Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

4D Publishing Group

Research on Data Producing Services across Compute Engines

Qi Wang[1,*], Rui-yun Wang[2], Ming Lei[1]
[1]China Justice Big Data Institute Co., LTD., Beijing, China
[2]China Electronics Technology Group Corporation 15, Beijing, China

[*]Corresponding author's e-mail:
zzyyfly@outlook.com

Abstract. Currently, data production services are facing the following challenges: achieving cross-computing engine scheduling, automated accurate labeling, and on-demand task scheduling. In this paper, cross-computing engine scheduling technology for business data production, domain-oriented business data automatic labeling technology and business data production visualization process scheduling technology are investigated, while high concurrency-related technical solutions are studied to improve the service processing performance. At the end of the paper, technical risk assessment analysis is also given in four aspects, including computing resource scheduling, labeling reliability, process scheduling optimization and thread safety.

Keywords. Cross-computing engine scheduling; Automatic data labeling; Data production task flow scheduling; High concurrency

1. Introduction

Data production service is an important part of data service[1], which is oriented to the data production service requirements of different users under different task requirements and support environments, and effectively realizes data product production requirement collection, work order management, data format conversion, processing of self-built data resources, processing of Internet and outsourced data resources, business data production model building, visualization of business data production tasks process configuration, business data production scheduling management, visual display of business data production results, cataloging and registration of data products, and finally further realization of user-oriented visual management and task-oriented processing of business data. Currently, data production services face the following challenges: cross-computing engine scheduling for business data production, automated and accurate labeling of business data, and on-demand configuration of data product production task scheduling.

In recent years, the problem of scalability in cross-computing engine scheduling has become a difficult point of concern. On the one hand, the production of data is often not a uniform process, usually the data production rate will continue to increase to reach the peak, and then fall back from the peak, the data volume will be huge during the peak period, and may last for a period of time, if not timely scheduling will cause a large accumulation of data, which may eventually lead to server downtime failure, so the scheduling apparatus needs to have good scalability and can dynamically adjust the system resources, so as to adapt to the growing production data and ensure the stability of system operation; on the other hand, when the data production rate falls back to a lower level from the peak, the excess system resources need to be recovered in time so as to avoid the waste of resources, while the user is indifferent to the whole resource dynamic adjustment process. Second, the cross-computing engine scheduling needs to be fault-tolerant: first, the flow of data is unidirectional and has a certain degree of continuity and real-time when producing data, so it is more expensive to replay data again if errors occur during the production process; second, during the scheduling computation, many low-value data are discarded and only a small amount of valuable data is saved, so incomplete data needs to be considered in the fault-tolerant design.

In addition, it is necessary to design fault-tolerant mechanisms for different data usage scenarios, so as to save system resources as much as possible while meeting application requirements. Meanwhile, cross-computing engine scheduling needs to support data state consistency management: firstly, there is a part of data in a large-scale distributed environment that needs to maintain its state consistency, and how to identify this part of data from the huge amount of heterogeneous data is a major difficulty faced by the cross-computing engine scheduling system; secondly, how to achieve state consistency of this part of data among different computing engines to meet the requirements of business data production under various scenarios is also a difficulty that needs to be solved. Secondly, how to achieve state consistency of this data among different computing engines to meet the requirements of business data production under various scenarios is also a difficult problem to be solved at present.

Although YARN[2]developed by Apache open source community provides unified resource management and scheduling for upper layer applications, which brings great benefits to clusters in terms of utilization, unified resource management and data sharing, and is an important support in scheduling research across computing engines, YARN, S4 and other systems currently do not have better solutions in terms of scalability, and in terms of fault tolerance mechanism, YARN, Puma[3], S4[4-5] and other systems only support partial fault tolerance, and in terms of consistency, YARN, S4 and other systems currently do not support or only partially support the management of state consistency.

