Pisco Med Publishing

Original Research Article

Design and Implementation of Real Time Data Processing System Based on Spark Streaming

Huang He

Guangdong Innovative Techical College, Dongguan City Guangdong Province,523960 China

Abstract: With the rapid development of Internet technology, people have more abundant ways to obtain information through the Internet.More and more information is spread in the network, which will produce a large amount of data.The big data type is more complicated. In order to implement data implementation, the industry is generally combined with different data types and business scenes, special development and design of different processing subsystems.This method has a large research cost and learning cost, and a unified computing system platform is relatively lacking, which further makes developers unable to maintain and expand the system formed by different sets of technology systems.Based on the above situation, this paper designs a universal real-time data analysis and processing system based on Spark Streaming.

Keywords: Spark Streaming; Real-time data processing; The system design

1. Introduction

With the development of great technological progress, the public attaches more importance to the value contained in data.With big data technology, the public can effectively excavate the value of the value, thereby providing corresponding guidance for life, production and learning.Real-time data is generated and updated all the time, but the public’s processing of that data is inadequate. Therefore, more and more researchers pay attention to effectively improve the efficiency of data processing by using different technologies from the large amount of data acquired.

2. Significance of Spark Streaming real-time data processing system design and implementation

The value in data is very rich. In order to fully mine the value in data, it is necessary to use effective modern tools to mine it, including real-time analysis and data mining.Because the data on the Internet is more complex and diverse, the general solution in the industry is to combine different data types and business scenarios to develop different processing systems, including Hadoop, Storm and Hive, and then use caching, message queues and other ways to connect different links.Although the above practices are applied to specific practice, there is a relatively large research cost and a learning cost, and there is no unified computing platform.Therefore, it is difficult to maintain and expand the system. However, Spark can effectively solve these problems.With the further development of the times, people also have higher requirements for data processing analysis.The problems of different computing modes and frameworks are more obvious. As an emerging distributed computing framework, Spark can replace Hadoop and MapReduce to some extent.

Compared with Spark, the MapReduce programming mode mainly processes a large amount of data offline and can meet the requirements of some real-time analysis scenarios, but the execution efficiency is low. Spark provides relatively fast memory calculation.For machine learning algorithms, iterative training is generally required. The application of Hadoop MapReduce computing mechanism is mainly to repeatedly store intermediate results on disks.In the next calculation, the data needs to be sent to different sub-nodes,this method is not suitable for machine learning algorithms because it requires long practice.Spark is good at iterative calculation and easy to expand. It can incorporate a variety of classical machine learning algorithms, laying a foundation for data mining in a big data environment.Real-time data stream processing means that the data received by the system needs to be processed in a short time, and the value of the data will decrease as time goes by.Although the Storm distributed flow calculation framework is also based on memory calculation.At the same time, it is also possible to improve data processing capabilities, but the disadvantage is that it can only be applied to a single environment.Spark Steaming solution proposed based on Spark can process data stream in a shorter time interval. It can be said that it is a quasi-real-time system, which can conduct streaming processing for real-time data stream, it has good scalability and has high fault tolerance and throughput.

3. Design of real-time data processing system based on Spark streaming

Data access, transmission, calculation verification and storage are the main processing processes of S real-time data processing system.The first is to collect the data access, the need to transfer the data to the corresponding location, waiting for the data calculation and verification work. Verify the data before it can be stored in the database.

doi: 10.18282/l-e.v10i3.2451

3.1 Access to the data

Data comes from different sources. For real-time data acquisition in the database, you can enable binlog in the database.In order to achieve the data synchronization between the target library and the source library,you can configure the instances• properties profile of deployerde in canal,configure hbase• yml and application • ymi configuration files in the adapter.The principle of implementation is mainly to carry out master - slave replication between databases.If the obtained data is mainly new data in log files, Flume can be used as a tool for collecting, aggregating and transferring massive logs to detect changes in log files. Flume can obtain the new data if the contents of log files are changed.Sink, Channel and Source are the main components of Flume. Sink is used to fetch data from Channel and store it in the corresponding file system, Kafka or database. Channel is mainly used to cache data provided by Source. Source mainly collects data.

