Distributed Data Platform System Based on Hadoop Platform

Jianwei Guo, Liping Du, Ying Li, Guifen Zhao and Jiang Jiya

Abstract Associated with the proposed rapid development of Web2.0, cloud computing, networking concepts, and technologies of the information age increasingly reflect the characteristics of its "big data." In order to exert the value of large-scale data, data mining technology in many areas of commercial, military, economic, and academic received more and more attention. At the same time, the huge scale of the data is a major challenge to the traditional data mining technology. A combination of data mining and cloud computing is becoming a trend in the industry rely on the robust-processing power provided by cloud computing and other distributed computing platform, and this kind of combination is constantly showing its strong advantages and potential.

Keywords Distributed computing · Hadoop

J. Guo (✉) · L. Du · Y. Li · G. Zhao · J. Jiya
Department of Information Technology, Beijing Municipal Institute of Science
and Technology Information, Beijing 100044, China
e-mail: vipherovip@163.com

L. Du
e-mail: duliping_419@163.com

Y. Li
e-mail: shai_wang@hotmail.com

G. Zhao
e-mail: gfzh@hotmail.com

J. Jiya
e-mail: jiya_jiang@sina.com

S. Patnaik and X. Li (eds.), *Proceedings of International Conference on Computer Science and Information Technology*, Advances in Intelligent Systems and Computing 255, DOI: 10.1007/978-81-322-1759-6_61,

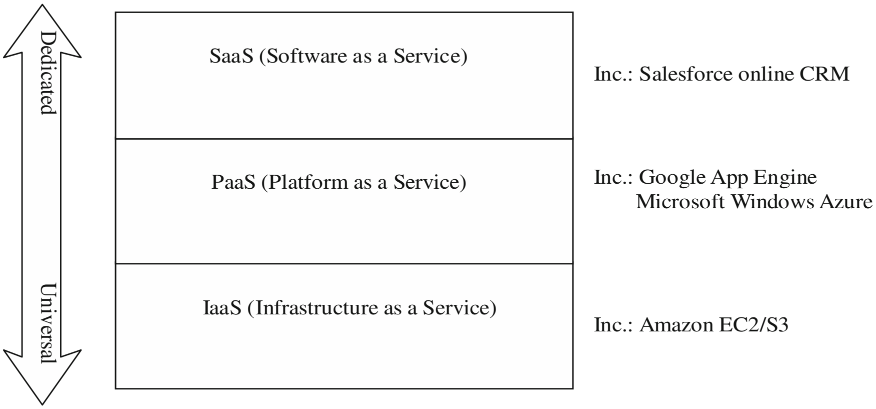
Dedicated
Universal
SaaS (Software as a Service)
PaaS (Platform as a Service)
IaaS (Infrastructure as a Service)
Inc.: Salesforce online CRM
Inc.: Google App Engine
Microsoft Windows Azure
Inc.: Amazon EC2/S3

Fig. 1 The three major service types of cloud computing

1 Introduction

Cloud computing is generally regarded as a commercial computing model that uses the resource pool composed of a large number of computers for computation. This kind of resource pool is vividly called as "cloud" [1], from which the user can utilize the computing power, storage space, or information services according to requirements. The user of cloud computing can dynamically apply for some resources and submit various tasks to the cloud for autonomous management, operation, and maintenance by cloud services. The user and application developer can ignore the detailed distribution at the bottom layer and focus more on implementation of tasks to raise efficiency, reduce the costs, and promote technological innovation. The resource pool of cloud computing is also virtualized as computing and storage resources. Different resources can be dynamically allocated and organized according to demand, and the resources applied by the user can be recovered and reused by the system. This mode of operation can fully utilize the computing resources and improve the service quality.

Cloud computing is the development of parallel computing, distributed computing, and grid computing, or the commercial implementation of the computing concepts. It combines the concepts of virtualization and utility computing and provides a series of services from hardware to software. The three major service types of cloud computing are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), as shown in Fig. 1.

Hadoop [2] is an open source project of Apache foundation that provides the software framework for distributed computing environment, uses Hadoop Distributed File System (HDFS) and MapReduce parallel programming model as the core technology, integrates database, cloud computing management, machine learning, and other platforms, and gradually becomes the standard platform for application and research of cloud computing in the industrial and academic circles. Currently, Hadoop is widely used in Facebook, Twitter, Yahoo!, and other famous companies and runs well in large-scale computer cluster with millions of computational nodes.

Hadoop was originated from the open source search engine Apache Nutch, one of the subprojects of Apache Lucene launched in 2002 [3]. In order to adapt to the sharp rise of data size, raise the data-handling capacity and ensure the search speed and accuracy of Nutch, an efficient distributed computation structure was in urgent need. In 2004, Google released the parallel data processing technology—MapReduce, one of its key technologies at the symposium on Operating System Design and Implementation, and published a thesis titled "MapReduce: Simplified Data Processing on Large Clusters [4]." At that time, Doug Cutting, the responsible person of Apache Nutch, saw the opportunity, lead his team to develop the open source MapReduce computation framework, combined it with the Nutch Distribution File System (NDFS), and integrated into the foundation platform of Nutch search engine. In February 2006, it was separated and became an independent project of Apache named as Hadoop. The core technologies of Hadoop are MapReduce parallel programming model and HDFS [5].