Automatic data annotation techniques always target a certain domain or a certain language, while methods for automatic extraction of language-independent keywords are still rare and the accuracy rate is low, not even more than 20%. Statistical methods are more general, but less accurate; linguistic methods and artificial intelligence methods are more accurate, but less general. How to solve the problem of generality of automatic keyword citation system will be a research direction of automatic keyword citation technology[6]. In recent years, automatic data citation methods based on semantic analysis have developed very rapidly, while semantic analysis relying only on semantic dictionaries is still far from enough, and better semantic analysis knowledge system is needed for intelligent annotation capability. In addition, supervised and semi-supervised machine learning methods need to use a large number of labeled data samples for training, but traditional data labeling requires a lot of labor costs and time costs, which leads to some constraints on the construction of the whole system. Currently, in most application areas, only a small portion of the data is labeled, while a large amount of unlabeled data is available. How to use a small number of labeled data samples to learn and train and obtain a classification model with high accuracy will be the intelligent development direction of the automatic keyword labeling method based on machine learning.

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

Data production visual process orchestration aims to orchestrate and design the processing process of data production by means of tola drag and drop. Traditional data production process orchestration is usually designed by hard coding or configuration files, which is less flexible and difficult to maintain and modify later. Data production visual process orchestration is closely related to computational image technology, data visualization technology, process engine technology, distributed scheduling technology, etc. Currently, it is an extremely active and critical aspect of industrial production in foreign countries. Data visualization process orchestration which realizes the unification of the field of scientific visualization with business process orchestration.

To address the above problems, this paper investigates the cross-computing engine scheduling technology for business data production, the domain-oriented business data automatic labeling technology and the business data production visualization process orchestration technology. In order to improve the overall processing performance of the system and utilize the system resources more effectively, this paper also investigates the high concurrency technology solution. Finally, this paper gives a technical risk assessment analysis from four aspects, including computational resource scheduling, marking reliability, process scheduling optimization and thread safety.

2. Cross-computing engine scheduling technology for business data production

Spatial cross-computing engine scheduling is the key to realize data production services. The distributed computing capability relies on the cross-computing engine scheduler to provide core support and key technical foundation for massive and complex data processing in data service systems. The nodes of cross-computing engine scheduler are deployed in multiple data centers and are an important part of the data center, which mainly provides powerful distributed computing capability to support massive and complex data processing for a wide range of users and various development applications, especially for incoming cleaning and screening, large-scale analysis and mining, complex relationship analysis and intelligent processing, etc. It is generally divided into basic computing services, value-added computing services and auxiliary support services. The system is the core support and key foundation for the real-time processing capability, analysis and mining capability, value-added processing capability and intelligent processing capability of big data engineering. The system relies on special cloud operating system and data organization and management system to provide strong support for data service system, business application support service system, etc.

