Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 1
AIEA 2023

Research on Guangdong Province Meteorological Data Analysis System under Hadoop Environment

Zhuxin Liu

Guangdong University of Science &Technology, Dongguan 523346, China

1424112136@qq.com

Abstract. In order to better monitor and manage the real-time transmission of meteorological information in Guangdong Province and improve communication quality, the Guangdong Meteorological Information Center has organized the development of the Guangdong Meteorological Data Analysis System. The system has the characteristics of intuitive and easy to use, strong real-time performance, and convenient query and statistics, which better meets the real-time monitoring and time management requirements of meteorological information transmission in Guangdong Province. Hadoop is a popular big data computing oriented, open-source big data processing system. Its characteristics of high throughput, high fault tolerance and easy expansion make it widely used in all walks of life. Therefore, this article focuses on the characteristics of real-time reception of meteorological information in Guangdong Province, and introduces the development ideas, use of some new technologies, and page implementation of the Guangdong Meteorological Data Analysis System under Hadoop environment. In the Hadoop environment, it is possible to dynamically exchange data with the server without refreshing the entire page, improving the efficiency of data transmission between the server and client, making browsing more intuitive and convenient. Enriched the methods and capabilities of data processing.

Keywords: Hadoop; Guangdong Province; Meteorological Data; Analysis System.

1. Introduction

With the development of meteorological cause, there are more and more kinds of real-time observation data, and the requirements of timeliness are getting higher and higher. The original monitoring method based on querying and receiving data directory files can no longer meet the needs. In order to better monitor and manage the real-time transmission of meteorological information in Guangdong Province and improve the communication quality, Guangdong Meteorological Information Center organized and developed the Guangdong Meteorological Data Analysis System [1]. The system is intuitive, easy to use, real-time, convenient for query and statistics, and meets the requirements of real-time monitoring and timeliness management of meteorological information transmission in Guangdong Province. Storing data in multiple independent devices, so that these devices can share the storage load, not only solves the bottleneck problem of single storage server in traditional centralized storage system, but also improves the reliability, availability and expansibility of the system [2-3]. The more data you choose, the more computation you need. Some studies are limited by too much calculation, such as spatial inspection in quality control of ground meteorological data. How to deal with massive meteorological data quickly and efficiently has become an important problem to be solved in meteorological science research. Hadoop is a popular big data processing system with big data operation and open source. Its high throughput, high fault tolerance and easy expansion make it widely used in all walks of life. The core of Hadoop lies in HDFS distributed file system and MapReduce distributed computing model. Its advantage lies in that it can use relatively low-end hardware configuration to achieve high-throughput file access, especially suitable for meteorological departments, which have huge meteorological data sets [4]. Therefore, according to the characteristics of real-time reception of meteorological information in Guangdong province, this paper introduces the development idea of meteorological data analysis system in Guangdong province under Hadoop environment, the use of some new technologies and the realization of the page [5]. More importantly, the system combines the functions of mathematical statistics, text analysis and

graph mining with traditional data mining tools, which enriches the methods and capabilities of data processing.

2. How Hadoop Works

Hadoop, developed by the Apache Software Foundation, is a distributed infrastructure architecture. Hadoop is currently a popular distributed file system, which has been widely used in various industries due to its high throughput, high fault tolerance, and easy scalability. Its advantage is that it can utilize relatively low-end hardware configurations to achieve high throughput file access, especially suitable for business applications with massive meteorological datasets in meteorological departments [6]. In 2005, the distributed file system used by Hadoop was introduced to the market as a support for low-level storage. One of the characteristics of HDFS is that it has high fault tolerance, and is designed and deployed on relatively inexpensive hardware devices to improve access rate, which is more suitable for programs with large data. Use meteorological sensors to collect meteorological data from around the world, which are distributed in various parts of the world. As time goes by, the accumulation of data is increasing, so using Mapreduce for data analysis is optimal [7]. Its own streaming storage mode is designed for this "write once, read multiple times" file storage, which can be customized and has extremely high efficiency. Each directory contains all the meteorological data packages of a meteorological station for the year, and in real life, there are many meteorological stations, so these data are all composed of data collected by many meteorological stations. In general, we can easily handle some large files.

3. Research on Meteorological Data Analysis System

At present, all kinds of meteorological data in relational database are mainly based on observation items and data classification, and rarely involve complex multi-table associations such as foreign keys. This simple relationship between database and table provides great convenience for the migration of meteorological data from relational database to database. Before the establishment of the meteorological data analysis system, after the meteorological data files are sent by the observation stations in all cities and counties in the province, it is mainly through accessing the receiving directory of the data communication file server of the information center or calling the personnel on duty in the business system of the information center to check whether the data files are successfully sent [8]. If the data files are missing for some reasons, the business personnel on duty in the information center need to call the station to follow up the report, which is inefficient and difficult to ensure the transmission timeliness of real-time data. In the past, the use of meteorological data was mainly for processing and accessing single or multiple data files. At present, a considerable part of historical data in this province is compressed and stored on disk every month, and it is inefficient and inconvenient to view and operate the data files after copying and decompressing [9]. According to incomplete statistics, the information center has stored more than 40TB of historical data and business information. With the rapid growth of data volume, the efficiency of this data archiving and use method will become lower and lower. Therefore, the reasonable design of meteorological data analysis system in Hadoop environment is shown in Figure 1, which can monitor real-time meteorological data in real time and publish information in the form of website, which is convenient for the reporters in the observation stations of various municipal bureaus to check in time and is conducive to improving the transmission timeliness of real-time meteorological data in Guangdong Province.