2 Structure of Hadoop Platform

The distributed system of Hadoop adopts a "Scale Out" mode to enhance the computing power. A comparative mode is called "Scale Up," which is represented by large stand-alone server. In the past decades, the development of computer and advancement in computing followed the Moore's law. With increase in data size, it was found that the problem of large-scale computing cannot be solved by relying on larger server only, but a new path should be opened, and more attention was paid to scaling out. In the case of Hadoop, scaling out means to organize low-end or commercial machines together and form a dedicated distribution system as shown in Fig. 2.

3 Distributed File System

3.1 Please Name Node and Data Node Architecture

Hadoop adopts the master/slave structure for distributed computation and storage, which includes two types of node, Name Node and Data Node. The two nodes play the master and slave roles, respectively in Fig. 3.

Name Node is at the master terminal of HDFS and plays the master role. Generally, there is only one Name Node in a Hadoop cluster. Name Node acts as the center of data management, but is not used as the hub of data transmission. It is

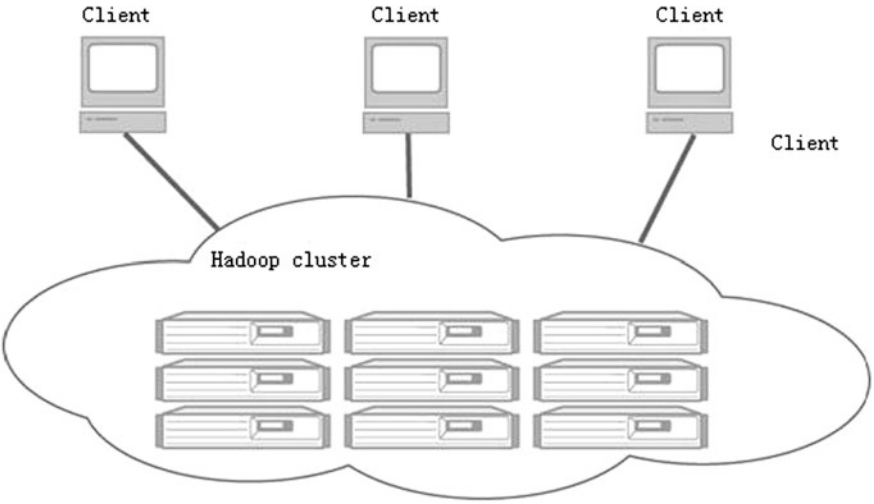

Fig. 2 Interactive mode of Hadoop cluster

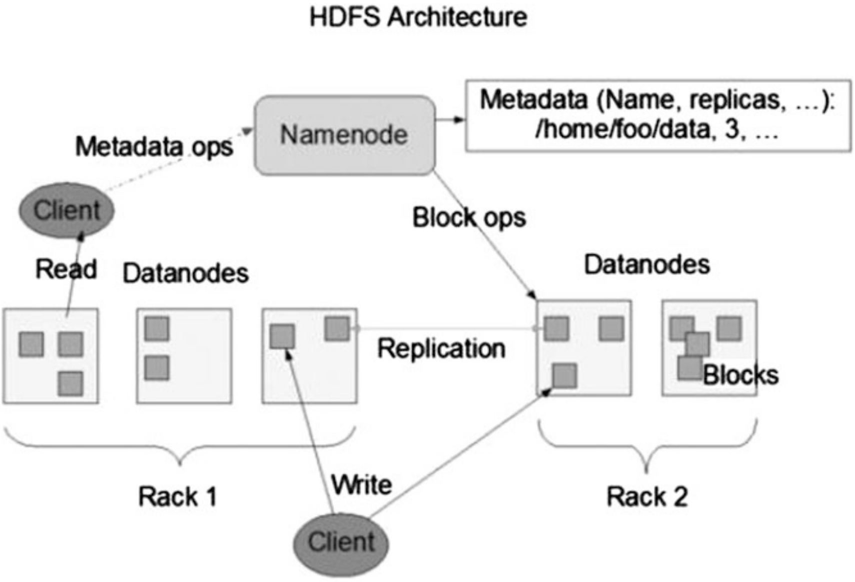

Fig. 3 HDFS architecture

used for management of the namespace of file system, storage of the directory structure of the whole file system and index node of files, and providing metadata service for HDFS.

3.2 Online Publication in Springer Link

All papers will be published in our digital library www.springerlink.com. Only subscribers to Springer's eBook packages or to the electronic book series are able to access the full text PDFs and references of our online publications. Metadata, abstracts, and author e-mail addresses are freely available for all users.