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

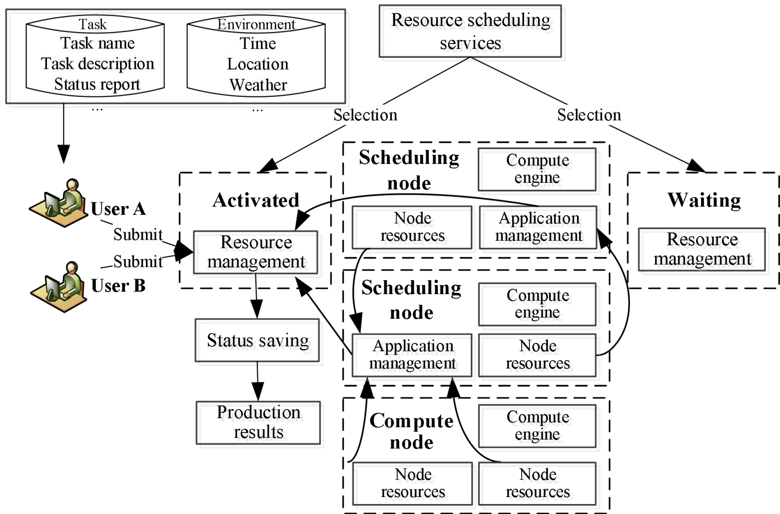

Figure 1. Scheduling across compute engines technical implementation diagram

Based on various processing frameworks such as real-time computing, offline computing, and graph computing, the cross-computing engine scheduling service system can carry out analysis and mining based on metadata and multi-factor correlation analysis of data; provide distributed machine learning and deep learning algorithm support based on multilingual processing frameworks, support exploratory analysis and mining for massive business data, and feature analysis of multi-source heterogeneous data such as video, image, voice, and text; and provide value-added service capability in accordance with multiple dimensions such as rules, space-time, and topics.

Cross-computing scheduling includes two aspects: one is how to solve the bottleneck of storage management in a large amount of industry data environment; the other is to solve the scheduling bottleneck of massive task cross-computing engine scheduling in the domain environment, so as to produce a cross-computing engine scheduler suitable for business people.

This system takes into account the complex execution environment of domain tasks and makes efficient use of system computing resources. Since task execution queues may be congested in a specific domain environment, if there are remaining resources in one queue, this remaining resource can be allocated to other queues, and when new computing tasks keep arriving in that queue and the existing resources are not enough, the resources previously allocated to other queues are reclaimed to that queue. This flexible allocation of resources can significantly improve resource utilization. Data production administrators can set a lower and upper limit for each queue to use the resources that will be shared by the compute tasks submitted to that queue.

Among other things, different data production tasks support multiple applications running simultaneously and multiple users sharing the cluster. To prevent a single application from taking up all the cluster computing resources, the data production administrator can set the maximum resources each application can use and the maximum number of tasks it can create, and to ensure security, the system restricts the users of each queue through a security access control list, allowing only privileged users to view the execution status of tasks or perform related operations.

Specifically, the following frameworks are required. Designing a multidimensional, fine-grained resource allocation framework; a resource scheduling framework for task prioritization; and a task recovery framework for automatic node failure discovery. The multidimensional, fine-grained resource allocation framework is a scheduling strategy. The basic idea of this strategy is that a data production user can label each resource node with a tag, such as p1, p2, etc., to indicate the characteristics of that resource node; at the same time, a data production user can label each queue in the scheduler with a tag, so that a job submitted to a certain queue will only use the resources on the node labeled with the corresponding tag. The resource scheduling framework for task priority allows users to define tasks with different priorities, with later-started high-priority tasks able to access resources released by running low-priority tasks; compute containers not started by low-priority tasks are hung until the high-priority tasks complete and release resources before being restarted. In the manual mode, resource management will enter a waiting state after discovering node failure and needs to be switched to an active state by an administrator through a command before the compute engine scheduler cluster can provide services to the outside world; while the automatic mode is implemented based on the application coordination service, the basic implementation idea is that after resource management finds a node failure, it will create an application in the same directory under the coordination service, the creator will enter the active state and the others will automatically switch to the wait state.

3. Domain-oriented automatic marking technology for business data

Although the amount of data that can be collected is huge, there are still a series of problems: first, the quality of data varies; second, there is a lack of correlation between data; third, data labeling is extremely difficult, and a large amount of data still relies on manual research and evaluation, and in the process of manual research and evaluation, there is also a lack of effective tools and systems to accumulate the results of research and evaluation, which cannot provide data support for the subsequent development of intelligent applications. Research on intelligent data annotation technology to achieve effective annotation of big data so that it can be directly used for intelligent model development and training. These include collaborative labeling, interactive labeling, semi-supervised labeling, and unsupervised labeling. Collaborative labeling is the simultaneous labeling of target samples by multiple people, interactive labeling is the interactive data labeling business conducted by online labelers through the front-end UI interface, and semi-supervised labeling and unsupervised labeling are intelligent labeling services with the help of semi-supervised learning or unsupervised learning machine learning methods.

