Research on Load Query of Grid Data Stream Based on Storm

Xudong Li, Hao Li, Airong Zhang, Dongdong Wang, Xudong Li, Hao Li, Airong Zhang, Dongdong Wang

State Grid Henan Yanshi Power Supply Company, Luoyang, 471911, China

Keywords: Data Flow, Storm, Continuous query, Histogram.

Abstract. In recent years, with the development of data flow applications, various fields required to respond data flow quickly and real-time. Using Storm frame to process data flow, the time of calculation can be shortened significantly. In this paper, Storm platform will be used for continuous queries of data flow. Firstly, this paper has a brief introduction for Storm framework and running process, then, it describes the model of continuous query, and then it introduces two generation algorithm of histogram, last, it experiments for continuous query on Storm cluster. The experimental results obtained are analysed to verify the validity of the data flow continuous query using Storm platform.

Introduction

With the development of computer network technology, it has emerged in a new model that called streaming data in many application areas, such as wireless sensor networks monitoring data, financial transaction data, network monitoring data, location-based services (LBS). Stream data with traditional data patterns are different, which is characterized by continuous uninterrupted data and large scale. In current applications, the stream data needs to be processed in real time, in order to explore their potential information. Continuous query of data stream processing is a typical application. With a continuous query is the arrival, the query results are constantly updated, is driven by the conventional event data query, the query is driven by the continuous stream of data.

Traditional data queries are operating against static and stable data sets. It cannot be directly used to solve fast, unlimited, real-time continuous query applications. Currently, the persons and institutions engaged in the data stream study abroad more and more, more significant achievements have Stanford's STREAM project team to study the general data stream management systems; UIUC research oriented data stream mining system, which developed mainly MAIDS for data stream classification, aggregation and excavation work frequent item sets. Domestic research started late data stream, still in the experimental prototype stage.

Storm framework

Storm is a real open-source distributed computing platform by Twitter in 2013 as a sub-project under Apache. Topology can be sustained by creating Storm, uninterrupted data stream in real time. Due to Storm's work and the definition of continuous query is very fit, and Storm has the characteristics of real-time processing, use it for continuous query, is an effective solution. Storm cluster divided into the main nodes (Master) and work node (Worker). On the master node running a Nimbus process, responsible for operating junction distribute code, assign tasks and monitor task operation, running a Supervisor working process node is responsible for part of the Topology of execution. After Topology submission will run, when the data stream is processed on arrival, unless manually stopped.

Topology is directed acyclic graph composed of a number of spout and blot, between spout and blot through the "Subscribe - send" mechanism to connect with Stream Grouping. Spout and bolt is the most basic stream process primitives. Spout is the generator of message. It reads the data from the external and sends the tuple to the blot. Bolt is the processor of the data, all the process are encapsulated in the bolt. Bolt can also be sent for further processing to other bolt.

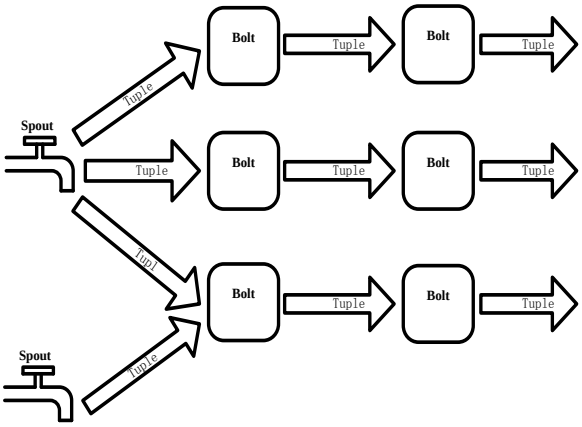

Figure 1 Topology Structure

Topology is a directed acyclic graph. It is the highest level of abstraction in Storm. It is submitted to the cluster to run a topology converting a data flow diagram. Tuple sequences are abstract as stream, it is a value list. Each value in list has a name. Storm tracks each Tuple using ACK mechanism to ensure that each tuple are complete treatment in topology

