The Advance of Distributed Computing Methods

Shuo Chen *

College of Physics and Electronic Engineering, Chongqing Normal University, Chongqing, China

* Corresponding author email: chen0990@nenu.edu.cn

Abstract. As data computing techniques continue to advance and change, distributed computing has become more and more mature and widely used, and it has become an important and effective method for computing data in today's era. Since human beings entered the information age, effective data processing has always been a topic of concern, in the face of complex and huge data, distributed computing has always played its own role, and gradually has a significant impact on other fields such as the Internet of Things, medical care, artificial intelligence and other fields, and has a positive effect on making today's human life more convenient and faster. Starting from the development background of distributed computing, starting from the three typical distributed frameworks of distributed computing, this paper not only introduces distributed computing as a data processing method, but also uses divergent thinking, describes different computing methods, and analyzes them together. In addition, this paper briefly points out the advantages and disadvantages of distributed computing, as well as the challenges and challenges that still exist in distributed computing and puts forward expectations for the future development of distributed computing.

Keywords: Distributed Computing; Cloud Computing; Parallel Computing; Computational Paradigm.

1. Introduction

With the continuous development and evolution of data computing methods, distributed computing has become more and more mature and widely used, and it has become an important and effective method for computing data in today's era. As mentioned in reference [1], distributed computing and systems have always been the basis for carrying important information infrastructure. In this day and age of big data, distributed computing has also had a significant impact on other fields such as the Internet of Things, medical care, artificial intelligence, and other fields, and has a positive role in making today's human life more convenient and faster. Starting from the development background of distributed computing, this paper starts from three typical distributed frameworks of distributed computing, and not only introduces distributed computing itself, which is a data computing technology, but also uses divergent thinking, describes different computing methods, and analyzes them together. In addition, this article also briefly points out the advantages and disadvantages of distributed computing, as well as the challenges and challenges that still exist in distributed computing and puts forward expectations for the future development of distributed computing.

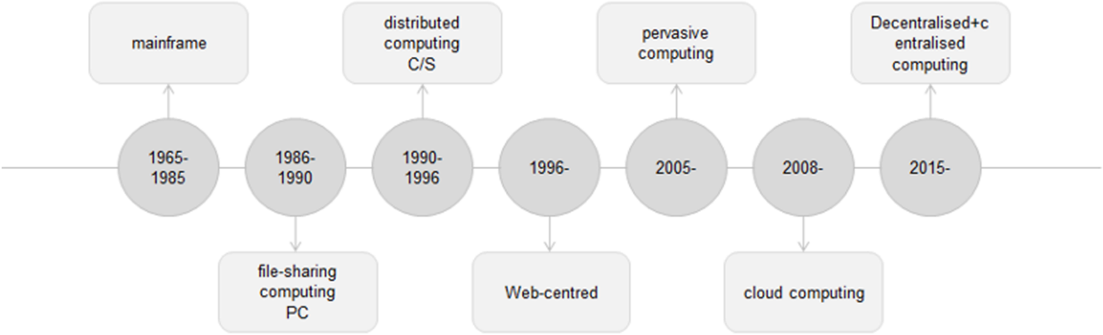


Figure 1. Evolution of different computing modes

Web 2.0 era, the data explosion exponential growth big data distributed computing needs frequently through a single computer memory expansion to enhance computing power can no longer carry the massive amount of data computing distributed computing (Distributed Computation) model is in two or more software to share information this software can be run on the same computer or multiple computers connected by a network. It investigates how to divide a problem that requires a considerable amount of computing power into many small parts, then distributing 51:4, these parts are too many computers for processing and combining the results of these calculations to obtain the result.

The purpose of the distributed computing model is to spread computing work across multiple computers, reducing the load and possible risks of concentrating computing on a single computer to provide high scalability, reliability, manageability, and flexibility.

Distributed computing has several advantages over other algorithms: firstly, scarce resources can be shared. Secondly, distributed computing allows balancing the computational load across multiple computers. Third, programs can be placed on the computer best suited to run them, where sharing scarce resources and balancing the load is one of the core ideas of distributed computing.

This paper mainly starts from the computing paradigm and development platform of distributed computing technology, then introduces the classic frameworks of distributed platforms, Hadoop, Spark, and Storm, and then compares with other computing models (mainly combined with parallel computing), has a deeper understanding of the distributed mode, and finally extends to the combination of distributed computing and other fields, indicating that distributed computing is playing an important role. Finally, the existing advantages and disadvantages of distributed computing are briefly introduced, and the user privacy of distributed computing is mainly prospected.