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

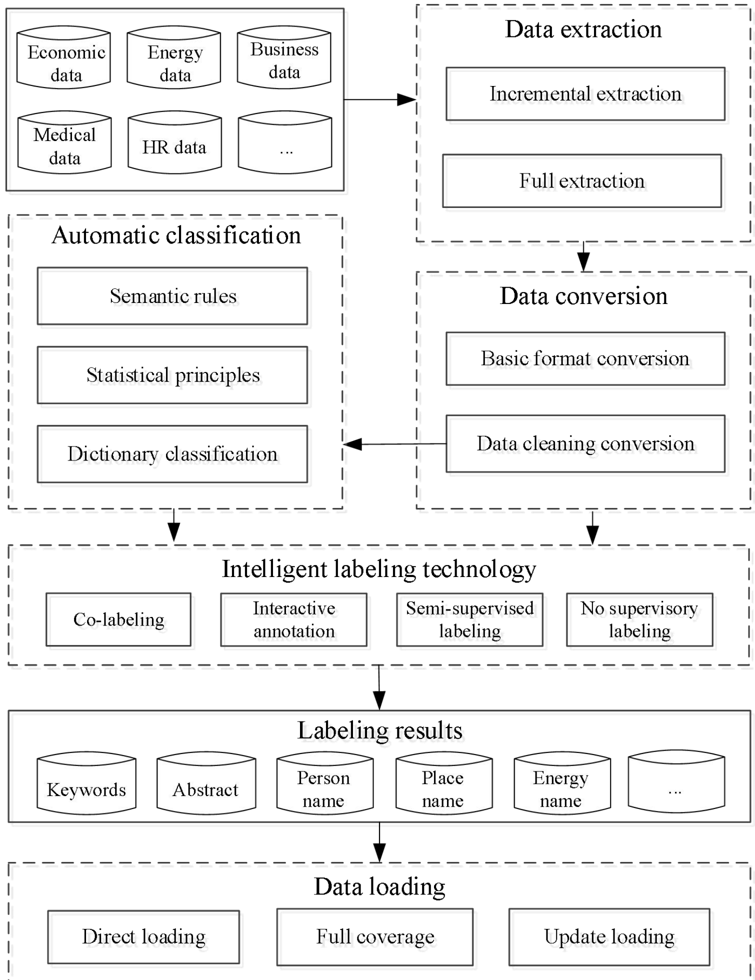

Figure 2. Data marking technical implementation diagram

The specific ideas are as follows: (1) In order to support intelligent data labeling, it is necessary to build a labeling system suitable for the intelligent application, and the intelligent data labeling system will be constructed by using a combination of semantic networks, knowledge graphs and a priori rules; (2) Based on the labeling system and then use collaborative labeling to obtain a certain amount and quality of labeled data; (3) Training the acquired annotated data, constructing automatic data annotation models, grading the automatically annotated data, introducing manual determination for annotation results with low confidence, and further promoting the optimization of the automatic annotation models for the determined results; (4) Establishing a quantifiable mechanism for judging the labeled results, and preferably selecting a combination of labeling patterns to put them into long-term operation for building intelligent computing datasets.

Data marking technology first needs to be loaded based on data extraction transformation. Data extraction imports data from the data service to the predecessor in a simple drag-and-drop manner through the graphical adaptation of the business system in the user's business. The data from the business system is

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

converted into the internal format of the data integration application system through extraction. The extraction methods are incremental extraction, complete extraction, etc.; the timing of extraction is to avoid the peak hours of user business as far as possible; the extraction period is for different types of data sources, and a reasonable extraction period is formulated by integrating the business needs of data marking and system cost. The implementation supports automatic and manual data extraction methods, file extraction and loading of various common formats, and graphical interface-based configuration file verification rules and view verification results. Data conversion is divided into basic data conversion and data cleaning conversion. Basic data conversion is a simple data conversion method, such as file format conversion, XML parsing, HTML parsing, etc. Flexible data processing can be performed, such as decompression, record merging, conditional filtering, de-duplication, de-duplication of rows, etc. Data cleaning and transformation requires the use of automatic classification functions. It is divided into three classification functions: semantic rule-based automatic classification, statistical principle-based automatic classification and lexicon-based classification.