Continuous query model

Traditional data query is instantaneous when finished will no longer be valid after continuous query as a new query mode, when after being submitted, it has been run, it will continue to deal with the arrival of the new tuple, as data constantly updated stream output new query results. Such as traffic management, the use of continuous query processing for electronic police information, traffic monitoring equipment acquisition, and can effectively avoid traffic jams; transactions in the financial sector, continuous queries can also be used to predict the trend towards price discovery trading opportunities. Continuous query is not like a traditional data query to draw precise query results, which allows the existence of acceptable error for the trend of judgments.

Continuous data stream query range can be divided according to the timing landmark model, sliding window model and snapshot models:

(1) Landmark model: In this model, the range of the data stream query algorithm from start of a particular time stamp t n to the current timestamp all data, the data before a specific time stamp t is not considered.

(2) Sliding window model: In this model, the range of the data stream query algorithm is ended within the first time of the current W interval data timestamp n, W is also known as the sliding window size, with the emergence of new elements to the window before moving, constantly updated range, but the size is always kept at W interval.

(3) Snapshot model: In this model, the query range of data flow algorithm is to define a data between two specific timestamp, take a snapshot, and then check.

Since the landmark model and the sliding window models take into account new data continues to come more in line with the characteristics of streaming data, applications are more widely. This paper selects the sliding window data stream processing model as a continuous query. Sliding window, there are two data streams based on the order defined time-based sliding windows and sliding window definition. Sliding window according to its sliding partition granularity, the former sliding granularity is a tuple, which is the time period granularity slide.

Usually a data stream can be defined as a tuple sequence: x1, x2, ..., xn ..., since the data streams continuously generated, the amount of data, cannot save all be processed, so the algorithm need continuous query stream data again scanned generate summary data, when a new tuple arrives, an outline of the data update. Summary data for continuous query algorithms operate in ensuring the necessary accuracy, the substantial increase in the efficiency of the continuous data stream queries.

Way to build data flow summary data structure also has a variety. There are histograms, sampling, wavelet, hash. This paper selects technical summary data generated histogram data stream. A histogram is the range of attribute a relation R is divided into several intervals, the interval is called

a straight party, then the number of tuples belonging to various statistical straight parties. When R is A relationship attribute query, the available results of a query histogram value estimate. Within the rated range of error estimates can be used as a query result.

Continuous query algorithm based on histogram

Algorithm parameters

A relationship between the assumed value of the distribution properties of R can be a collection of T represents a binary, T = {(a_1, v_1), (a_2, v_2), ..., (a_n, v_n)}. A possible where ai represents the attribute values corresponding to each a_i v_i denotes A property value is equal to the number of ai tuples. With a four-tuple set Z = {(ab_1, ae_1, nai_1, att_1), (ab_2, ae_2, nai_2, att_2), ..., (ab_m, ae_m, nai_m, att_m)} to represent R histograms on the A's . Where [ab_i, ae_i] ($min \le ab_i \le ae_i \le max$) represents a range or histogram, ab_i and ae_i as a starting point and end point of the straight side; nai_i represents the current value of the number of straight party; att_i fall into the histogram indicates the number of tuples within. According to the definition of the relationship can get data on the properties of R A distribution histogram, by definition can know Quads Z must meet the following conditions:

(1) $[ab_1, ae_1] \cap [ab_2, ae_2] \cap ... \cap [ab_m, ae_m] = \Phi$;

(2) $[ab_1, ae_1] \cup [ab_2, ae_2] \cup ... \cup [ab_m, ae_m] = A$;

(3)$\sum_{i=1}^{m} att_i = n$; Here, the relation between the number n R of the instance.

In the process of generating the histogram, the histogram is defined within the error of each are within the acceptable range and thus can be more accurately the distribution of data in the R-response relationship. For different data stream query algorithms need to build a different histogram summary data.

