Scholar Publishing Group
Distributed Processing System
https://doi.org/10.38007/DPS.2021.020305
ISSN 2790-0916 Vol. 2, Issue 3: 41-48

Structured Data Processing Method and Distributed Processing System Construction

Turchet Luca*

Guys & St Thomas NHS Fdn Trust, Royal Brompton & Harefield Hosp, Harefield Resp Res Grp, London, England

**corresponding author*

Keywords: Structured Data, Distributed Processing System, Text Preprocessing, Information Extraction

Abstract: Today, data is gradually changing in popular culture in a continuous, rapid and time-varying direction. According to the new application framework, people put forward higher requirements for the performance and functions of data systems, and also put forward higher requirements for the practice and understanding of data processing. This paper aims to study the processing method of structured data and the construction of distributed processing system. This paper first analyzes the research status of distributed data systems at home and abroad, then designs a typical distributed data processing system and discusses two basic technologies of the system, and finally applies it to the project of postal centralized delivery system. This paper summarizes the structural and linguistic features of the text. According to these characteristics, this paper proposes an organizational structure. The method has three main parts: text prioritization, new text discovery and information extraction. This paper analyzes the characteristics and memory space allocation of continuous query, and designs a dynamic window query based on the greedy principle, which improves the efficiency of continuous query processing. Proven by experiment. The new word rate with a word length of more than 5 is 0%. In the case of different data volumes, the load change rate of the system is less than 0.05, and the system is stable.

1. Introduction

In recent years, with the development of digital information technology and the development of network technology, the requirements for the transmission and processing of information are getting higher and higher. The information that can be represented by data or a unified structure is called structured. Data, such as numbers, symbols. Structured data has specific fields, namely row data, which are stored in the database, and the implemented data can be logically expressed in a two-dimensional table structure. For example, a user uses a social software to post a comment,

which has fields such as a post identifier (Identifier, ID), time, title, and text. With the rapid development of distributed data stream processing technology, a large number of experts, scholars and research institutions are devoted to the research of distributed data stream processing technology. Academia and industry fully realize that distributed data stream processing technology has broad application prospects and development space [1] -2].

In the research on the processing method of structured data and the construction of distributed processing system, many scholars have studied it and achieved good results. For example, Lin WD has deeply studied how to extract and detect abnormal patterns in the data flow environment, and proposed An algorithm for detecting and predicting the trend of abnormal patterns over time is developed, and a measurement framework and corresponding algorithms are constructed to accurately extract abnormal patterns [3]. Strzalka D deeply studied the method of processing data stream with continuous query [4].

This paper first analyzes the research status of distributed data systems at home and abroad, then designs a typical distributed data processing system and discusses two basic technologies of the system, and finally applies it to the project of postal centralized delivery system. This paper summarizes the structural and linguistic features of the text. According to these characteristics, this paper proposes an organizational structure. The method has three main parts: text prioritization, new text discovery and information extraction. This paper analyzes the characteristics and memory space allocation of continuous query, and designs a dynamic window query based on the greedy principle, which improves the efficiency of continuous query processing.

2. Research on the Processing Method of Structured Data and the Construction of Distributed Processing System

2.1. The Difference between Distributed Systems

If we regard the datasets as a special data stream, then we can define DDSMS as an extension of a traditional database system. Next, we first make an inductive comparison between DDSMS and DBMS.

Several functional and performance differences between traditional DBMS and DDSMS:

(1) The basic calculation model is not consistent. Traditional database management system assumes that DBMS passively stores data units, while users actively initiate queries and other operations. This is a user-active and DBMS passive model. The DDSMS obtains the data from an external data source and returns the data to the user when the system detects the data that meets the query conditions. This is a DDSMS-active and user-passive model.

(2) The DBMS query is an accurate query, and currently no DBMS provides a built-in function to support the approximate query. And DDSMS, due to the large amount of data and rapid changes, can often only provide approximate query results.

(3) DBMS provides a query, a query to obtain the query results, and DDSMS is a continuous query, as long as the user has registered a query, and does not cancel the query, then the query will always be valid, DDSMS constantly returns the query results to the user.

(4) DBMS usually does not consider the time and space constraints associated with transactions, and its scheduling and processing decisions do not consider the various time characteristics of the data. The design index of the system does not emphasize the adaptability of real-time and query service quality, while real-time and adaptability are just necessary for data flow applications [5-6].

Accessibility, transparency, openness, and scalability are the four main characteristics of distributed systems.

(1) Accessibility. Achieving convenient and fast resource sharing and allowing users to access remote resources more conveniently is the fundamental purpose of a distributed system. There are various types of shared resources, and various devices such as computers, scanners, servers, storage, and networks can be shared and used by other users as resources.