The meteorological data analysis system has strong applicability and the potential for business and large-scale applications. Taking the national level automatic station ground meteorological element data in structured data and the new generation Doppler weather radar product files in unstructured data as typical examples, a distributed storage scheme for the two in the framework is proposed and implemented. Starting from the server, formulate a load balancing policy, such as when the server has a high load at a certain moment, and a new user visits the server, the load balancing device will

allocate this connection to the device with a smaller load [10]. Among them, the real-time monitoring page can display the timely arrival and sending rates of various data on time or on the same day. The information query page can query the sending and receiving status of certain data at a certain time. The information statistics page conducts detailed statistics on the collected real-time meteorological data by day, month, and year. Every time a client page makes a request, whether it is simple or complex, the server responds and sends the updated entire page to the client, resulting in low operational efficiency. In the Hadoop environment, it is possible to dynamically exchange data with the server without refreshing the entire page, improving the efficiency of data transmission between the server and client, making browsing more intuitive and convenient.

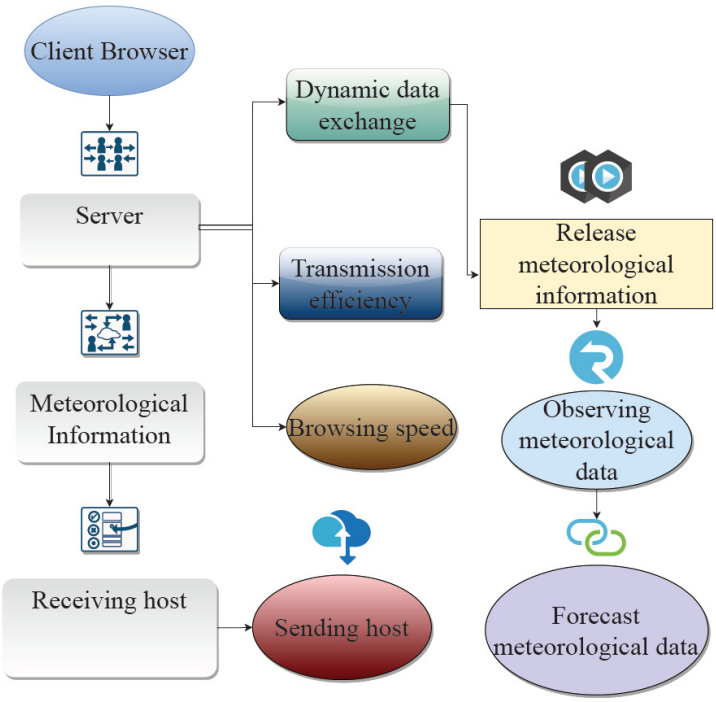

Figure 1. Flow Chart of Meteorological Data Analysis System

4. Meteorological Data Statistical Efficiency Testing

4.1. Main Configuration Parameters

Meteorological observation data has authenticity and objectivity, and is a typical data type of "write once, read multiple times". The vast majority of meteorological data comes with geographic and temporal information attributes in file names or file contents. Among them, geographic information attributes are specifically represented by station numbers, time attributes are represented by observation time, and the granularity of observation time is usually accurate to seconds. In terms of high-altitude meteorology, recorded data refers to the average symbol, time, wind direction, wind speed, wind layer, etc. of humidity, pressure and other elements in the sounding system at and between the specified layers, forming a record table. The recorded data mainly specifies or converts the original coefficients, and the purpose of this data type is to provide a scientific and comprehensive database for the data. The design in the system implementation process needs to meet the principle of hashing, which is used to improve the probability of multi machine load balancing, reduce the concentration of hotspots at the same time, and cleverly arrange the station numbers of meteorological data in reverse order. In order to further verify the effectiveness of the meteorological data analysis system, the distributed computing method is mainly used to calculate the amount of Big data, so the efficiency of large clusters is unquestionable. However, due to the limited test conditions, the efficiency of dual computer clusters can only be tested. Table 1 shows the comparison of some software and hardware configurations between the two hosts.

Table 1. Main Configuration Parameters

Name	CPU main frequency	Cache	Internal storage	Operating system
Compute-A	800*2	214	523	Ubuntu
Compute-B	1865.145	136	259	Ubuntu

Push the queue to the merge task; After sorting the pushed queue information once, the merge task merges the queue into a large file and writes it, and finally writes the file information into the first-level index. In practical work, we can make full use of the middle function to query batch data within the time range, which greatly improves the query efficiency of long sequence set data.

4.2. Test Result

The statistical time-consuming based on meteorological data analysis system is obviously superior to the traditional relational database in performance and efficiency, and the retrieval and downloading time of unstructured data index database based on meteorological data analysis system is obviously superior to the traditional index database. Table 2 is a comparison of test efficiency from l MB data volume to l GB data volume.

