Performance Evaluation and Optimization of Big Data Processing Platform Based on Cloud Environment

Zhang Xiaohong, Wang Jing

Jiangxi University of Engineering, Xinyu, 338000, China

Keywords: cloud environment; big data; platform; performance

Abstract: With the advent of the era of big data, high-performance big data processing platforms are urgently needed to deal with ever-increasing types of data, thereby improving the application efficiency of big data in various industries. Based on the strategic considerations of the application and development of big data processing platform, it is included in the cloud environment for research, considering scalability, flexibility, fault tolerance and other indicators for performance evaluation, and proposes targeted performance optimization strategies, to provide technical reference for building a high-performance big data processing platform in the cloud environment.

1. Introduction

In recent years, with the rapid development of Internet technology and its wide application in various industries, the scale of data generated has been expanding, and the era of big data has been ushered in. The explosive growth of data puts higher demands on data processing technology. At present, some big data processing tools such as Hadoop, Spark, Hive, etc. have been developed. Some big data processing platforms such as E-MapReduce, EMR, Azure HDInsight, etc. have also been built. However, in a heterogeneous cloud environment, it is necessary to pay attention to the overall performance of the big data processing platform. The scholars at home and abroad have conducted in-depth research. For example, Li Tianzhuo, Wei Binbin and Yang Chao conducted performance tests on Hive, Spark, Kylin and other [1] big data processing platforms. Wang Lei, Gao Wanling, and Zhan Jianfeng [2] evaluated the data processing platform using the basic operands per second (BOPS) metric. However, none of the above documents incorporates the performance research of big data processing platform into the cloud environment. Only Zhang Chao [3] built a big data processing platform in the cloud environment based on Hadoop cluster, and Huang Wei [4] proposed a performance optimization method for big data processing platform. This provides a strong reference for the research in this paper. From the current research status at home and abroad, the performance evaluation of the big data processing platform mainly considers the performance indicators such as throughput and delay. However, these performance metrics are not enough in a cloud environment. In addition to considering traditional performance indicators, it is also necessary to measure indicators related to cloud computing, such as scalability, flexibility, fault tolerance and reliability, this will be the focus of this paper. In addition, in the cloud environment, in the face of different characteristics of the load and application, the construction of a better performance big data processing platform is more difficult. This is also the problem that this article will solve. The main content of this paper is to try to propose a targeted performance optimization strategy by evaluating the big data processing platform in the cloud environment.

2. Key technologies of big data processing platform

With the development and maturity of Hadoop technology, the big data processing tool has evolved into an ecosystem. The Hadoop related technology architecture is shown in Figure 1. As a mature distributed file system, HDFS is widely used in the big data ecosystem. HBase is a set of column storage systems built on HDFS that can be used for large-scale structured data storage and supports horizontal scaling. MapReduce is a big data computing engine for Hadoop, Spark, etc. It provides computing power for big data processing. It uses the divide and conquer method to divide

DOI: 10.25236/eeems.2018.021

large-scale data into small data sets, hand them over to multiple nodes for processing, and finally summarize the intermediate results to output the final result. Hive is a SQL-like big data query tool. Hive compiles SQL statements and converts them into MapReduce programs for processing. The role of Sqoop is to convert the data table into a data file. As a basic component of Hadoop, YARN solves the problems of the first generation Hadoop architecture and plays an important role in resource scheduling. Zookeeper is a coordinator in the big data ecosystem, mainly for solving issues such as unified naming services, state synchronization services, and distributed cluster management.

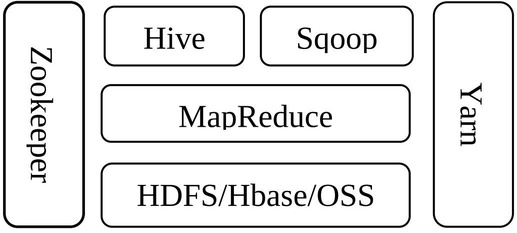

Figure 1 Hadoop technology architecture

Spark is an important computing engine in the big data ecosystem, technology architecture is shown in Figure 2. Spark has several different modes of operation: Local mode, Standalone mode, Mesos mode, and YARN mode. Local mode refers to the mode of running Spark tasks on a local machine, generally used for development testing. Standalone mode is a stand-alone cluster operation mode with a master-slave architecture. This mode of operation uses Spark's own resource scheduling framework and guarantees high availability of the master node through Zookeeper. Mesos mode is a mode of operation that uses the Mesos scheduler as a cluster resource scheduler. The Spark client will connect to Mesos, and users do not need to set up a Spark cluster by themselves. The YARN mode is similar to the Mesos mode. The Spark client can work by connecting to the YARN scheduler without any additional clustering. The YARN mode is the most widely used in practical work. Spark has a wide range of application scenarios. It can process large-scale data sets by calling Spark's API. It can also use SparkSQL and other tools to write SQL statements to process large amounts of data. The final result is to convert SQL statements into Spark tasks.