Defines the equivalent error continuous query algorithm

In limited error histogram, for any equivalence query, the results obtained are the size of the error does not exceed the set histogram is generated when the nominal errors. When generating a histogram of each straight side, the recording side of the current maximum and minimum values of the straight, its comparison with the new tuple, if the minimum and maximum distances from the current exceeds the rated error deviation, then the end of the direct, and will set a new straight vi square boundary.

Algorithm 1: define the equivalent error deviation of the histogram generation algorithm

```
Input: T = {(a1, v1), (a2, v2)... (an, vn)}
Output: error is equivalent query histogram deviation of Z.
i=0;Flag=true
for j=1 to n do
  if Flag=True do
i=i+1;low=vj;high=vj;abj=aj;Flag=False;
else
if vj-low>deviation||high-vj>deviation do
aei=aj-1;Flag=True;j=j-1
else
if vj<low do low=vj;
if vj>high do high=vi;
atti=atti+vj
endif
endif
endfor
```

According to the algorithm for data collection T were again scanned, you can generate a summary query error within the deviation of the histogram. Generated using the above algorithm guarantees to meet the equivalent query in the case of error in the least number of straight side, while the number of data streams and query efficiency histograms positive correlation.

Define the scope of continuous query algorithm error In defining error histogram attributes of relation R. A range queries, such as a, b∈ A, and a <b range queries a <r.A <b, guarantee the results obtained by mistake difference is within limits.

When the histogram is generated for each of the straight tuple calculated within histograms ensure queries within the rated error, as straight as possible so that the calculation Side contains more elements.

Algorithm 2: define the scope of continuous query error histogram generation algorithm

```
Input:T={(a1, v1), (a2, v2) ,…,(an, vn)};
Output: Continuous query histogram limited range error Z.
i=0;Flag=true
for j=1 to n do
if Flag=True do
i=i+1;avg(zi)=vj;start=j;Flag=False;
else
avg(zi)=(avg(hi)*(j-start)+vj)/(j-start+1);
k=start;
 if (vk<=avg(zi))
Nsta=i;
Nsum=avg(zi)-vk;
Psun=0;
 Else
Psta=i;
Psum=vk-avg(zi);Nsum=0;
 endif
for k=start+1 to i do
Nsum=max{0,Nsum+avg(zi)-vk};
  Psum=max{0,Psum-avg(zi)+vk};
 if(Nsum>deviation/2||
Psum>deviation/2)
j--;Flag=True;break;
update the histogram parameters;
endif
endfor
```

If the start and end of a range queries are in an inside straight side, the algorithm to ensure that the error does not exceed deviation / 2, if not with a straight side, the guarantee error does not exceed deviation. Two variable used in the algorithm N sum amount and are used to indicate the straight P sum prescription anti- and forward to the cumulative deviation control within a small deviation histogram in deviation / 2.

Experimental simulation and evaluation

Experiments were simulated equivalent query definition Topology and scope of the data stream query. A similar range queries and equivalence queries were conducted for the community and the next equivalent query during range queries. We can get the desired range queries. Experiment data flow simulation program generates a temperature measured by the sensor, the acquisition frequency of 100HZ, with Spout receiving stream data and transmitted tuples scanned to create a histogram calculation Bolt summary data structure, when the continuous Bolt subsequent query. Experiments tested the impact of the size of the data flow velocity and the sliding window on the efficiency of the algorithm.

Experiment sliding window size is set to 20 seconds, and then changing the flow rate of the data stream, respectively in the two continuous query summary histogram algorithms, the time to perform the recording of continuous queries delay. Topology will be submitted to the Storm cluster, using the Storm UI right to monitor the operation of Topology, recording each Topology of the individual components, the experimental results were compared in Figure 2.

