An analytical method of electric power consumers behavior based on Storm

Dewen Wang[a], Liping Yang[b]

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, Hebei Province, China

[a]wdewen@gmail.com, [b]1062969800@qq.com

Keywords: Large data ,Storm , real time ,clustering

Abstract. The smart grid is one of the important fields of large data applications. The study of power behavior of the user in the big data environment has important significance for the demand side management ,load forecasting and so on . Aiming at the problem of insufficient real-time response capability of massive power data, the introduction of distributed real-time computing platform Storm is used to analyze the user's behavior. In this paper, the k-means algorithm is implemented under the Storm framework. Through the experimental test and comparison analysis, it is verified that the Storm computing system can improve the real-time processing of the data, and can deal with large-scale data.

Introduction

In recent years, with the rapid development of smart grid, smart meter and acquisition terminal has been widely used. Collecting and dealing with the data in everyday is exponential growth, more structure type, interactive strongly, and gradually form the power user data [1]. The vast amount of electricity used in data hiding the user's behavior habits, mining the hidden value, for guiding the user's electricity behavior and energy saving, has important significance.

Clustering analysis [2] is a common method in data mining, which can deal with a large amount of data and get the distribution characteristics of the data. It has been applied to the analysis of the smart grid user's behavior. For example, in[3] self organizing map neural network and hybrid visual clustering technique is used to calculate the characteristic properties of the cluster, which can realize more effective information mining. In [4] the user behavior analysis is achieved based on cloud computing platform and parallel K-means clustering algorithm. However, the method of the above research can not meet the demand of large scale stream data processing.

For the vast amounts of data, the real-time data processing and analysis method is an urgent problem for the smart grid. Storm is an open source distributed real-time computing framework, which can handle a large amount of stream data [5]. Based on Storm, the distributed, high performance and scalable cluster environment can be used to analyze and process the massive data in real time. In summary, this paper will take home users as an example, and put forward the cluster analysis based on Storm platform, which provides a new way for the analysis of power users' behavior.

Analysis of user behavior based on stream computing framework

Stream computing framework Storm

Storm is an open source distributed real-time computing system, and is differert from the Hadoop that developed rapidly in recent years, which is focused on the computational framework of streaming data processing.Storm is mainly used in the field of real-time analysis, such as continuous computing, online machine learning, distributed remote invocation, data extraction, conversion and loading.

There are two kinds of nodes in the Storm cluster, which are the control node and the working node. The control node is run on the Nimbus's background program, which is responsible for

distributing the code inside the cluster, assigning tasks to machines, and monitoring the cluster state. The working node is running a background program known as Supervisor, and the Supervisor will monitor the work assigned to it and start or close the work process as needed. All the status of information about Nimbus and Supervisor stored in Zookeeper, Nimbus conduct task scheduling and allocation according to the state information.

In the Storm framework, the logic of computing tasks is encapsulated into the Topology object, which is called the calculation of the extension. Topology is a graph of different Spout and Bolt connected by data stream, as shown in figure 1. Spout is a message producer of Topology, which usually reads data from external data source, and then sends it to Storm in the form of Tuple. Bolt in the package is the processing logic, after receiving the Tuple can perform filtering, aggregation, computing, function operation, etc. Spout and Bolt is a subscription relationship, can be flexible to achieve data orientation and diversion.

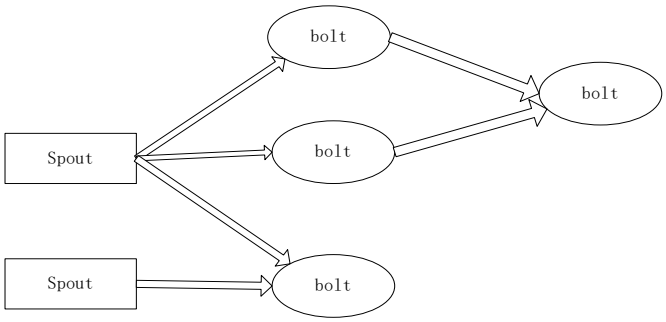

Fig. 1 Schematic diagram of the Topology model

k-mean algorithm theory

K-means algorithm [6] is easy to understand, and it can be run in a distributed environment. The basic idea is that the K instance is selected as the initial cluster center, and then the nearest cluster is calculated. The working process of the algorithm is as follows:

(1) the total number of categories is determined by selecting the K value according to the data set D.

(2) the K instance is selected as the initial center C_1, C_2, C_3, in the data set..... C_k.

(3) This K classes centers and other remaining examples of simple was caculated the Euclidean distance. The kind of distance as a measure of similarity between instances of the class will be the center of high similarity to the class divide, and become a category.

(4) It is to judge whether or not satisfy the number of iterations ,and which is stopped.Otherwise, the objective function is calculated with the error square sum function:

$$S = \sum_{j=1}^{k} \sum_{i=1}^{|D|} \left\| D_i - C_i \right\|^2$$

In function: k is the number of clusters, C_i is the first K center of mass.|D| is the number of the data set .

(5)It is to judge ΔS whether or not meet the threshold . if it is satisfied, the calculation is stopped. Otherwise, it is to execute (6).

(6)It is to calculate the average value of each new class to update the center, and then execute (3).

