Research on the reliability of network traffic data collection based on Hadoop

ZONG Feng

Shandong Yingcai University, Shandong, 250104, China

nfxzf@163.com

Keywords: Hadoop; Big data; network traffic; data collection; reliability transmission

Abstract. In this paper, the problem of how to improve the integrity and reliability of the data acquisition system based on Hadoop is introduced, and the distributed network fault detection technology is introduced; And based on this technology, a distributed node monitoring framework (DNMF) is designed, which is suitable for the data acquisition system ,the node fault detection and processing algorithm and node load balancing algorithm are proposed in this framework; At the same time, the load of DNMF is used to balance the load, and the load of some nodes is prevented.

1 Introduction

Mass network data acquisition is the basis of the Hadoop based massive network data processing platform, only to ensure the integrity and correctness of the data collected, based on the analysis results of the data has its value and significance. The collection of massive network data needs to collect data from each key link nodes in the mobile Internet, which is the mobile Internet traffic collected from the domestic carriers. After data collection, it needs to pass through many levels of forwarding nodes; the data is transferred to the data, the data storage area of the massive network data processing platform, which includes the network service quality analysis, mobile network user behavior analysis, user traffic control and so on.

However, in recent years, the rapid development of mobile Internet, the backbone of the operator's backbone network traffic growth is also growing, in the process of collecting data generated within a few minutes will be able to reach the GB level. At the same time, in the process of data transmission, because of the large number of forwarding nodes in the middle, the quality of the network transmission has a certain amount of instability, etc., the data acquisition process is prone to flow data in the transmission process, data transmission error and other issues. Therefore, it is necessary to study the reliability of large scale distributed data acquisition process, and to improve the reliability of data acquisition. The mechanism needs to be able to monitor the whole distributed data acquisition system in a timely and effective manner. When a node has a problem or even a failure, the mechanism can fault tolerance processing in time. At the same time, the mechanism needs to be able to control the traffic flow of the distributed data, to prevent data loss caused by the load of a node, so as to ensure the efficiency, integrity and stability of data transmission.

2 High reliable data acquisition mode based on distributed fault detection mechanism

Data acquisition is the precondition of mass network data processing platform. The data monitoring and acquisition based on the network nodes is the important method to collect the data of mass network data processing platform, which involves a large scale of the network, data

Fund Project: Project of National Natural Science Foundation of China(61501284);The 2015 Key project of statistical research of Shandong Province(KT15090); The 2015 Research on the development of soft science in the national economy and social informationization of Shandong Province(2015EI063);The 2015 Business management research project of Shandong Province(J201525); The 2014 annual scientific research project of Shandong Yingcai University(14YCYBZR05); The 2015 Key project of statistical research of Shandong Province(KT15089); The 2015 Key project of statistical research of Shandong Province(KT15087).

transmission needs to flow through multiple nodes. In order to guarantee the integrity and reliability of the collected data, the data acquisition system is designed to ensure the safety and reliability.

At present, based on Internet, the large-scale distributed cooperative technology is widely used, such as grid computing, P2P, the data acquisition system of the Internet of things and so on. This kind of application system is completed by the cooperation of much software, with the characteristics of the distribution of the region, the large number of nodes, high dynamic and complex environment. The traditional fault detection system has been unable to meet the needs of the system fault detection in this complex environment [1]. Therefore, how to improve the reliability of distributed application system is a research hotspot. Distributed data acquisition and application is a typical representative of this kind of distributed application, and it is also a prerequisite for meeting the massive network data processing platform. Research on fault detection in the environment of large-scale distributed data acquisition has very important significance; to ensure the system run reliably, and improve the reliability of data acquisition, at the same time, the nodes of the diverse, dynamic and highly dynamic, distributed fault detection mechanism research has huge application space and practical value.

3 Research on distributed network fault detection technology

The basic principle of fault detection technology is the use of external forces to intervene to confirm the existence of a system failure and the final detection of the possible occurrence of fault location. In 1971, Dr. Beard, the Massachusetts Institute of Technology, proposed the concept of fault detection in his doctoral dissertation [2], followed by Wilsky and Himmelblau, respectively, published a review of the fault diagnosis and academic works, this area has been an unprecedented development.

Fault detection technology applied to distributed systems is the first to consider the scalability and detection efficiency of fault detection in large-scale distributed applications [3]. Distributed fault detection based on hierarchical mechanism is an effective method to improve the detection efficiency and has good scalability in distributed systems. Different from the traditional fault detection technology, distributed fault detection technology in the application of node fault detection, it is also the need to focus on the problem of network fault detection in distributed systems. In distributed application system, the location of the nodes, the diversity of the terminal types, the instability of the network transmission and so on, will result in the probability of the occurrence of the distributed application system greatly increased [4-6].

Current research shows that detection technology based on gossip protocol and hierarchical gossip protocol can effectively on distributed system fault detection, and has good scalability, and improve the fault detection accuracy at the same time, the efficiency of gossip based fault detection technology will have a certain degree of decline, near real time fault detection cannot be realized.

4 High reliability mass data acquisition framework

4.1 Data acquisition framework topology

Data acquisition is based on the Hadoop's massive network data processing platform, which is based on the network traffic monitoring equipment to collect the network traffic, and to store the data in the form of a stream. Traffic monitoring equipment collector is mainly deployed in the metropolitan area network traffic monitoring equipment, such as the key nodes of the export link, the provincial network export link, the backbone network interconnection link, etc... Collector will capture the data transmitted to the various transmission channel, Channel to receive the mobile Internet traffic data for preliminary processing, according to the type of data transmitted to different Uploader, each Uploader to receive the same type of data collection and eventually transferred to HDFS. The topology structure of the data acquisition framework is divided into the following parts.