(2) Transparency. The transparency of a distributed computer means that the interface finally presented to the user is single, and the user cannot see the complex underlying structure and resource allocation of the system. There are four types of transparency: access, location, replication, and fault transparency. The transparency of the system improves the compatibility and portability of the system, hides the underlying complex architecture, and improves the user experience.

(3) Openness. The openness of a computer program is the property that determines whether the program can be extended and copied in different ways. The openness of a shared system depends primarily on the extent to which new resource-sharing services can be added and used by multiple client systems.

(4) Scalability. Distributed systems can operate effectively and efficiently at different scales. The system still works if the number of resources and users surges [7-8].

2.2. Algorithm Selection

The structured data processing method selected in this paper is based on the Dirichlet distribution algorithm. The Dirichlet distribution is a set of continuous multivariate probability distributions, just as the Beta distribution is the conjugate prior probability distribution of the binomial distribution. The Ray distribution is used as the conjugate prior probability distribution of the polynomial distribution [9].

$$Dir(\mu|\alpha)=\frac{\Gamma(a_0)}{\Gamma(a_1)\mathrm{K}\ \Gamma(a_k)}\prod_k K=1\mu_k^{a_{k-1}} \tag{1}$$

In layman's terms, the Dirichlet distribution is actually, under the premise that the probability set of the result set of R{R1, R2...Rn} obtained by event A in an experiment is p{p1, p2...pn}, Continue to do experiments to find out what is the probability P` of this probability set P, which is equivalent to finding the distribution above the probability distribution of event A. And such a distribution is the Dirichlet distribution. This paper still uses the latent Dirichlet model for word clustering to extract text topics [10].

$$p(\omega_i,z_i,\theta_i,\varphi|\alpha,\beta)=\prod_j N=1p(\theta_1|\alpha)p(z_{i,j}|\theta_1)p(\varphi|\beta)p(\omega_{i,j}|\theta_{i,j}) \tag{2}$$

2.3. Database Design

The role of the database is to store relevant data and make it available to users. The database design of this system is divided into three levels.

The third-level database refers to the database of each mail car, which stores the data information of all the parcels passing through the mail car.

The secondary database is between the monitoring system and the mail car, and is specially used to receive and classify the information of the mail package. The main function of collecting and storing the parcel data information of each postal car is to reduce the load of the monitoring system, improve the speed at which users can acquire data, and play a caching role.

The primary database is the database of the monitoring system. When the user queries data, the

monitoring system directly queries the secondary database for the required data and stores it for the user to find. In addition to storing postal data information, this database can also store some data that users often need to query, such as area code correspondence, postage costs, etc. [11-12].

3. Structured Data Processing Method and Distributed Processing System Construction Research Design Experiment

3.1. Overall System Design

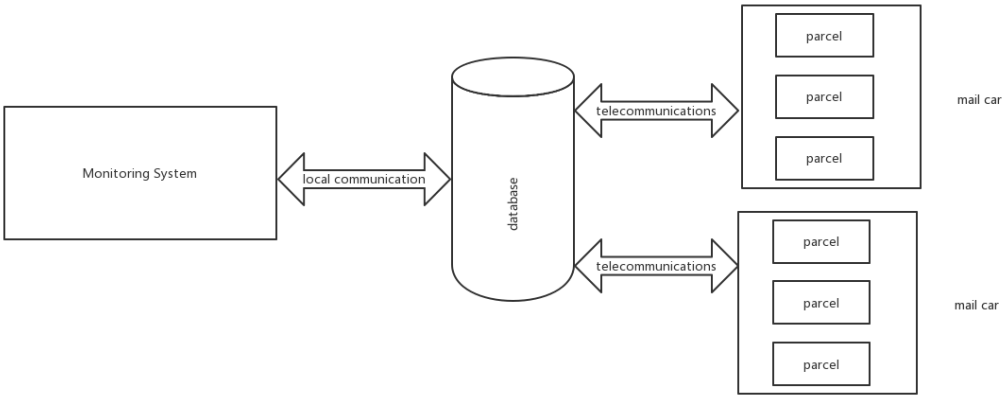

Figure 1. System flow diagram

(1) The parcel is an analog signal;
(2) The library stores the postal information on the basis of the communication protocol;
(3) The detection system is used for on-demand data information;
(4) Two-way communication is adopted between the library and the postal package, and the local two-way communication is used between the detection system and the library.