Implementation of K-means in Storm

In order to better take the advantages of Storm platform, build a fully distributed environment on Linux cluster. According to the standard K-means algorithm flow, this paper will divided the K-means algorithm into four stages based on Storm,data preprocessing, initialization phase, clustering calculation, storage results. Storm framework itself is not responsible for of the preservation of the calculation results, can achieve different Bolt to complete the diversification of storage of the calculation results.It can be directly written to data files, can be stored in a persistent store to the database, can also be achieved through the rapid storage of Redis cache memory, can also be saved to the cloud platform, etc..

In Storm, the Spout and Bolt components are composed of data stream Topology. According to the actual demand, this paper designes as follows, Spout component is the data source of whole framework, and users access data through IRichSpout interface. Kafkaspout read data from the Kafka message queue, and read data into the storm. Splitbolt part the data according to the algorithm , and then send the message to Initbolt by the way of shuffleGrouping, the bolt is responsible for the initialization of the cluster state including the cluster center, cluster and error square sum, and then pass the information Kmeanbolt bolt by the way of shuffleGrouping. In this bolt it is to execute the K-means mean value calculation and classification, and then send the results to Resultbolt by the way of filedsGrouping.

Experiments and results analysis

Experimental environment

In this paper, Six computers is used to build a Storm cluster environment, one of which is running Nimbus and Supervisor, and the other five are running Supervisor, and configure the operating environment that is needed. Experiments were performed under the condition of the cluster without other tasks, and the results were analyzed. Take the actual measurement electricity data of the city residents as the sample, about 2700000 per household.It is to analog data stream on the master node.

Experimental test

In this experiment, it adopts K-means algorithm based on Storm cluster (SK) to execute cluster analysis of sample data, and compare the efficiency of K-means algorithm with Java algorithm. Sample data is limited in the experiment, so the data set is constantly increasing, and then the relationship between the size of the data set change and the time , precision of the clustering is tested . In order to improve the accuracy of the test results, each data is obtained by the average value of three times test. The results of the experiment are shown in Figure 2:

Table 1 results of user clustering analysis

category	correct rate(%)of SK	correct rate(%) of single K-mean
idle housing users	98	95.2
old man family	86.3	74.4
commercial user	98	80.2
office worker and old man	93.2	87
office workers and school family	89.2	71

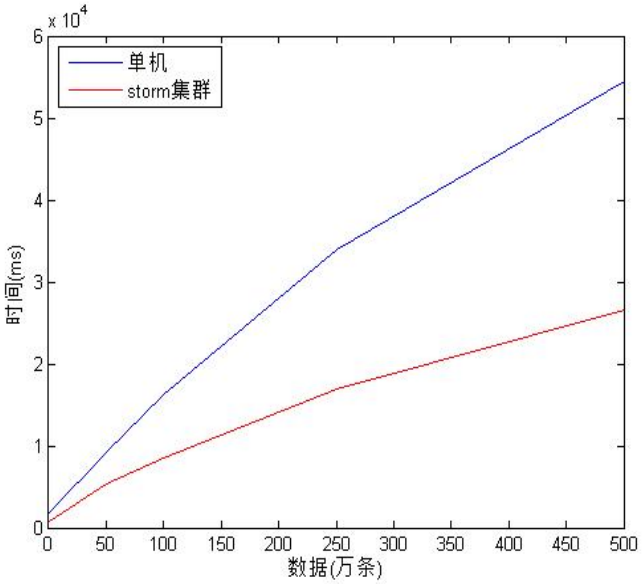

Figure 2 Storm cluster and single run time comparison

From figure 2, it can be seen that the processing of the same data, processing efficiency based on storm cluster is better than single machine about standard K-means, and the program running the average consumption time is shorter.With the increase of data, the processing ability of Storm cluster is gradually emerging, processing time has a greater advantage, high real-time. The reasons for this result are: (1) computing tasks in Strom are divided into different components (spout and bolt). (2) the system based on the Storm cluster version has a shorter latency. At the same time to receive the new data on the old data processing, do not wait for all the data to reach the reprocessing.

Conclusions

In view of the problem of low efficiency about intelligent data analysis, a new method based on Storm is proposed. Under the environment of Storm cluster, the K-means algorithm is designed and implemented under Storm framework, which is used to cluster analysis of user's data. The results show that Storm can be competent for real-time processing of large scale monitoring data, and the cluster is running in good condition and load balance. Storm provides the basis for the power grid real-time flow processing business, and has a broad application prospect in the electric power big data.

References

[1] Song Yaqi，Zhou Guoliang，Zhu Yongli．Present status and challenges of big data processing in smart grid [J]．Power System Technology, 2013, 37(4): 927-935．

[2] Zeng Feng，Zhang Xiaoning．Application of cluster analysis to preventive maintenance scheme design of pavement[J]．Journal of Harbin Institute of Technology，2009，16(4)：581-586．

[3] Liu Youbo，Liu Junyong，Zhao Yan，et al．Calculation of characteristic attributes of consumer aggregations based on multi-objective clustering[J]．Automation of Electric Power Systems，2009，33(19)：46-51(in Chinese)．

[4] Zhang Suxiang，Liu Jianming，Zhao Bingzhen，et al．Cloud computing-based analysis on residential electricity consumption behavior[J]．Power System Technology，2013，37(6)：1542-1546．

[5] Leibiusky J, Eisbruch G, Simonassi D. Getting started with storm [M]. O'Reilly Media, Inc., 2012.

[6] DE AMORIM R C，MIRKIN B．Minkowski metric，feature weighting and anomalous cluster initializing in k-means clustering[J]．Pattern Recognition，2012，45(3)：1061-1075．