Intelligent big data service for meteorological cloud platform

Jie Zhang

Shanghai Meteorological Information and Technical Support Center,
Shanghai Meteorological Bureau,
No. 166, Puxi Road,
Shanghai 200030, China
Email: jzhangchina@126.com

Shengjun Xue* and Tao Huang

School of Computer Science and Technology,
Silicon Lake College,
Suzhou, China
Email: sjxue@163.com
Email: nuisthuangtao@163.com
*Corresponding author

Abstract: With the acceleration of meteorological informationisation, meteorological data has gradually become a typical industry big data. In view of the challenge of storage, index and processing, cloud computing technology provides the technical support for meteorological big data. We design a framework of meteorological big data service in the cloud computing environment which contains meteorological services, meteorological scientific research services and public meteorological services. As the most popular distributed processing technology, MapReduce is effectively used for distributed processing of meteorological big data. Finally, based on MapReduce, the daily meteorological data of Baoshan Station in Shanghai is analysed, and the statistical results and corresponding examples are provided. The application research of meteorological big data in the cloud environment can not only improve the overall meteorological service level, but also play a significant role in accelerating the meteorological informatisation process in the big data era.

Keywords: cloud computing; meteorological mata; meteorological mervices; Hadoop; MapReduce.

Reference to this paper should be made as follows: Zhang, J., Xue, S. and Huang, T. (2019) 'Intelligent big data service for meteorological cloud platform', *Int. J. Intelligent Internet of Things Computing*, Vol. 1, No. 1, pp.23–31.

Biographical notes: Jie Zhang received her Bachelor's and Master's in Computer and Software from the Nanjing University of Information Science and Technology in 2010 and 2013 respectively. She has worked in the Shanghai Meteorological Bureau since 2013. Her research interests include issues related to meteorology, big data and cloud computing.

Shengjun Xue graduated from the Zhejiang University in 1983 with a major in computer application. He was promoted to Professor and was hired as a doctoral supervisor, in 2000 and 2012 respectively. He is currently an academic leader in high performance computing at the School of Computer and Technology at Silicon Lake College. His research interests include issues related to big data and cloud computing.

Tao Huang received his Bachelor's and Master's from the School of Computer and Software, Nanjing University of Information Science and Technology, in 2010 and 2013 respectively. He worked in the Shanghai Jiading District Meteorological Bureau from 2013 to 2018, and started his career as a Lecturer at the School of Computer and Technology at Silicon Lake College in August 2018. His research interests include issues related to artificial intelligence algorithms, big data and mobile cloud computing.

1 Introduction

With the continuous acceleration of meteorological informatisation, the historical meteorological data are accumulated (Abdullahi et al., 2016). At present, the total amount of meteorological data saved by China National Meteorological Administration reaches PB level, and the new data volume per year is also close to PB level (Abdelbaky et al., 2012). Meteorological data are usually collected through observation stations, and has various types, including surface observations, satellite data, radar data, numerical prediction products, weather products and so on (Chen et al., 2014). At present, China has more than 50,000 automatic weather stations, more than 2,000 surface observation stations and more than 1,000 other kinds of observation stations (Wang et al., 2017). The historical data and real-time meteorological data collected by these automatic weather stations constitute the meteorological big data (Abdelbaky et al., 2012; Chen et al., 2014).

Meteorological big data has rich application and research value, and can provide a variety of meteorological services (Alarabi et al., 2018), including business services, research services and public services. Meteorological service mainly provides various services including meteorological data query, forecast product production, data loading (Chen et al., 2013). Research services mainly refer to the numerical analysis of meteorological big data for risk assessment of meteorological disasters (Cordeiro et al., 2011; Xu et al., 2018). Public services mainly refer to the processing and analysis of meteorological observation data, which is benefit to people's daily production, life and other industries. However, we are faced with huge technical challenges in the management, storage, analysis, processing and retrieval of meteorological big data (Hu et al., 2018).

Cloud computing can provide technical support for meteorological big data services (Ismail et al., 2017). Due to elastic computing, virtualisation, on-demand services, remote disaster preparedness, distributed storage and processing technologies, more and more users choose to deploy their applications to the cloud platform (Lakshmanan and Humphrey, 2014; Li et al., 2012). The *National Institute of Standards and Technology* (NIST) defines cloud computing as: cloud computing technology can easily acquire resources on-demand service and pay-as-you-go via the internet, such as computing, storage and network bandwidth, etc. which can be quickly acquired and released through the internet. With the emergence of Hadoop which is an open source cloud platform, the *private cloud* or the *public cloud* in the meteorological industry can be come true (Li et al., 2014; Li and Shen, 2017; Li et al., 2018). As a distributed processing model in Hadoop, MapReduce model can effectively support the analysis and processing of massive meteorological data (Ma et al., 2017), and as a distributed database in Hadoop, HBase can provide technical support for the retrieval of massive meteorological data (Rui et al., 2014; Shuai, 2017). The research of meteorological big data in the cloud environment can not only solve the problems of storage and processing of massive data perfectly, but also improve the service level of the meteorological industry.