(1) Data acquisition layer. Collector is used to monitor and collect relevant data in the network, and the traffic monitoring equipment Collector is around the core of the network. The monitoring equipment is used to monitor the flow information flowing through all the links in the network. Collector collects raw data information and stores it in the form of fixed size, which is stored in the

physical storage medium, and then transmits the data center of mass network data processing platform. And the key data information of the original flow, then transmitted to the corresponding Channel in real time, directly based on the network transmission converge to the massive network data processing platform. In the transmission process, each file as a transaction unit, only when the transaction is fully transferred to the Channel, that is completed a complete transmission, the data will be removed from the Collector.

(2) Data forwarding layer. After receiving the data from Collector, the transmission channel of Channel is classified as a specific classification. The classification method is the requirement of the data processing platform. After data classification, the transmission channel Channel is encapsulated in a transactional approach to the data stream. The Channel is encapsulated as a transaction, which is stored in Channel, and then transferred to a specific Uploader node. When the Uploader node is fully received by the Channel node, the node is then removed to ensure the integrity of data transmission. Each Uploader node collects the data information of a specific type. The Uploader node collects the data from each channel node in order to collect the data. Then the data is transmitted to the HDFS in the form of transaction.

(3) Data storage layer. After data transfer data, the data is stored in the data storage area .The current use of HDFS to store massive network data acquisition framework to collect all kinds of traffic data, for the follow-up data processing and analysis.

4.2 Data acquisition framework monitoring technology

In the framework of the massive data acquisition, data needs to be carried out among different nodes. With the increase of nodes, the probability of failure, failure of the system and the load of nodes will be greatly increased, and these will lead to the loss of data. Therefore, in order to guarantee the data transmission security, improve the fault tolerance and reliability of data transmission, it needs to design a reasonable distributed fault detection mechanism, the data flow through real-time detection, and provide some necessary node data statistics, node configuration management, start standby node application to replace the application of failure. This distributed fault detection mechanism should be able to detect all kinds of faults effectively, in addition to meet the basic requirements of distributed systems for fault detection services, but also need to have a high degree of flexibility, good scalability, lightweight detection, dynamic load balancing, etc.. According to the above requirements, this paper establishes a distributed node monitoring framework, referred to as DNMF.

DNMF's design is based on hierarchical structure of the fault detection technology and Gossip technology, the nodes of the distributed system according to a certain rule, and then monitor node management server to assign the appropriate monitoring nodes to monitor each level, each monitoring node uses Gossip technology to spread the monitoring information, so that each monitoring node to its own node monitoring information to provide services for follow-up. The distributed node monitoring framework is defined as the control node in the distributed system, and then divided by the control node, the location and the network location.

In the framework of mass network data acquisition, the DNMF technology includes monitoring node management, monitoring data communication between nodes and monitoring nodes, monitoring nodes, communication and fault handling, the starting mechanism of standby node application, and the configuration management of monitoring nodes based on HTTP.

DNMF is mainly used to deal with the faults in distributed system, and to achieve load balance, so as to improve the reliability of the distributed system. So in view of the current massive data acquisition system, the paper designs the corresponding fault processing algorithm and node load balancing algorithm to reduce the data loss rate in the high-speed network data acquisition process, in order to improve the reliability of data acquisition. DNMF uses the information transmission mode based on Gossip protocol, so that all the monitoring nodes obtain the global fault detection information of the distributed system. It uses the method of random spread to ensure that information can reach the nodes in the topology; the method has the advantages of wide coverage, low network consumption.

5 Conclusions and Prospect

In this paper, based on the Hadoop of massive network data processing platform, how to improve the integrity and reliability of data acquisition, the paper introduces a distributed network fault detection technology, and based on the technology; design a distributed node monitoring framework, which is suitable for the data acquisition system. At the same time, the load of DNMF is used to balance the load, and the load of some nodes is prevented.

However, in the massive network data processing platform based on Hadoop, the massive network data is in a complete and reliable collection, and the storage and processing of the data scale is much larger than the traditional data processing business, how to store data efficiently and fast processing is an important problem to be further studied and solved.

References

[1] Lin Wen-hui. Research on Key Technologies of Massive Network Data Processing Platform Based on Hadoop [D].Beijing: Beijing University of Posts and Telecommunications, 2014.4.

[2] Beard R V. Failure Accomodation in Linear Systems through Self-Reorganization. Massachusetts Institute of Technology, 1971.

[3] Horita Y, Taura K, Chikayama T,A Scalable and Efficient Self-Organizing Failure Detector for Grid Applications, Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing :IEEE Computer Society, 2005,202-210.

[4] Foster I'Kesselman C, Nick J Met al., Grid Services for Distributed System Integration, Computer, 35 (6) 2002 37-46.

[5] Foster I, Kesselman C, The Grid 2:Blueprint for a New Computing Infrastructure, Access Online via Elsevier, 2003.

[6] He X, Wang Z, Zhou D H, Robust Fault Detection for Networked Systems with Communication Delay and Data Missing, Automatica, 45 (11) 2009 2634-2639.