Table 2. Running time of different files in different test modes

Filesize	Test mode			
	Single machine	Pseudo distribution	Distributed A-B	Distributed B-A
4	0.7857	3.4215	5.7542	2.8562
18	1.5647	5.1236	14.2365	5.4521
65	4.3569	14.5687	56.5415	16.3524

From Table 2, it can be found that with the increase of the span of meteorological data time series, the time-consuming of traditional relational database in dealing with high-frequency reading and writing will increase continuously and even lead to inability to handle it; The distributed storage technology based on, through multi-node parallel processing, can easily handle huge data; Play the role of data warehouse in, which provides the interface of classes for the data stored on the internet to facilitate the user's processing. Although the time consumption increases irregularly and linearly, it will eventually complete the requested task. Meteorological data analysis system can select data category, subcategory, municipal bureau, county, station number, date, time and minute parameters to query the required information. The display area will display the data category and subcategory of the query, including the number of arrival, the number of actual arrival, the timely rate of reporting and the timely rate of sending. This data type has some limitations and big errors at first, and it is gradually becoming perfect after continuous induction and summary. Although empirical data is not formal data, it has strong usability and has been widely used in meteorological forecasting business. This data type can also be converted into data and specified data under certain conditions. If the arrival or transmission time exceeds the limit, blue will be displayed between the arrival and transmission time; If there is a missing report, it will be displayed in red at the time of arrival or transmission, which is very intuitive and convenient.

5. Conclusion

This article conducts further research on the Guangdong Provincial Meteorological Data Analysis System under the Hadoop environment. Since its trial operation on the provincial business network, the Guangdong Provincial Meteorological Data Analysis System has been widely used by meteorological staff in all provinces, cities, and counties. Business duty staff can directly browse the sending status of information files in real-time through the website, providing data organization

functions, solving the problem of data sources and increasing the processing capacity of similar databases; The business engine can freely combine various analysis algorithms to meet different levels of requirements, making it easy to develop typical applications. Therefore, this system has superior performance, rich functionality, and broad commercial prospects. It is both fast and convenient, which has played a good role in improving the transmission efficiency of real-time meteorological data in our province. The distributed architecture in the Hadoop environment can realize the high throughput, high concurrency, high fault tolerance, high reliability and low cost of the system. As the basic framework of cloud computing, it is of great significance to solve the problems faced by meteorological data such as massive expansion, high concurrency read and write, and data retrieval efficiency of long series Big data sets. Therefore, it is necessary to provide more authentic, scientific, and objective data for meteorological forecasting, scientific research, etc., so as to effectively improve the quality of high-altitude meteorological observation and better serve the meteorological industry, thereby promoting the sound and rapid development of the meteorological industry.

References

[1] Xiaofang W, Zhengquan C, Liping J,et al Analysis of a Rare Cryogenic Freezing Rain and Snow Event in Cold Wave Weather over Guangdong Province in 2016[J].Meteorological Science and Technology, 2022, 32(16):31-39.

[2] Zhang Y, Chen X, Bi S, et al. Design and Application of Integrated Monitoring Model for Provincial Meteorological Observation Data Transmission[J]. Journal of Earth Science and Environmental Protection, 2019, 7(5):11-20.

[3] Lin-Gen Y, Jing-Yong Z, Computer D O. Design of LTE Performance Analysis and Prediction System Based on Hadoop[J]. Modern Computer, 2019, 12(4):16-20.

[4] Huanli L, Zenglu F, Chan L. Comprehensive Assessment and Analysis of National Surface Meteorological Observation Stations: A Case Study of Hebei Province[J]. Meteorological and Environmental Research: English Version, 2019, 10(4):7-16.

[5] Jinqing Y, Yaobo H, Xiqing L U, et al. Application of Agrometeorological Proverbs in the Climate Prediction of Precipitation in Flood Seasons[J]. Chinese Journal of Tropical Agriculture, 2021, 25(4):19-26.

[6] Hua W, Yaodong D, Yu Z, et al. Insurance Risk Zoning of Late Rice Cold Dew Wind in Guangdong[J]. Meteorological and Environmental Sciences, 2018.

[7] Pang Guqian, He Jian, Liu Chang, et al. Hazard Analysis of Severe Convective Weather in Guangdong Province, China[J]. Journal of Tropical Meteorology: English Edition, 2021, 27(2):8-18.

[8] Cui W, Chen-Xian Z, Si-Yuan H. Analysis of A Warm-sector Heavy Rain in Qingyuan, Guangdong[J]. Guangdong Meteorology, 2022, 18(5):16-24.

[9] Xiaowei R, Zhi H, Liqun Z. The application of Hadoop in the storage of meteorological data[J].Journal of Meteorological Research and Application, 2022, 36(20):25-36.

[10] Zhi H, Liqun Z, Xiaowei R, et al. Query and Statistical Analysis of Mass Automatic Station Data Based on SparkSQL in Hadoop Environment[J]. Meteorological Science and Technology, 2019, 32(5):11-16.