Automatic classification based on semantic rules in data production services for domain data freely maintain the classification word list, manually add or modify rules;there is no limit to the size of the word list. The rule-based classification supports multi-conditional with, or, and non-relationships, has the function of setting the number of word frequencies, and provides a convenient interface for rule definition. The classification method can also be updated at will, and there is no limit to the number and structure of categories. Support multi-level classification. Support category re-sorting. Users can assign different confidence levels to the rules, and select the rule with higher confidence level for matching in case of multiple rules conflict to improve the accuracy rate.

Automatic classification based on statistical principles in data production service data apply a variety of statistical methods for automatic classification, specific statistical classification algorithms are mainly: K-nearest neighbor method, decision tree, plain Bayesian classification, support vector machine SVM, random forest, AdaBoost algorithm, neural network algorithm, etc. The above classification algorithms can achieve good classification results in text classification tasks, and the accuracy rate can reach more than 70%. However, in practical applications, the accuracy rate is not enough to meet the business needs, so the algorithms need to be optimized and the parameters of the algorithms need to be continuously adjusted to obtain better classification results.

The lexicon for new things in the field should also be kept established and updated, and the classification method based on knowledge lexicon control also has a good automatic classification effect. The dictionary-based automatic citation module is used to cite the data. According to the characteristics of the data, the platform will develop the corresponding automatic citation rules for subject words by combining the industry citation experience and Chinese intelligent information processing technology. The automatic citation module is customized and optimized according to the multidimensional classification system to achieve identification based on the subject lexicon, and also identifies high-frequency strings based on statistical methods, and then filters the high-frequency strings to obtain keywords.

After automatic classification, effective annotation of big data needs to be realized by data intelligent annotation technology, so that it can be directly used for intelligent model development and training. The specific techniques include collaborative labeling, interactive labeling, semi-supervised labeling and unsupervised labeling. Collaborative labeling is the simultaneous labeling of target samples by multiple people, interactive labeling is the interactive data labeling business conducted by online labelers through the front-end UI interface, and semi-supervised labeling and unsupervised labeling are intelligent labeling services with the help of semi-supervised learning or unsupervised learning machine learning methods. The specific ideas are as follows.

① To support intelligent data labeling, it is necessary to build a tagging system suitable for cyberspace

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

force business applications, which can be constructed using a combination of semantic networks, knowledge graphs, and a priori rules.

② Based on the tagging system and then using collaborative annotation to obtain a certain amount and quality of annotated data.

③ Training the acquired annotated data, constructing an automatic data annotation model, grading the automatically annotated data, introducing manual judgments for annotation results with low confidence, and further promoting the optimization of the automatic annotation model for the judged results.

④ Establish a quantifiable mechanism for judging annotation results and preferably select a combination of annotation models to put into long-term operation for building data sets.

The data labeling function should also support the secondary development of data processing functions, provide standard plug-in development specifications, allow developers to develop special plug-ins to achieve personalized automatic labeling functions, and support plug-in "plug-and-play" system.

Data loading is to realize the extraction and transformation of data tags stored in the target database. When loading data, it is usually necessary to consider both loading frequency and data appending strategy. The data loading strategy needs to be determined according to the data extraction strategy and business rules, mainly including three types of append loading, overlay loading and incremental loading.

Append loading means that each load keeps the existing data in the destination table and appends the extracted data to the destination table. Overlay load firstly clears the existing data in the destination table, and then loads the extracted data directly. Incremental loading means that when loading, it first determines whether the loaded data already exists in the destination table, and if it does, it updates the original data with the new data, and if it does not, it inserts the data into the data table.

