Research on big data processing of water conservancy automation

Zhaochunlei
Central information of Tianjin University of Technology
email: leichunzhao12@163.com

Abstract: with the continuous development of computer technology, data processing technology continues to bring forth the new, especially with the development of big data, distributed cluster and cloud computing technology, digital water conservancy began to change to smart water conservancy. One of the important technologies to transform from digital water conservancy to smart water conservancy is the processing of water conservancy big data, which is the core technology to realize wisdom. A complete work-flow of big data processing includes data collection and importing, data cleaning and quality control, data management and storage, data analysis and visualization, data modeling and model management. This paper mainly proposes a solution to the real-time streaming big data processing of water conservancy automation, which can effectively process the real-time high frequency streaming big data reported by water conservancy automation equipment.

1.Introduction
With the development of intelligent sensors in water conservancy industry and its gradual but large-scale application, together with the maturity of the Internet of things technology, the data collection capacity of water conservancy industry has been improved continuously, thus forming the big data of water conservancy. The big data of water conservancy is composed of data from water conservancy business and the related industries and fields. Gongqihui [1] et al. put forward that the big data of water conservancy has the characteristics of large data volume, diverse data sources, continuous growth of data, high data value and real-time or quasi real-time requirements. Chen Beiqing [2] et al. think that the big data of water conservancy has the characteristics of large data volume, complex data type and complex calculation process and time consuming.

The traditional water conservancy data processing is based on the internal data of the specialty and department for sampling analysis, while the processing of big data of water conservancy is the overall analysis and processing of the massive data of cross departments, fields and multi dimensions. The big data analysis of water conservancy generally refers to the use of distributed computing cluster to process, mine and visualize the data, and there are two main methods of batch processing and flow processing[3]. MapReduce is a commonly used batch engine at present, but it has a large throughput when processing massive high-frequency real-time data, which can not achieve low latency and is not suitable for real-time processing. Spark streaming is an extension of spark core API, which can realize high throughput and fault tolerant real-time stream data processing[4].

2.RocketMQ
RocketMQ is an open source message middleware from Alibaba, which is designed to learn from Kafka. Open source in 2012 and top Apache project in 2017.
It consists of producer, topic and consumer. The message is passed through topic. Topic stores the logical address of the message. Producer sends the message to the specific topic. Consumer subscribes to topic, actively pulls or passively accepts messages.

Topic is a logical concept. Each logical queue stores a part of the message data, but the message data saved is not the real message data, but the message index pointing to the commit log. The top can be created in cluster mode or through a single broker mode. Each topic created has the same role.

In order to achieve high throughput and high concurrency, usually a topic is assigned to multiple agents. Each agent contains multiple topic partitions, and each topic partition holds the same type of queue.

As the core component of rocketmq high concurrency system, message queuing can help the business system structure improve the development efficiency and system stability.

3. Spark Streaming
As a main model in the field of big data processing, streaming computing is a mainstream stream computing framework as Storm developed by Twitter company, S4 by Yahoo company, Microsoft's Timestream and Spark Streaming by UCBerkeley AMPLab.

Spark streaming real-time streaming big data cluster consists of multiple working nodes, each node runs multiple spark executor and runs related business handler on spark executor. Maven application can be created to process PLC data reported by water conservancy automation, and complete packaging of application and related jar package. Jar package can be submitted to spark cluster node by spark submit command. The cluster node submitting the application is the driver node of the application, and the resource is obtained from cluster manager. The real-time streaming data will be formed in batches according to the time window to form the elastic distributed data set (RDD). Each time, the action of RDD will generate a new job, each job contains multiple tasks. Cluster manager mechanism will dynamically distribute Tasks to the Executor according to the cluster resource allocation. The final processing results of each worker node are collected to the driver node for summary output.

Spark streaming is a distributed streaming computing framework based on discrete data stream Dstream, which is built on spark computing engine. It has high throughput and fault tolerance mechanism for real-time flow data processing. Dstream is composed of time continuous elastic distributed data set RDD sequence. Each RDD contains data flow within a certain time interval, which is immutable and can be recalculated[5].

The fault tolerance of spark streaming is realized by Narrow narrow dependence and shuffle wide dependency with DAG. This fault tolerance is adopted mainly for the reason that the data in RDD is an immutable distributed elastic data set. If RDD is lost in the process of cluster processing, it can be recalculated according to DAG.

Narrow dependency is the partition one-to-one between the parent RDD and the child RDD, and the Map and Union operations are narrow dependency; shuffle dependency means that the correspondence between the parent RDD and the child RDD is not one-to-one, and the ReduceByKey operation is wide dependency.

Spark streaming supports data is acquired from a variety of data sources, such as Kafka, flume, zeromq, rocketmq, etc. Spark streaming reads data from the RocketMQ consumer API through defining interface and allocates it to spark streaming cluster node partition, so as to realize distributed processing of real-time streaming big data[6].

Because of the high frequency report of PLC data in water conservancy automation and the data continuously arriving, the traditional data processing framework of real-time big data processing with real-time reading and large-scale data scale can not meet the needs. Spark streaming can handle the real-time big data and off-line big data in a distributed way[7].