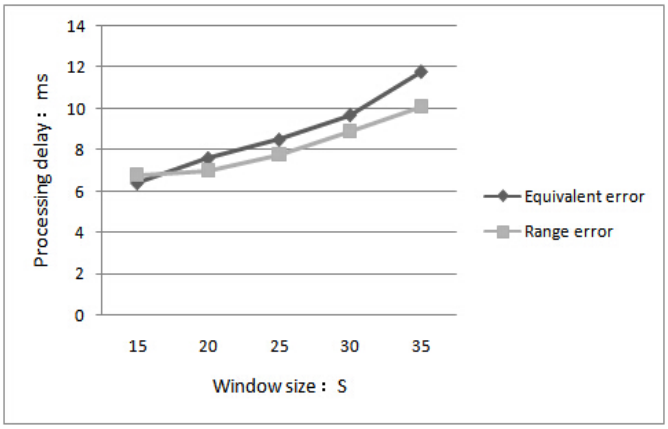

Figure 2.Different window sizes query processing delay

As can be seen from the figure the curve, the histogram-based summary data processing delay in milliseconds query can achieve near real-time requirements. Since the number of the histogram generated based on the error range of the histogram of the error slightly is less than the equivalent number of the histogram based on the data flow in the same conditions using shorter processing delay error range based on the histogram. With increasing of velocity data, the data faster query processing delay also increases with.

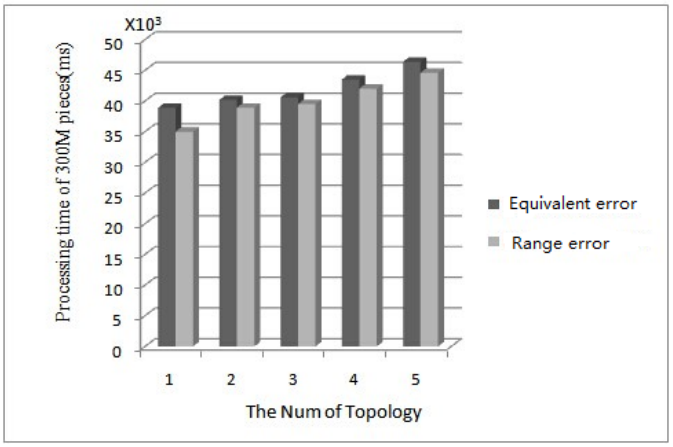

Figure 3. Processing delay time

As can be seen from the figure the curve, based on histogram processing delay summary data query in milliseconds, can achieve near real-time requirements. Since the number of histogram generated based on the scope of the error and histogram-based equivalence ratio error, head slightly less, at the same data flow conditions, the use of error based on the histogram processing delay will be shorter. With increasing velocity data, the data arrives faster query processing delay also with the increase.

Conclusion

With the generated stream data, stream data of various processing techniques are also produced. Storm distributed real-time computing platform in the process information in the form of data streams through the "Subscribe - send" pass, so with Storm stream data processing is very effective. This paper introduced the framework of continuous query Storm stream data, generate summary histogram data streams through the sliding window model, generate summary histogram-based equivalent range error and error based on two algorithms, using two different algorithms to generate histograms query, execution efficiency is also different. Text query only for a single data stream equivalent queries and range queries, the next step in this paper will connect multiple data streams, grouping operations research. With Storm framework, will shorten the processing of data streams in real time in response to the quasi-level research has important implications for various applications.

Reference

[1] Wu Lei based on real-time data flow algorithm and application of continuous query [D]. North China Electric Power University, 2006.

[2] Wang Weiping, Li Jianzhong, Zhang Dongdong. [J], Journal of Software data stream query processing methods JA query-based sliding window, 2006,17 (4): 740-749.

[3] Xuzheng Feng, Yang Geng .LBS continuous query model identifier [J]. Based Computer Technology and Development, 2011,21 (9): 237-241.

[4] Storm project page, Storm, distributed and fault-tolerant real-time computation [EB / OL] http://storm.incubator.apache.org/.

[5] Nathanmarz / Storm Wiki GitHub, Understanding the parallelism of a Storm topology [EB / OL] https:// github. com/ nathanmarz /storm/wiki /Under standing-the-parallelism-of-a-Storm-topology