4. Business data production visualization process orchestration technology

Based on the business data production visual process orchestration technology needs to meet the following conditions: the various components of the data production business are dragged and dropped in a visual environment, arranged into a complete business process in an intuitive way, and can be run in a visual state to get the output product of the data production business process. This requires the use of workflow engine technology and visual display technology. The visual process orchestration is based on a mature workflow engine, which supports users to edit customized workflows in a user-friendly visual editor to achieve visual process orchestration for business data production.

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

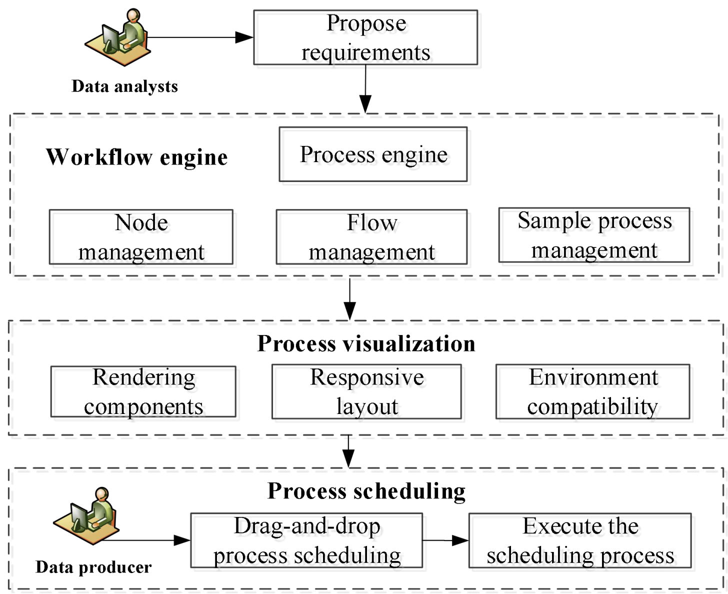

Figure 3. Visualizing process orchestration technical implementation diagram

Visual process orchestration for business data production is generally built based on workflow engine technology. The workflow engine can be used as part of the application system to provide core solutions for each application system, such as information flow routing, conditional judgment, on-demand approval, etc. The flow path of data can be decided according to the different roles, positions, posts and other conditions of users.

Based on the mature workflow engine, it can well meet the complex and flexible process needs of users. In the process design, an intuitive visual operation is adopted and the process can be defined by dragging and dropping. The whole workflow is described by XML specification and can be directly imported and exported. In the process drive and application, it supports various flow methods such as serial, parallel, counter-signature, single-signature and conditional flow, etc. Users can carry out various operations such as flow, sign-off, endorsement, rejection, request for rework, mandatory end, reassignment, etc. to meet the needs of various approval situations. The system also provides powerful process monitoring and process statistical analysis functions, which can record the complete process trajectory and help users master the work flow processing situation at any time. The visual workflow engine enables dynamic and efficient management of content workflows and provides a high degree of scalability and maintainability.

Data production-oriented visualization process orchestration business also requires the use of visualization theory, methods and technologies. Production data visualization is the study of techniques for presenting data in a visual representation. Visualization technology is the use of computer vision technology and image processing technology to convert data into images that can be recognized by the human eye, and display and interactive processing on the computer screen related methods and technologies. It involves many fields such as human-computer interaction, computer graphics, image processing, and computer-aided design, and is one of the most commonly used technologies in applications such as big data processing and intelligent decision analysis. Data visualization can effectively represent data and can visualize key features, thus helping users to discover the hidden values in the data.