2. Paradigms and Development Platforms for Distributed Computing

2.1 Paradigm

2.1.1 The Model For Passing Messages

Communication at its most basic level is message conveyance between processes generally consisting of a receiver and a sender, which can be multiple, and the basic operations essential for connection-oriented communication to send and receive. Connect, and disconnected also needed are operations. Development tools based on this paradigm include the socked API and the Message Passing Interface (MPI).

2.1.2 Client/Server Paradigm

The client/server paradigm, referred to as the C/S paradigm, is one of the most commonly used paradigms in the web today It is a simple concept that effectively abstracts service requests from the web with the server side as the service provider passively waiting for requests for services and the client side as the requester sending requests to the server.

2.1.3 Peer to Peer Paradigm

Peer-to-peer network architecture (P2P) is characterized by an Internet system without a central server and relying on exchanges between users. When referring to P2P in reference [2], it says that P2P cannot be given a definition. In P2P, each node is both a server and a node. Nodes cannot communicate with each other and must depend on the exchange of information between groups of users. Furthermore, in P2P, there is no hierarchy. All nodes are equal and have the same functions and responsibilities. Each participant can send messages to another participant and receive messages sent by other participants.

2.1.4 Message System Paradigm

The principle of the messaging system paradigm is simple. The paradigm centralizes the data sent by the sender into the messaging system. It forwards it to the receiver, who, as the sender, sends the message out without waiting for a response and can perform other operations immediately after sending. Messaging system paradigms can be divided into two broad categories:1. Peer-to-peer messaging paradigms. 2. Publish/subscribe to messaging paradigms.

2.1.5 Remote Procedure Call Paradigm

Suppose there are two computers, A and B. If process A of computer A wants to obtain request b from computer B, as with a local procedure call, the remote procedure call triggers a preset "action" of a procedure provided by process B. After execution, B returns b to A. After execution. B returns b to a in A.

2.1.6 Distributed Object Paradigm

Distributed object technology is a cross-platform cross-language, object-based distributed computing technology in a distributed environment that allows object users to access any helpful object on the network without knowing where the object is located. Distributed object technologies are the core technologies for building business application frameworks and software artifacts. They are represented by three categories: Microsoft's COM/DCOM/COM+ technologies, Sun's JavaBeans MI, and OMG's CORBA technologies.

2.1.7 Web Services Paradigm

The web service paradigm consists of a service requester, a service provider (object), and a directory service. The network service paradigm works as follows the service provider registers itself with a directory server on the network; when the service requester(process)needs to access the service, it communicates with the directory server while it is working then, when a connected service can work, the directory server provides a read of the directory service work progress about the service: finally, the process uses the reference to interact with the required service. However, it is not yet fully mature, and it is still in a state of experimentation and development. This is also noted in references [7].

2.1.8 Mobile Agent Paradigm

A mobile agent is a program or object that can be moved. As shown in Figure 2-13, in the mobile agent paradigm, an agent starts from a source host and then automatically moves between hosts on the network according to the execution route it carries with itself. On each host, the agent can access the data it needs and perform the necessary work to do its job.

2.1.9 Cloud Service Paradigm

Three service models have been identified by the National Institute of Standards and Technology (NIST) for cloud computing: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The three of them have different characteristics.

2.2 Several Development Platforms

(1) Coms: a series of publicly available protocols and software designed for computers around the world to be able to work together" With its various research projects in distributed computing can be created. It was officially launched on 21 September 2000.

(2) Sun Grid Engine: Searches the remaining resources of Sun Solaris systems on the LAN and applies them to distributed computing projects.

(3) Dcom: is a framework for C programs and archives that lets you convert your usual time-consuming and difficult mathematical programs into distributed computing programs

(4) The Dispense Package is a free and available software package that allows you to build a distributed computing project quickly and easily on the Internet

(5) DOGMA: is a search engine developed at Brigham Young University. It is Java-based and provides a web-based user interface for distributed computing, supporting firewalls and proxies.

3. Three Major Distributed Frameworks

The main popular distributed computing frameworks are Hadoop MapReduce, Spark Streaming, and Storm; each of these three frameworks has its advantages.