4. the overall structure
The framework design based on RocketMQ and Spark Streaming
Spark divides big data into RDD as the basic data structure of spark, and performs data processing in RDD. Spark schedules cluster resources for calculation according to operation. RDD operations are divided into two types: transformation and action. The transformation operation represents the process of transforming one RDD into another through a series of operations. Transformation does not trigger real computation, only DAG between RDD is established with a directed acyclic process. Action represents the end of a calculation, no longer generates a new RDD, and returns the result to the driver program. Each Action action calls the runjob method of sparkcontext to submit the request to the cluster.

The data source is that the water conservancy automation equipment can write the collected data by the Netty server to the RocketMQ producer interface by the data acquisition software of PLC through the collection and operation data of the lower computer to the rocketmq cluster node[8]. The RocketMQ cluster node writes the data continuously reported in real time into the corresponding topic according to the PLC data of different alarm types. The data destination after relevant business processing is divided into three categories. The first is that the real-time alarm information can be actively sent to the alarm system SMS (such as mobile, Unicom, Telecom, Netcom, alibabacloud, wechat, Tencent, etc.) through the message system SMS (such as mobile, Unicom, telecom, Netcom, alibabacloud, wechat, Tencent, etc.) The information is sent to the relevant person in charge in time, so that the alarm can respond quickly and timely; the second is to store the PLC data reported in real time into the Elasticsearch database. Because of its strong data retrieval ability, the system has high timeliness in data query; The third is to transfer the data of complex business processing to spark streaming for cluster operation, and persist the processing results to the time series database and relational database. The database can be used as the result database reported by water conservancy automation, which provides support for the system related business applications.

5. Spark streaming tuning
When the application of real-time big data flow type of water conservancy automation was deployed in the beginning of the cluster, there were some problems such as slow operation, high resource consumption and instability of application system. At this time, the cluster needs to be optimized to achieve the best performance. Tuning is a very specific thing, different application and scene are different in optimization methods, and there is no uniform standard for tuning.

5.1.parallelism of reception and processing
In order to improve the parallelism of data receiving and processing, the number of data segments should be considered. Only one Partition can be processed for each task. If the number of parts is too small, the data quantity for each Partition is too large, resulting in too much memory pressure, and the calculation ability of Executor cannot be fully utilized; if there are too many parts, it will lead to too many pieces and low execution efficiency.

5.2.data serialization
Data serialization has a significant impact on cluster performance in distributed data processing. This paper mainly uses the Java object input stream framework ObjectOutputStream Framework to serialize, and implements more detailed control serialization by using Java.io. External interface.

5.3.batch interval setting
The stability of data processing is directly affected by whether the frequency of real-time data reported by PLC in water conservancy automation can keep stable with the processing speed of cluster. If the frequency of PLC data reported is much higher than the cluster processing speed, data backlog will be generated, and the reported PLC data is too low than the cluster processing speed, waste of calculation happens. Reasonable batch time interval setting makes data reporting and processing maintain a relatively stable state, thus ensuring the stability of the system operation, otherwise, the delay will continuously increase and the system operation go unstable.
6. Conclusion
The experimental test has proved that the construction of the distributed cluster and the use of real-time flow big data processing technology can realize the efficient processing of real-time flow big data of water conservancy automation. At the beginning of the cluster running, the data processing efficiency of the cluster is very low, the CPU and memory occupancy rate are high, but the efficiency of data processing has been greatly improved after the cluster is tuned for many times, and the experimental design goal is achieved. There may be some shortcomings in the automatic flow big data processing of water conservancy, which needs further transformation, expansion and optimization in practice to meet the actual needs of the project construction.

reference:
[1] Gongqihui, Liu Wei, Li Kun, et al. The construction of water conservancy data center based on big data [C] / / paper collection of the third China water conservancy information and digital water technology forum, Nanjing: Hehai University Press, 2015: 243-248
[2] Chenbeiqing, Tan Debao, Tian Xuedong, et al. Discussion on the application of big data technology in water conservancy industry [J]. Journal of the Yangtze River academy of Sciences, 2016, 33 (11): 59-62
[3] Chenjunfei, Dengmenghua, Wanghuimin. Summary of water conservancy big data research [J]. Progress in water science, 2017 (7): 622-631
[4] Yangxiuhui. Research on real-time analysis system of urban traffic congestion area based on Hadoop [D]. Wuhan: Wuhan University of technology, 2016
[5] Tan Liang, Zhou Jing. Real time traffic data processing platform based on spark streaming [J]. Computer system application, 2018 (10): 133-139
[6] Yangxiuhui. Research on real-time analysis system of urban traffic congestion area based on Hadoop [D]. Wuhan: Wuhan University of technology, 2016
[7] Peiguocai. Flow computing and its application in real-time marketing of Telecommunications [J]. Information communication, 2018 (3): 239-241
[8] Zhuyoupu, Lu Jun. design and implementation of distributed energy efficiency management platform based on Kafka [J]. Computer and digital engineering, 2018 (12): 2620-2623