Advanced Journal of Engineering(ISSN Online: 2771-9863)
October 2022 Vol.1, No.4, pp. 57-68
Doi: https://doi.org/10.55571/aje.2022.10011

After integrating the business requirements of each domain, the rendering of the business process module visualization of this system requires a JavaScript library based on data manipulation documents to help bring the most intuitive presentation to the data. By using HTML, SVG and CSS, the JavaScript charting library needed for business process orchestration visualization needs to be able to run on desktop computers and handheld mobile devices, and needs to be compatible with the mainstream browsers currently on the market, including Chrome, Edge, Firefox, Safari, etc., and to provide a variety of charts for data visualization, including Bar charts, line charts, pie charts, heat maps, dashboards, etc. Each process incorporates powerful visualization components that manipulate the DOM in a data-driven manner and can bind any data to a DOM node and then apply a data-driven transformation to a document. For example, an array can be converted to a table on a web page, or the same data can be used to create a scalable vector graphic with good adaptability. It also supports visual presentation of large volumes of data and dynamic interaction.

5. High concurrency technology for improving overall processing performance

High concurrency technology refers to the technology that enables multiple threads to execute concurrently from software or hardware. Computers with multi-threading capability are able to execute more than one thread at a time due to hardware support, thereby improving overall processing performance. Systems with this capability include symmetric multiprocessors, multi-core processors, and chip-level multiprocessing or simultaneous multi-threaded processors. In the domain data production business, computers with multi-threading capabilities are able to execute more than one thread at a time due to hardware support, thereby improving overall processing performance.

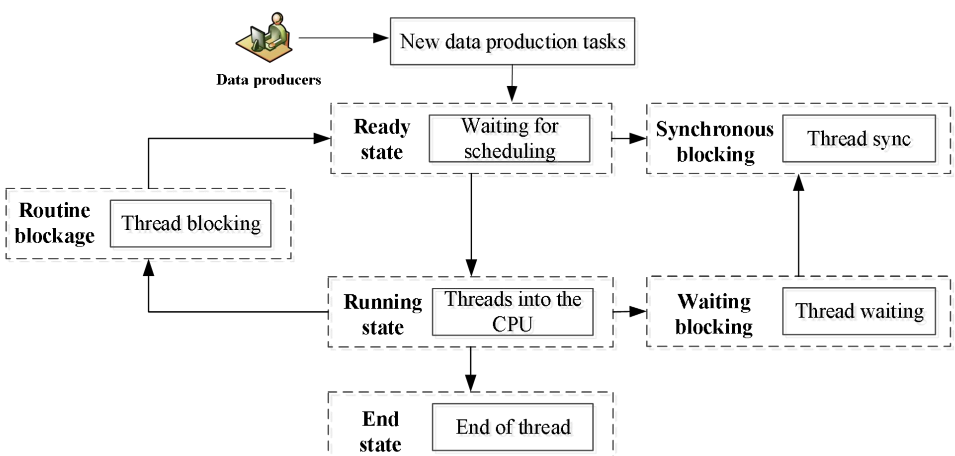

Figure 4. Multi-threaded operation technical implementation diagram

Multi-threaded hardware needs to support fast switching between ready and executing state threads. To achieve this, the hardware needs to implement saving and restoring registers that are visible to the program and some control registers that have an impact on program execution (e.g., program counter PC, program status register SR). Switching from one thread to another means for the hardware to save the value of a set of registers of the current thread and to restore the value of a set of registers of the upcoming thread. These thread switches can be done in one CPU cycle, and each thread appears to run as if it were alone, i.e., not sharing hardware resources with other threads. For the operating system, each thread is usually treated as if it had a separate processor, which simplifies the design of the system software.

In order to switch efficiently between threads, each thread needs to keep its own set of registers (register set). Some hardware is designed to have two register files per processor core to enable fast switching between multiple threads. The advantage of multi-threading over single-threading is that it is asynchronous.