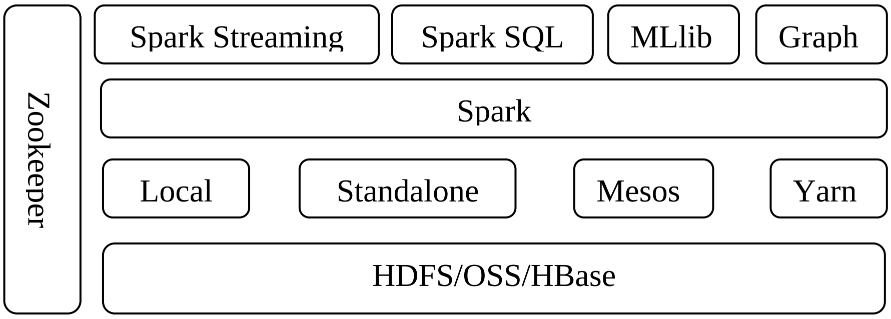

Figure 2 Spark technology architecture

Alibaba Cloud launched the Object Storage Service (OSS). It separates data from storage and is a remote storage system that is easy to use and manage. There are several reasons for the introduction of the OSS system. First, the types of machines required for calculation and storage are different. Storage requires a larger disk, but computing requires a powerful CPU and more memory. Users can choose to have a more powerful CPU and more memory for Hadoop and Spark. Second, when the storage is not enough to save larger data, the traditional HDFS is difficult to expand, and the OSS is easy to expand. Third, the user's data sometimes needs to be used multiple times, and the E-MapReduce feature is used on demand and released. The drawback of this approach is that it is difficult to store data on the cluster, and it needs to be regenerated the next time it is used. OSS avoids this defect. Moreover, OSS can be used for storing large files such as pictures, audio and video, and logs. Various terminal devices, Web site programs, and mobile applications can write or read data directly to the OSS. With BGP bandwidth, OSS can achieve ultra-low latency data downloads, and can also be used with Alibaba Cloud CDN acceleration service to provide the best experience for the update distribution of pictures, audio and video, and mobile applications. Alibaba Cloud provides a platform-independent Restful API for use by users. And the user can use the OSS API as the file path of the HDFS, and copy and migrate files between OSS and HDFS. This paper is based on the Alibaba Cloud E-MapReduce platform to evaluate and analyze the performance of big data processing tools Hadoop and Spark.

3. Big data processing platform performance evaluation

In the performance evaluation of the big data processing platform, a variety of loads are selected, including WordCount, Sort and Grep loads.

(1) WordCount. WordCount is a CPU-intensive load that is used to count the number of occurrences of each word in the payload file. The application has high requirements for the computing power of the cluster, and the CPU usage is relatively high at runtime. The WordCount application has a small amount of output (word statistics) while reading large amounts of data. Its test data is generated by the RandomTextWriter. RandomTextWriter is widely used by users and is an example application of Hadoop that generates text data.

(2) Grep. Grep is similar to WordCount and is also a typical CPU intensive load. It is often used to find out a string in a data or a character that satisfies a regular expression. It is also used to find out how many times a string appears in a file. It has been widely used and Grep has fewer output files. Test data is generated by BigDataBench, and BigDataBench uses Wikipedia terms to generate data.

(3) Sort. Sort is a typical evaluation load on Hadoop and Spark platforms. The difference between Sort and TeraSort is that TeraSort sorts text data, and Sort sorts random binary sequence files.Sort is a typical I/O intensive task, and the Sort application not only has a large number of I/O read operations, but also a large number of I/O write operations. It needs to write the sort result to the file system. Therefore, Sort is often used to test system I/O read and write capabilities, and its test data is generated by BigDataBench.

