Symantec has proposed two solutions: Critical System Protection provides feature-free, host-based security for IoT device. Device Certificate Service provides a solution for industrial equipment data protection based on the Public Key Infrastructure (PKI) implemented by Elliptic Curve Cryptography (ECC)13.

## 7. Summary And Outlook

This paper summarizes the content and current research status of data governance in the Industrial Internet. According to above sections, data governance is a complex activity that includes many different sub-dimensions. Implementing data governance from these sub-dimensions can promote industrial data as a core treasure.

However, as an emerging technology, data governance will inevitably encounter various problems in the development, thus it will continue to face new challenges. With the scale of industrial data increasing, diverse data types and the new concept that data transfer to knowledge, which increase the complexity of data governance greatly. Under this circumstance, whether the current data governance content can meet the expectations of various enterprises?

The integration of multi-source data expose the correlation between data and other data, which may lead to the leakage of technology, reducing the confidence of enterprise to develop Industrial Internet. Thus, research the initiative strategies to reduce the risk of privacy data disclosure is a very interesting directions.

Because data governance is a long-term and complex activity, a short-term investment in a large amount of manpower and money may not be able to quickly achieve data quality improvement. Therefore, data governance also faces the risk of investment costs14.

As a enterprise, it also requires a standardized management method to execute data governance. It needs a long-term practice in actual.

## References

[1] S.Zhang, R.Pan, Y.Zong. Big Data Governance and Service. The first edition, Shanghai: Shanghai Science and Technology Press, 2016.1-224.

[2] W.Gao, J.Qi. Analysis of Security Problems in Big Data Governance. Wireless Internet Technology, 2018.

[3] What are the scientific and technological problems facing the development of the Industrial Internet? https://www.iyiou.com/p/98199.html.

[4] J.Zhang, X.Wu, Z.Yang, et al. Research and Application of Industrial Data Acquisition Technology Based on Industrial Internet of Things. Telecommunications Science, 2018, 34(10): 130-135.

[5] F.Liu, M.Li, H.Ren, et al. Data quality assessment method based on rule base. Journal of Computer Systems, 2017(11): 167-171.

[6] F.Wu, G.Xing, Y.Xing. Research on ETL design based on ontology. Computer Engineering and Design, 2007, 28(7): 1517-1919.

[7] Q.Zhang. Research on decision support technology for road transportation safety management based on OLAP. Chongqing University, 2007.

[8] Y.Liu, G.Liu. Discussion on the Management of Metadata in Geographic Information Industry. Surveying and Spatial Geography Information, 2009, 32(5): 147-149.

[9] Data view. Metadata management analysis and data warehouse and main data introduction. http://www.cbdio.com/BigData/ 2016-04/14/content_4803027.htm

[10] CEchina. When people talk about industrial big data, what are they talking about. http://www.ccw.com.cn/ Make/2019-07-30/8500.html

[11] China Electronics Technology Standardization Research Institute. Industrial White Paper (2019 Edition).http://www.cbdio.com/BigData /2019-04/02/content_6066048.htm.

[12] Siemens secures digital value chain with data encryption. Network Security and Informatization, 2017(12): 13-13.

[13] L.Du, S.Chen, Y.JiAng, et al. Research on Key Technologies of Industrial Internet Security. Information and Communication Technology and Policy, 2018, 292(10): 17-20.

[14] X.Wu, B.Dong, X.Cao, et al. Data governance technology. Journal of Software: 1-27. https://doi.org/10.13328/j.cnki. Jos.005854.