In a real-world application scenario, Task 1 and Task 2 are two completely independent and unrelated tasks. Task 1 is waiting for the remote server to return data for post-processing, and the CPU is always in a waiting state, running "idle".The program is running in a single-task environment, so Task 2 has a very long wait time, which significantly reduces the efficiency of the system. Single-tasking is characterized by queued execution and significantly lower CPU utilization.

In multi-threaded operation CPU can completely switch back and forth between task 1 and task 2, so that task 2 does not have to wait for 10 seconds before running, and the operation efficiency of the whole data production system is greatly improved.

6. Technology risk assessment

The technical mechanism involves four issues: scheduling of computing resources, marking reliability, process scheduling optimization, and thread security.

The scheduling of computing resources includes infrastructure, big data computing capacity, data, services and other resources, and the technical security risks in the scheduling process need to be considered in view of the existence of multiple users and multiple resources associated with each other. To this end, container technology can be used to isolate the relevant resources in an independent operating space, so as to achieve the isolation of resources to protect and secure scheduling.

The biggest difficulty in marking reliability problem is the insufficient data support, especially the insufficient sample data that can be used for machine learning and training. Although the amount of data that can be collected is huge, there are still a series of problems: first, the quality of data varies; second, the lack of correlation between data; third, data labeling is extremely difficult, and a large amount of data still relies on manual research and evaluation, and in the process of manual research and evaluation, there is also a lack of effective tools and systems to accumulate the results of research and evaluation, which cannot provide support for the subsequent customization of data product development.

The process scheduling optimization problem includes interactive response time optimization and resource scheduling optimization. When multiple users open the drawing board for visual process scheduling at the same time, it makes the users intuitively feel no lag and jam during the scheduling process as much as possible. For this reason, in-memory scheduling and automatic saving are needed to enable high reliability and reduce the risk of business loss for multiple users when drawing process scheduling tasks.

Thread safety issues include the possibility of data contamination and other unexpected situations in the execution of programs with multiple threads sharing domain data. This can be done by adding locks to shared resources to ensure that each resource variable is occupied by at most one thread at a time. Or let the threads own the resources as well.

7. Conclusion

To address the problems and shortcomings of current data production services, this paper investigates the cross-computing engine scheduling technology for business data production, the domain-oriented business data automatic marking technology and the business data production visualization process orchestration technology. At the same time, in order to improve the overall processing performance of the system, a high concurrency technology solution is studied to utilize the system resources more effectively. Finally, this paper gives a technical risk assessment analysis from four aspects, including computing resource scheduling, marking reliability, process scheduling optimization and thread safety.

Acknowledgment
The successful completion of this study could not have been achieved without the diversified output of Apache and other open source technology communities in terms of technical capabilities and open culture. The continuous sharing of cutting-edge technical ideas by a large number of researchers has enabled this technical study to go deeper and deeper, and to achieve as detailed an argument as possible. We would like to express our sincere gratitude.

Conflicts of Interest
All authors disclosed no relevant relationships.

References

[1] Ouyang R.B, Wang Q.Y., Long X.Z. (2016). Design of data service framework based on microservice. *Journal of Huazhong University of Science and Technology (Natural Science)*.S1:127-128.
[2] Zhan W.T., Ai Z.L., Liu L.Z., Li C.B., Gao Z. (2016). A high priority job scheduling scheme based on YARN. *Software*, (03):84-85.
[3] Sun D.W., Zhang G.Y., Zheng W.M. (2014). Big data Streaming Computing: Key technologies and System examples. *Journal of software*, 25(4):839-862.
[4] Neumeyer L., Robbins B., Nair A., et al. (2010). S4: Distributed stream computing platform. *Proc.of the 10th IEEE Int'l Conf.. on Data Mining Workshops*, 170-177.
[5] Niu C.C. (2017). Research on key technologies of big data streaming computing. *Journal of Zaozhuang College*, 02:113-114.
[6] Gao Y. (2012). Keywords Automatic indexing method review. *Electronics World*, (06):120.