4 MapReduce Parallel Programming Model

MapReduce is an abstract programming model originally developed and used by Google, which can solve the problem of data-intensive operation under many big data environments and deliver the programs that are designed in distributed environment and ideal for parallel computing.

5 MapReduce Implementation

The part of key-value pair of intermediate data with sequence or marking information (defaults of authentication key) can usually be hashed, and the data will be transferred to the corresponding Reducer according to the result of hashing. Certainly, Hadoop will be a convenient method for definition of the allocation rule by the user. In the process of implementation, each data output by Mapper will be redirected on the network, and rule judgment of each data will be performed by Partitioner. Data with the same characteristics will be put into the same Reducer for processing. The process of transferring data into the correspnding Reducer is generally called shuffle as shown in Fig. 4.

6 MapReduce Framework

Each request for computation in the Hadoop cluster is called a job. Hadoop is used to complete the job in a distributed environment. Like HDFS of master/slave structure, MapReduce adopts the similar architecture composed of three types of server, JobTracker, TaskTracker, and JobClient. The master process in the MapReduce framework of Hadoop is a Hadoop called JobTracker, which is responsible for management of all jobs under the framework, and as a scheduling core, allocates tasks to various jobs as shown in Fig. 5.

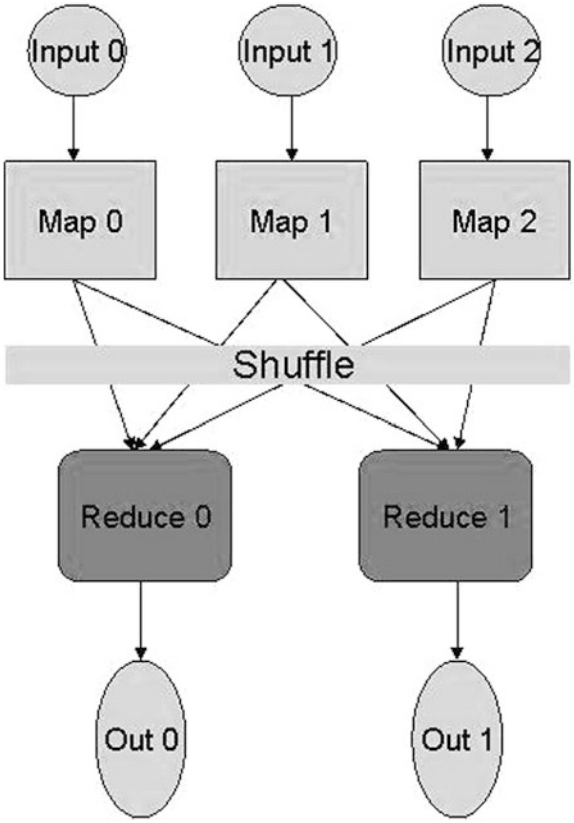

Fig. 4 Shuffle process of MapReduce

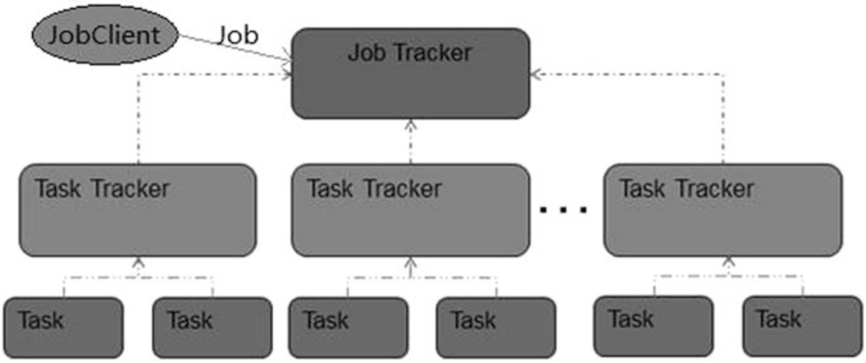

Fig. 5 Basic framework of Hadoop MapReduce

7 Conclusion

This paper has studied on the software architecture of Hadoop for building a distributed platform, focused on the core framework HDFS and the parallel programming model MapReduce, and introduced the design concept of HDFS and the basic method for file storage and management by practice. In particular, the parallel programming model MapReduce is systematically studied for understanding its powerful problem processing and implementation model, and an application template based on MapReduce is written for data mining.

References

1. Xu, Q., Wang, Z.: Cloud Computing: Application Development Practice, p. 12. China Machine Press, China (2011)
2. The Apache^{TM} Hadoop$^{®}$ Project. http://hadoop.apache.org/
3. Lam, C.: Hadoop in Action. Manning Publications Co., New York (2011)
4. Dean, J., Ghemmawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating System Design and Implementation, pp. 137–150. ACM Press, New York (2004)
5. Apache. Hadoop distributed file system, 2010-09-24. http://wiki.apache.org/hadoop/HDFS