(1) Hadoop MapReduce is the earliest of the three and the most popular distributed computing framework. MapReduce is a typical distributed data computing framework, and its principle is based on GFS-based HDFS distributed file system and HBase data storage system, as shown in Figure 2.

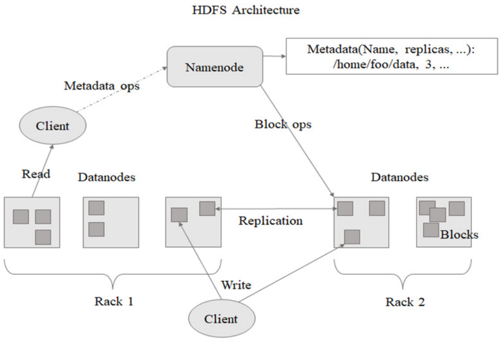


Figure 2. Hadoop's principle framework

(2) Spark is another essential framework with some architectural improvements over Hadoop. Spark works a hundred times faster than Hadoop, and in principle, the reason for this significant difference is that Spark uses memory to place the data it records. Spark may not be able to work continuously in scenarios that require long periods of work, because as soon as there is a power outage, the data stored by Spark will disappear, as shown in Figure 3.

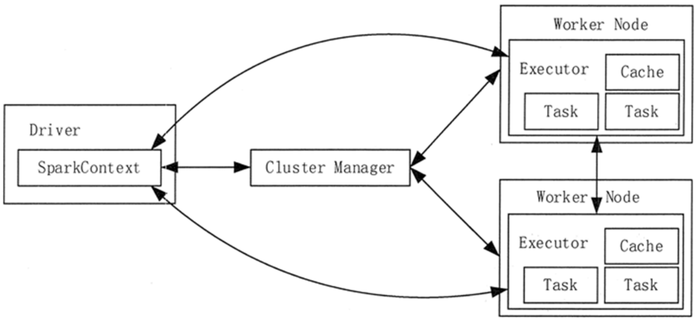


Figure 3. Spark's Principle Framework

(3) Storm framework: this is a framework that can synchronize data processing, it can be connected with the Internet, read the data on the network in time, and when its work is completed, it can transmit the work results in time, which is its biggest difference: real-time, as shown in Figure 4.

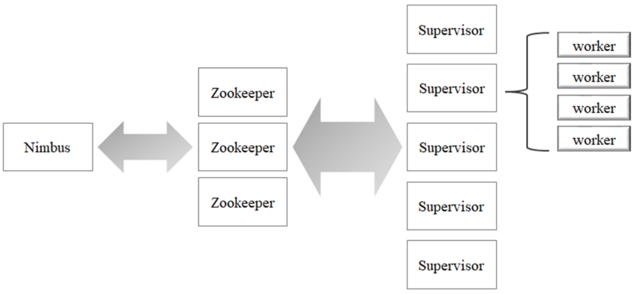
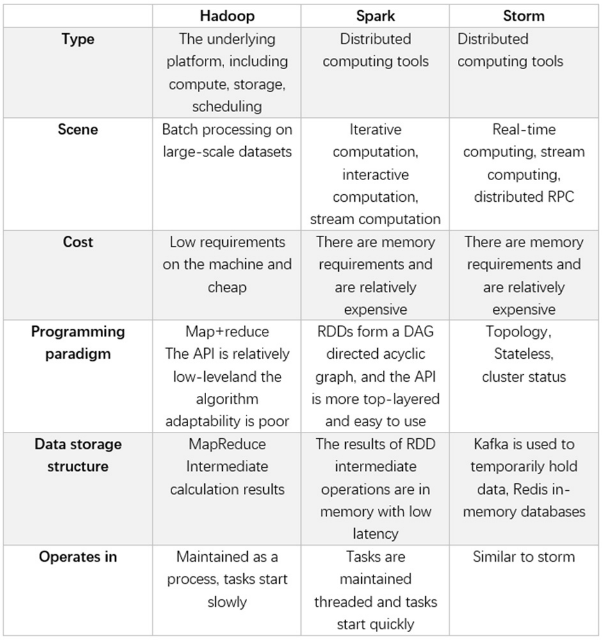


Figure 4. Storm's principal framework

4. Comparison Table

Hadoop, Spark and Storm are currently the three main distribution computing systems, each with the following advantages. These three frameworks are used in different scenarios, when offline data processing, Hadoop and Spark are generally selected, and because Spark is faster, it will be used instead of Hadoop when big data needs to be processed quickly. Storm's ability to transmit incoming data synchronously also shows that it can process data in real time in more everyday situations. Due to the use of different scenarios and needs, the choice of computing framework is also different. The following table lists several aspects. The following table lists some of the attributes of Hadoop, Spark, and Storm, which better demonstrate their advantages, form a starker contrast with each other and are more intuitive, as set out in Table 1.