3.2. Experimental Design

This paper designs experiments for the structured data processing method and distributed processing system constructed in this paper. The first is to analyze the information extraction word length of structured data, and select a variety of word lengths to verify the relationship between new words and registered words.. The second is to study the load change rate of the distributed processing system to explore the stability of the system.

4. Structured Data Processing Method and Distributed Processing System Construction Research Experiment Analysis

4.1. Word Length

The first group of experiments to verify the necessity of setting the word length L threshold does not need to conduct experiments in a stand-alone environment and on the Spark platform. Because although the sizes of the two datasets are different, they describe the same thing and have the same structure, the login words contained in them are fixed, and the word length will not change due to

the different sizes of the datasets, so the first One set of experiments was performed only in a stand-alone environment. In order to verify the necessity of setting the word length L threshold, 1000 sample records are randomly sampled, and word length statistics are performed on these 1000 records. The length of new words and the length of registered words in the 1000 sample records are counted separately. If a word can form a new word with other words, the length of the word is mostly 1 or 2. If the length of the word exceeds 5, it cannot be part of a word, that is, it cannot form a new word. The experimental data are shown in Table 1.

Table 1. Proproportion of new words under different words

	1	2	3	4	5	>5
Login word	100	98	59	91	96	100
neologism	0	2	41	9	4	0

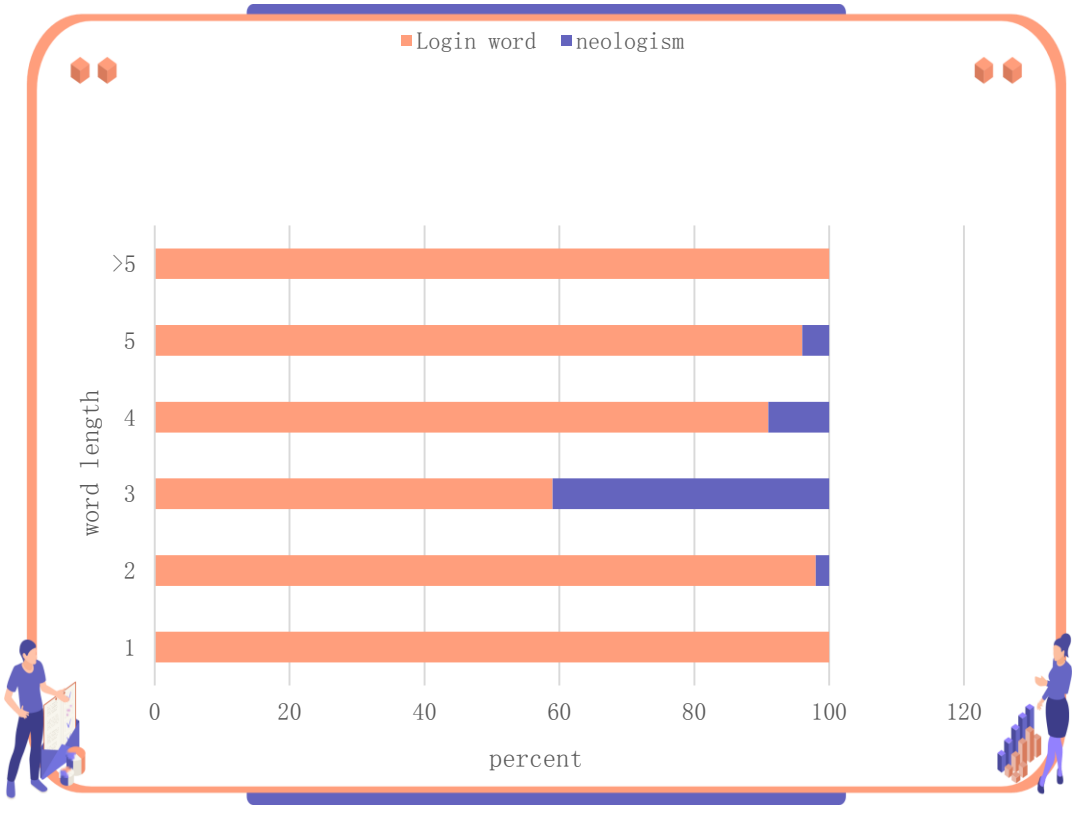

Figure 2. The ratio of new words and login words on L

As can be seen from Figure 2, the L values of most Chinese words are less than or equal to 5. Therefore, this paper sets the threshold value of L to 5. Words with L more than 5 default to the common combination of multiple registered words, not a new word. Preliminary screening based on

the size of L can improve the efficiency of new word discovery.

4.2. Analysis of Load Change Rate

The experiment mainly tests the load processing capability of DDSMS in the network.