When users use the big data processing platform, in addition to the difficulty of accurately measuring scalability, there are other challenges. The choice of file storage system affects the performance of the big data processing platform, which is often overlooked by users. When the current cluster is not enough to handle the current load and the cluster needs to be expanded, the cluster expansion mode also affects the processing performance of the platform. There are differences in performance obtained by using multiple low-configuration nodes and enhancing configurations on a small number of nodes. Therefore, evaluating and analyzing the performance of big data processing tools Hadoop and Spark on the Alibaba Cloud E-MapReduce platform can be divided into two parts. 1 is a comparison of the performance of Hadoop and Spark in the local file system HDFS and remote storage system OSS. 2 is to use the evaluation load comparison of different characteristics to evaluate the scalable performance of Hadoop and Spark on Alibaba Cloud E-MapReduce. In the evaluation process, the following matters need to be noted:

(1) On the measurement scenario, two scenarios are selected: performance metrics of the big data processing platform on different file systems and performance metrics of the big data processing platform in different extension modes.

(2) In the evaluation load, three different evaluation loads were selected: WordCount, Sort, and Grep.

(3) In terms of data size, three different data sizes were selected for each load application. Therefore, the performance impact due to different data scales can be shielded, and the real performance of the big data processing platform can be measured more accurately.

(4) In the performance measurement of HDFS and OSS, in order to ensure accurate measurement, 2, 4, 8 and 16 data nodes are selected respectively on the cluster size.

(5) Due to the volatility of the cloud platform, all measurement results are stable data selected after multiple measurements.

(6) Due to the different parameter settings of Hadoop and Spark, the performance difference will

be caused. In order to shield the influence of parameters, the parameter settings are consistent in the evaluation.

4. Big data processing platform performance optimization strategy

Through performance evaluation, it is found that users have some key factors affecting performance when dealing with big data. For example, cluster configuration, network factors, parameter settings, disk I/O performance, and so on. Faced with a variety of influencing factors, it is difficult to accurately locate performance bottlenecks, and this is one of the reasons why users are difficult to solve performance problems. In the cloud environment, performance is subject to more interference factors, and optimization performance is more difficult. Performance optimization strategies can be taken for these parameters.

(1) Compress the result data of the Map side. The I/O performance of the data shuffling phase can be optimized by compressing the result data of the Map side. Data shuffling is a key factor in performance in big data processing. During the data shuffling process, the Reduce thread needs to pull the output generated by the Map side on the network. Network transmission is one of the key factors affecting performance. Therefore, it is a good strategy for users to compress the results generated by Map in the process of using. By combining and compressing multiple Map results, the amount of data transmitted by the network can be reduced, and the task execution time can be shortened.

(2) Appropriately increase the file block size. Performance evaluation found that file block size can affect the performance of CPU-intensive tasks. For example, on the Hadoop platform, the Map number can be reduced by appropriately increasing the file block size to reduce data shuffling time and improve running performance. On Spark, the size of the file block also has a certain impact on performance.

(3) Configuration tuning. On Hadoop, for I/O-intensive workloads, such as TeraSort, Sort, etc., users can improve the performance of the Reduce process by appropriately increasing the number of threads that pull the Map results from the Reduce side. On Spark, properly increasing the number of virtual cores and memory size on the Executor process can optimize the performance of Spark execution.

5. Conclusion

The evaluation and optimization of big data processing platform performance in cloud environment is directly related to big data processing efficiency. This paper analyzes the specific platform and provides a reference for the evaluation and optimization of big data platform performance. In the actual performance evaluation and optimization process, multiple performance indicators and differentiated operating environments should be considered in order to improve the accuracy of the evaluation results and the effectiveness of the optimization strategy.

Acknowledgement

Science and Technology Research Project of Jiangxi Education Department "Research on Image Enhancement Technology in Mathematical Modeling " (No. GJJ171177).

References

[1] Li Tianzhuo, Wei Binbin, Yang Chao, Yang Xinkai. Research on performance optimization of big data processing platform [J]. Telecommunications Information, 2018, (10): 22-27.

[2] Wang Lei, Gao Wanling, Zhan Jianfeng. BOPS: A performance indicator for data center computing [R]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2017.

[3] Zhang Chao. Research on new big data processing platform under cloud environment [D]. South

China University of Technology, 2016.

[4] Huang Wei. Performance Analysis and Optimization of Cloud Based Big Data Processing Platforms [D]. Hangzhou Dianzi University, 2018.