Table 1. Comparison table of the three distributed frameworks



	Hadoop	Spark	Storm
Type	The underlying platform, including compute, storage, scheduling	Distributed computing tools	Distributed computing tools
Scene	Batch processing on large-scale datasets	Iterative computation, interactive computation, stream computation	Real-time computing, stream computing, distributed RPC
Cost	Low requirements on the machine and cheap	There are memory requirements and are relatively expensive	There are memory requirements and are relatively expensive
Programming paradigm	Map+reduce The API is relatively low-leveland the algorithm adaptability is poor	RDDs form a DAG directed acyclic graph, and the API is more top-layered and easy to use	Topology, Stateless, cluster status
Data storage structure	MapReduce Intermediate calculation results	The results of RDD intermediate operations are in memory with low latency	Kafka is used to temporarily hold data, Redis in-memory databases
Operates in	Maintained as a process, tasks start slowly	Tasks are maintained threaded and tasks start quickly	Similar to storm

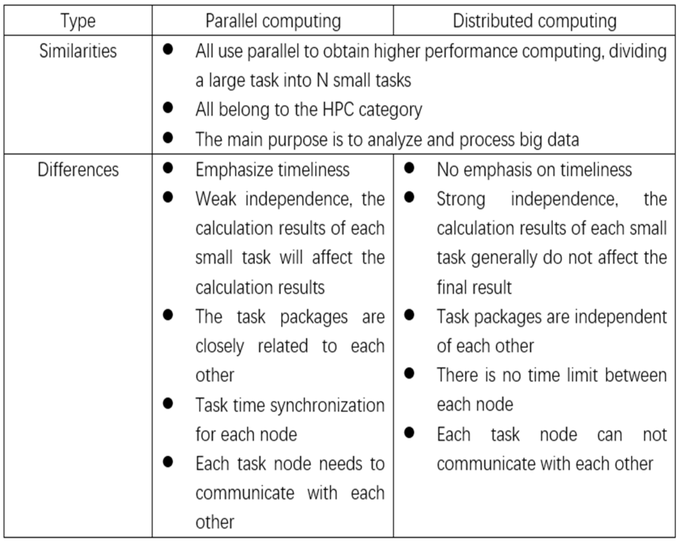
5. Five Types of Computing Technology

In addition to distributed computing, parallel computing, cloud computing, cluster computing, and grid computing, broadly related to distributed computing, are also described below. The aim is to give a clearer picture of distributed computing in different computing models.

The basic idea is to decompose the problem into multiple sub-problems and use different processors to compute each sub-problem simultaneously, thus effectively saving computation time. In the early days, parallel computing could only be done by specially designed mega-parallel computers or clusters of computers. However, massively parallel computers with many processors were often expensive to manufacture and consumed a lot of power, so they were not very practical for general research purposes. The use of graphics processors, or GPUs as they are often called for parallel computing, is becoming a trend. To improve computational efficiency, parallel computing processing problems are generally divided into the following three steps:(1) separating the work into discrete independent parts that help to solve them simultaneously, (2) executing multiple programs simultaneously and promptly means (3) returning the finished results to the host computer for displaying the output after certain processing.

The introduction to distributed computing is not repeated here, and the principles are partially covered in the above article. However, there are similarities between distributed and parallel computing and significant differences, which are presented below in a table for representation, as shown in Table 2 and Figure 5.

Table 2. Similarities and differences between distributed and parallel computing



Type	Parallel computing	Distributed computing
Similarities	● All use parallel to obtain higher performance computing, dividing a large task into N small tasks ● All belong to the HPC category ● The main purpose is to analyze and process big data	
Differences	● Emphasize timeliness ● Weak independence, the calculation results of each small task will affect the calculation results ● The task packages are closely related to each other ● Task time synchronization for each node ● Each task node needs to communicate with each other	● No emphasis on timeliness ● Strong independence, the calculation results of each small task generally do not affect the final result ● Task packages are independent of each other ● There is no time limit between each node ● Each task node can not communicate with each other