3.2 Data transmission

Data access rates are different from data processing rates, so a buffer needs to be inserted between data processing and data access.The buffer needs to be high performance and can be used in real-time event scenarios, and Kafka meets these requirements. The open source stream processing platform under Apache is Kafka, which can effectively process the action stream data generated in users' daily life.Kafka's main operating mode is cluster mode, which can be used to deal with stress between servers or nodes.So the monitoring log file data obtained by Flume can be sent to Topic and cached in the Kafka buffer, prompting the collected data to be eventually applied.

3.3 Check data calculation

MapReduce is generally used to calculate data. However, MapReduce supports reduce and Map operations. Map and Reduce results are written to disks and HDFS respectively.However, MapReduce takes too much operation time, so it cannot be used in real-time computing scenarios.Spark is mainly used for memory calculation. Multiple calculation has certain advantages, and Spark has many RDD operations.Therefore, S is more suitable for the data calculation verification part.

3.4 Store data

The real-time data import system is mainly after Sflow processing, stores corresponding data in HBASE.HBase provides massive storage, high concurrency, easy expansion, and low cost.At the same time, HBase can also store multiple different versions of data, which has a higher data query response speed.Because it has a relatively special addressing mode, it can access the Timestamp URL obtained by metadata and also cache metadata related information.Therefore, the corresponding fast feature of HBase is shown. Rowkey design is very important in HBase. There are many regions in HBase. Each region has its own stopRowKey and startRowKey. Thus further leading to the dielectric properties of the situation.

4. Implementation of real-time data processing system based on Spark Streaming

The real-time data processing system uses Spark Streaming to invoke data in Kafka and process data in real time. The processed data structure is stored in HBase.Take Flume as an example. As mentioned above, Sink, Channel and Source are the main components of Flume, and the three components need to be configured and connected in series.At the same time, Kafka is set to the type of Flume Sink, and also pay attention to serialization mode settings to be transferred to topic.The data stored in Topic needs to create data collected by consumer consumption.Consumers are primarily based on Spark Steaming, with Spark Steaming to verify the calculation of collected data, it needs to connect Kafka and Spark Streaming.After reading the Topic data, you can verify the calculation.Separate the data string array by means of a space,if the data value in the array or the array length does not match the demand, it is not possible to store the corresponding data.like:WordCount, after reading the topic data, it needs to store it in an array separated by Spaces, then convert it into a tuple, and finally aggregate the tuple with the same key value to specify the number of occurrences of words.Data after processing, stored in HBase, then create a corresponding table and insert the data that has been calculated into the HBASE table. The first batch of data needs to be processed for the second batch of data. HBase can store data of multiple versions if key data is the same.Therefore, after inserting the database in the same Key data, it is not possible to implement the accumulated implementation. Therefore, it is necessary to delete the same data in the original table, and insert the sum of the new value and the original value. Therefore, you need to get all the RowKey acquisitions in the table. During the data insertion, you need to first detect the same RowKey.If the data does not exist, insert it. If the same data exists, record the original value and delete the data.

5. Conclusion

With the development of big data technology, enterprises attach more importance to the value in real-time data.The real-time data processing system based on Spark Streaming can access, transmit, verify, calculate and store real-time data.Therefore, for the company, in order to fully excavate the hidden value of the data, the corresponding data acquisition, application platform can be established, and the real-time data processing system can be introduced and applied.However, the system also has some deficiencies, which need to be further optimized and perfected.

References:

[1] Shi Wei. Design and Implementation of Real Time Data Processing System Based on Spark Stream [J]. Modern Information Technology, 2020, 4 (20): 10-12.

[2] Li Tianxi. Research and Implementation of Experimental Data Processing System based on Spark Streaming [D]. Shaanxi: Xidian University,2015.

[3] Xu Yifeng, Feng Dajun, ZHANG Hanwen, et al. Application of electronic technique,2017,43(9):98-100,105.

[4] Zhang Muwei, Dong Feng. Design of real-time data processing system for infrared camera based on FPGA [J]. Infrared,2016,37(4):1-6,32.