Change the load of the system by adjusting and assigning nodes with idle DDSMS processing capacity: when the amount of data input per unit time is τ, the system is just not overloaded, that is, the amount of data that the system can process per unit time is τ, and the load of the system at this time is M; if the amount of input data per unit time is increased to 2τ, the system load at this time is 2M, and so on. The load changes were recorded through repeated experiments, and the experimental data are shown in Table 2.

Table 2. System load change diagram

	100	200	400	500	1000
Test1	1	0.97	0.95	0.94	1
Test2	1	0.98	0.95	0.95	1

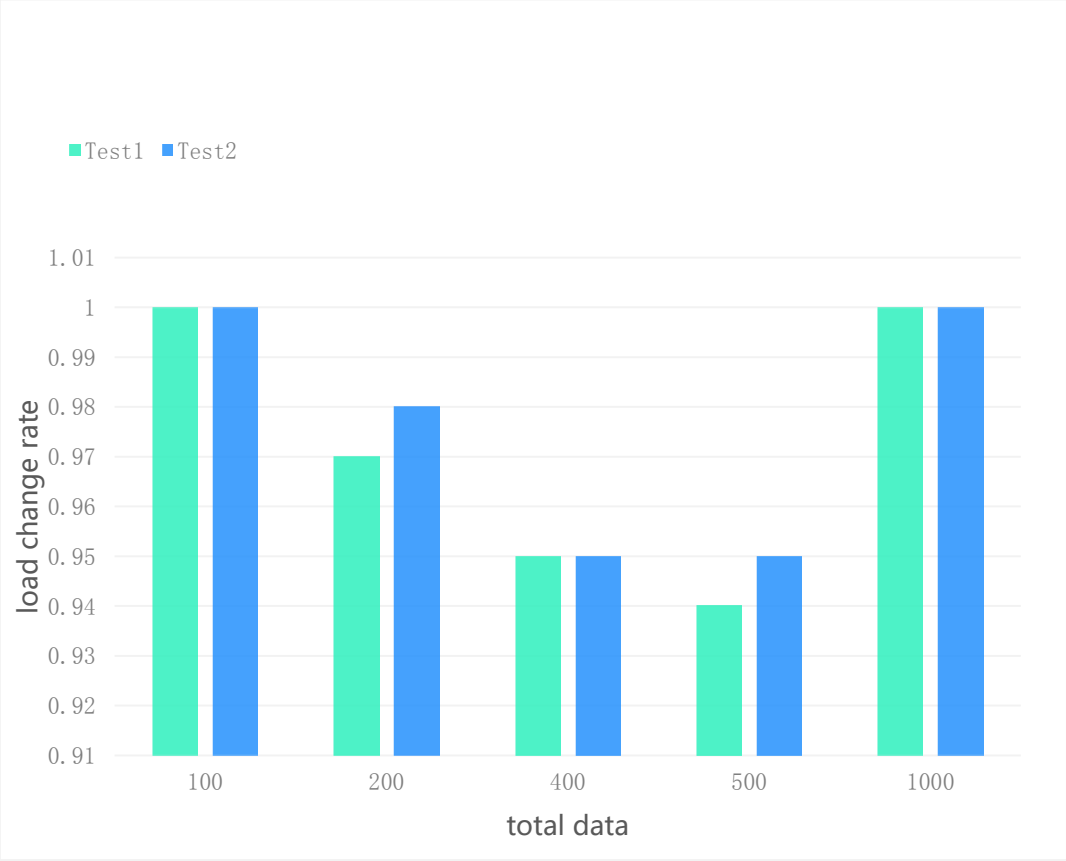

Figure 3. Data load under different total amount of data

It can be seen from Figure 3 that the load fluctuation rate increases slowly with the increase of the system load, and the difference between the peak value and the valley value is also about 0-05, indicating that the system has a good load handling capacity.

5. Conclusion

This paper studies several important aspects of distributed data processing systems, and analyzes the prototypes of related processing systems at home and abroad. A general DDSMS scheme is designed, and the basic functional modules are briefly analyzed. A dynamic query algorithm that controls the size of the sliding window is designed to limit the number of tuples stored in each input stream to minimize the consumption of memory space and processing time. The embodiments of this document provide a structured data processing method and a distributed processing system, which are used to improve the processing efficiency of write requests and reduce the queuing delay of the write operation queue. The distributed processing system is configured with a merge submission strategy for write requests, and the distributed processing system analyzes and judges multiple write requests stored in the write operation queue according to the merge submission strategy, so as to determine whether there is a At least two write requests of the write operation type. The merge commit strategy can include various implementations. For example, the write requests in the read and write operation queues can be polled regularly, so as to determine multiple write requests that are added to the write operation queue at the same time or in stages within a certain period of time. Whether batch processing is possible. The merge submission strategy can be determined by the operating user of the distributed processing system, configured in the distributed processing system through user configuration, or determined by the distributed processing system according to the queue storage situation of the write operation queue, for example, according to The number of write requests added to the write operation queue accounts for the ratio of the capacity of the write operation queue to determine whether to execute the merge-submission strategy in this embodiment.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Grembowski D, Ingraham B, Wood S, et al. Statewide Evaluation of Washington's State Innovation Model Initiative: A Mixed-Methods Approach.. Population health management, 2021, 24(6):727-737. https://doi.org/10.1089/pop.2020.0374