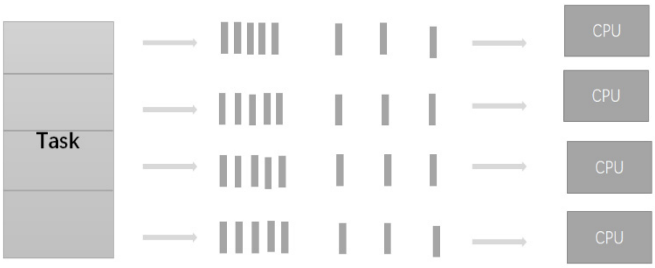


Figure 5. The general principle of parallel computing

Generally speaking, a computer cluster is also a large computer, and the software and hardware within the computer cluster are in a coordinated relationship with each other, and they cooperate with each other to accomplish the computational tasks together. In a computing cluster, each computer can be seen as a node, and the nodes are connected to each other so that they can also coordinate with each other at the same time and work together to accomplish their work, which allows them to have a more efficient working state, and at the same time, the accuracy increases. A clustered system is homogeneous if the architecture of the computers in the cluster system is similar, and vice versa. If the clustered computers are classified by function, they can be divided into the following three categories: high availability (HA) clusters load balancing clusters, high performance computing clusters, and grid computing.

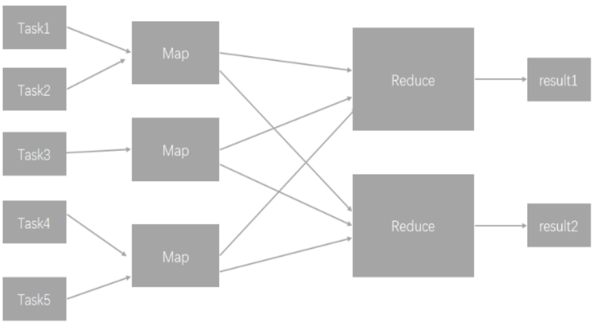


Figure 6. MapReduce-A typical distributed framework

Grid computing is a subset of distributed computing, which is essentially distributed computing, and it also has many similarities to cluster computing. Grid computing is a technology that has evolved over time, and each person has given a different definition to grid computing. In fact, one can give an explanation of grid computing like this: using grid computing technology to put a group of servers, storage systems and networks together to work, allowing them to produce higher efficiency and better results. Grid computing is like a huge virtual computing system for end users or applications. One of the differences between grid computing and distributed computing is that one of the main features of grid computing is that it can process data across administrative domains, something that distributed computing cannot do and something that traditional computer clusters do not have.

Cloud computing is the latest new concept to start it is not just a computer concept such as computing, but also an operational service concept now. This refers to the development of distributed computing, parallel computing and grid computing or a commercial implementation of these concepts. Cloud computing is also a kind of distributed computing, which can share IT resources through virtualization. It is not a brand-new technology; it is a technology to apply network with a brand-new concept. Its core idea is to unify management and scheduling of computing, storage, network, software and other resources through the network, to achieve resource integration and configuration optimization, and to meet the various needs of different users in the form of services to obtain and expand at any time, use and pay on demand, and minimize costs.

6. Integration of Distributed Computing with Other Technologies

(1) IoT field. The Internet of Things (IOT) can interconnect everyday objects with networks through sensors, giving them the "intelligence" to better serve humanity. IOT uses communication technology and RFID to collect data in real time through sensors, and a huge number of sensors will collect even more data and information in these everyday devices, which will be stored in the cloud, and processing these data makes IOT inseparable from distributed computing or other computing frameworks. Likewise, IoT will be combined with AI for better data collection, which makes it more efficient. Today, IoT has been applied in many places in human life, whether it is in smart retail, smart cities, smart agriculture, telemedicine, or in smart watches, smart bicycles, smart refrigerators, and other items that can be found everywhere, IoT is pervasive.

(2) Artificial intelligence, one of the hottest technologies of our time, lies at the heart of AI, which also makes it closely integrated with distributed computing, and only with an excellent set of algorithms can true AI become a reality. Artificial intelligence includes many elements, just like natural language processing, image recognition, language recognition, robotics, etc. Artificial intelligence is still in the age of development, but again, it can already be seen in our daily lives. Intelligence of artificial intelligence is the wisdom that is analyzed, calculated, and summarized through big data.

(3) Medical field. Some relevant examples of distributed computing are not only used in management systems in the medical sector but have also been successfully applied in current pathological research, the study of protein folding, erroneous aggregation and associated diseases (Folding @home), the search for effective drugs against cancer (United Devices).

7. Challenges of Distributed Computing Technology

(1) Compatibility issues. Almost all distributed computing technologies currently do not have complete and unified standards, although work has started in this area. The lack of standards makes the research on distributed computing technologies scattered, and it is difficult to form a stable research direction, thus restricting the development of distributed computing technologies to a great extent. As a result, interoperability, and interconnection compatibility between each other is a huge problem.

(2) Domain issues. Despite the fact that distributed technology has existed for a long time, its promotion and application are still developing. In many areas, there are still gaps.

(3) Heterogeneous issues. Nowadays, the network is a heterogeneous environment, and distributed computing technology must first solve the interoperability problem of heterogeneous environments. The first task in solving the problem of interoperability in heterogeneous environments is how to identify each other. At present, it is impossible to require all resources to be described in the same way, nor is there a way to intelligently identify these resources, which leads to any distributed computing technology being used only within a certain range.

(4) Security issues. The biggest challenge facing distributed computing technology is the growing size of the network, and the security aspects of the whole platform become extremely problematic.

8. Perspectives on the Future of Distributed Computing Technologies

The emergence of cloud computing technology has brought the grid storage and distributed computing model to fruition. All information data can be managed and shared in a unified manner, thus improving the ability to gather and organize information. Thanks to the rapid development of the Internet era, people's lives are becoming increasingly intelligent, and the data that needs to be processed is becoming increasingly complex, so people are concerned not only about the technology itself, but privacy and security are becoming a key topic. In the future development path of distributed computing, apart from purely technical issues, more attention should be paid to the privacy and security of users, as mentioned in the references, which is believed to become a core competition point of distributed computing in the future [10].

9. Conclusion

There is no doubt that distributed computing plays a crucial role in today's human life, enabling true sharing of information and breaking the spatial barriers between data sharing, while playing a vital role in the field of intelligent technologies led by the Internet today. The purpose of this paper is to introduce some discrete elements of distributed computing while providing a proper introduction to the principles of distributed computing technology. At the same time, some relevant suggestions are made in the hope that distributed computing can keep pace with the times and be improved and developed with the progress and development of the times, on top of which, as mentioned above, distributed computing should develop toward a more secure direction and effectively protect users' privacy. In addition, when distributed computing is secure enough, it is hoped that distributed computing can also be applied in military, financial and other more important fields.

References

[1] Liao S.F., Guo D.K. and Ye R. Reservations. Preface to the topic of emerging distributed computing technologies and systems[J] Computer Science2022,49(03):1-2.

[2] Zhou, X. F., Wang, C. J... A review of distributed computing technologies[J]. Computer Age, 2004(12):3-5+10.

[3] Wang X. Research on privacy protection for distributed computing[D]. Zhejiang University, 2020. DOI: 10. 27461/dcnkigzjdx2020.001427.

[4] Ying J. Y. Research progress on data security of smart grid based on loT technology[J/OL]. Electronic Science and Technology; 1-6[2022-10-17] DOI:1016180/jcnkiissn1007-7820.2023.03.012.

[5] Zhang, D. (2018, October). Big data security and privacy protection. In 8th international conference on management and computer science (ICMCS 2018) (Vol. 77, pp. 275-278). Atlantis Press.

[6] Sun Y. Research on data processing technology of computer information based on big data [Jl. Modern Industrial Economics and Informatization,2022.12(01):112-113+118DOl:1016525/jcnki.14-1362/n. 2022. 01. 039

[7] Haikui W. Overview of Distributed Computing Technology[C]//Proceedings of 2019 9th International Conference on Management and Computer ScienceICMCS2019) Francis Academic Press2019:259-262.

[8] Xiao, H. (2010). Towards parallel and distributed computing in large-scale data mining: A survey. Technical University of Munich, Tech. Rep.

[9] Bawankule, K. L., Dewang, R. K., & Singh, A. K. (2022). A classification framework for straggler mitigation and management in a heterogeneous Hadoop cluster: A state-of-art survey. Journal of King Saud University-Computer and Information Sciences.

[10] Liu X. D. Exploration of data security issues and protection initiatives in big data cloud computing environment [J]. Internet of things technology,2022,12(07): 77-79.DOI: 10.16667/j.issn.2095-1302. 2022. 07. 023.