[2] Zhou X G, Gong R B, Shi F G, et al. PetroKG: Construction and Application of Knowledge Graph in Upstream Area of PetroChina. Journal of Computer Science and Technology, 2020, 35(2):368-378. https://doi.org/10.1007/s11390-020-9966-7

[3] Lin W D, Lei L Y, Ting R D, et al. Research on massive information query and intelligent analysis method in a complex large-scale system. Mathematical biosciences and engineering :

MBE, 2019, 16(4):2906-2926. https://doi.org/10.3934/mbe.2019143

[4] Strzalka D. CGraph: a distributed storage and processing system for concurrent iterative graph analysis jobs. Computing reviews, 2019, 60(12):463-463.

[5] Nanayakkara S, Perera S, Senaratne S, et al. Blockchain and Smart Contracts: A Solution for Payment Issues in Construction Supply Chains. Informatics, 2021, 8(2):article 36. https://doi.org/10.3390/informatics8020036

[6] Solonar A S, Khmarski P A. General construction principles and performance features of trajectory processing by data from one radar data source. Journal of Physics: Conference Series, 2021, 1864(1):012138 (9pp).

[7] Gu J, Li Z. Research on Data Secure Transmission and Processing Method of a LoRa IoT System. Journal of Physics: Conference Series, 2021, 1802(3):032044 (10pp).

[8] Li Y, Li Q, Shen W, et al. Research on the Layout and Data Processing Method of Distributed Optical Fiber in Shield Tunnel Monitoring. Journal of Physics: Conference Series, 2020, 1626(1):012012 (6pp).

[9] Tsingas C, Almubarak M S, Jeong W, et al. 3D distributed and dispersed source array acquisition and data processing. The Leading Edge, 2020, 39(6):392-400. https://doi.org/10.1190/tle39060392.1

[10] Mahdi H. Implementing QFD in decision making for selecting the optimal structural system for buildings. Construction Innovation, 2020, Vol. ahead-of-print(No. ahead-of-print):16.

[11] Han S, Choi J I, Woo G. Case Studies and Trends in Data Reproduction and Distributed Processing using Event Sourcing and CQRS Pattern. Journal of KIISE, 2020, 47(12):1101-1110. https://doi.org/10.5626/JOK.2020.47.12.1101

[12] Konikov A I, Fedoseeva T A. Analog-to-digital data processing tools in the construction industry and in the transportation sector. IOP Conference Series: Materials Science and Engineering, 2020, 918(1):012069 (7pp).

[13] Kostikov Y A. Development of an information system for distributed processing of streaming data. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(3):3853-3864.

[14] Christopher G B, Kabari L G. Hybridized Concurrency Control Technique For Transaction Processing In Distributed Database System. Artificial Life and Robotics, 2020, 9(9):118-127.

[15] Ha Y S, Park S Y, Lee D H. Construction on Lot Tracking System for Failure Cost Reduction of a Small and Medium Precision Parts Processing Company. Journal of Society of Korea Industrial and Systems Engineering, 2019, 42(3):80-88. https://doi.org/10.11627/jkise.2019.42.3.080

[16] Cheolgi, KIM, Daechul, et al. An Efficient Block Assignment Policy in Hadoop Distributed File System for Multimedia Data Processing. IEICE Transactions on Information and Systems, 2019, E102.D(8):1569-1571. https://doi.org/10.1587/transinf.2019EDL8016

[17] Benitez P, Rocha E, Talukdar S, et al. Efficiency analysis of optimal inspection management for reinforced concrete structures under carbonation-induced corrosion risk. Construction and Building Materials, 2019, 211(JUN.30):1000-1012.

[18] Thanzeel F, Balaraman K, Wolf C. Streamlined Asymmetric Reaction Development: A Case Study with Isatins.. Chemistry (Weinheim an der Bergstrasse, Germany), 2019, 25(47):11020-11025. https://doi.org/10.1002/chem.201902688