2 Meteorological big data service framework in cloud computing environment

2.1 Meteorological big data feature

Big data is usually characterised by large data volumes, various data types, fast data processing speeds, and high data values, and these four characteristics are often considered to be the fundamental features of big data. In order to explore whether big data technology is suitable for meteorological data, it is necessary to determine whether meteorological data has the characteristics of big data.

Firstly, the current meteorological industry has accumulated a large amount of data, and the amount of data has reached the PB level, so this feature is consistent with the volume characteristics of big data. Secondly, the meteorological data are diverse, including various structured and unstructured data, which match the variety characteristics of big data. Thirdly, the meteorological data is usually generated continuously every minute, the rapid accumulation of meteorological data requires the fast processing speed. the requirements for processing speed. Finally, the processing, analysis and deep mining of meteorological data are beneficial to the services of the meteorological industry, such as weather forecasting, disaster warning, etc. so meteorological data has a high value.

Through the above analysis, meteorological data fully meet the characteristics of big data, so how to effectively store and process the meteorological big data has become an urgent problem to be solved. As an important technology for processing big data, cloud computing can be applied to the meteorological industry and help to develop and implement efficient meteorological big data services.

2.2 Meteorological big data service dramework

In the current meteorological industry, the meteorological regional centre is distributed with a large number of hardware devices, including the high-performance computers, the traditional storage servers and the network communication devices. These infrastructure devices can be integrated through cloud computing technologies to form a proprietary cloud, as shown in Figure 1. Therefore, the meteorological big data service framework can be deployed based on the proprietary cloud in the meteorological industry.

Based on the above description, we design a meteorological big data service framework in the cloud environment, as shown in Figure 2. The framework mainly consists of five levels, which are described as follows:

- *Infrastructure layer:* In the infrastructure layer, there are many servers, storage resource, network equipment, lighting system, refrigeration system and data centre site which are the physical facilities provided to store and process of meteorological big data. So that it can dynamically provide meteorological operators and researchers with computing and storage resources at the infrastructure level.
- *Platform layer:* In the platform layer, the *distributed file system* is used to realise the redundant storage of distributed files. The dynamic distributed retrieval of meteorological big data can be achieved by using the distributed database HBase. As the distributed computing model, MapReduce can be used for parallel computing of meteorological big data. The static meteorological data storage and the convenient retrieval can be realised by using the data warehouse Hive.
- *Application layer:* In the application layer, based on the software tools provided in the *platform layer*, a variety of meteorological applications are developed, including the *site monitoring*, the *cloud platform monitoring and management*, and the *weather services*. Among them, the *site monitoring* is mainly to monitor and manage various weather stations and their equipments. The *cloud platform monitoring and management* mainly manages and monitors the server nodes of regional meteorological centre dynamically. According to the different requirements, the *weather services* mainly include operational services, public meteorological services and scientific research services.
- *Big data services layer:* In the big data service layer, it mainly provides various types of meteorological big data services. For example, it can provide meteorological services by using real-time data retrieval based on HBase. And we can produce the numerical forecasting products based on MapReduce to provide the meteorological scientific research services and the public meteorological services.
- *User layer:* The users of meteorological big data services mainly include meteorologists, meteorological researchers and related industry personnel. Meteorologists can use obtain meteorological big data service through the intranet, the meteorological researchers can obtain some license data through the internet, and the related industry personnel can also obtain predicted products by browsing the weather website.

Figure 1 Deployment diagram of meteorological proprietary cloud (see online version for colours)

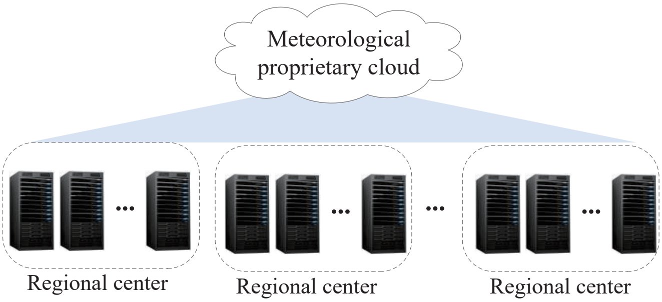
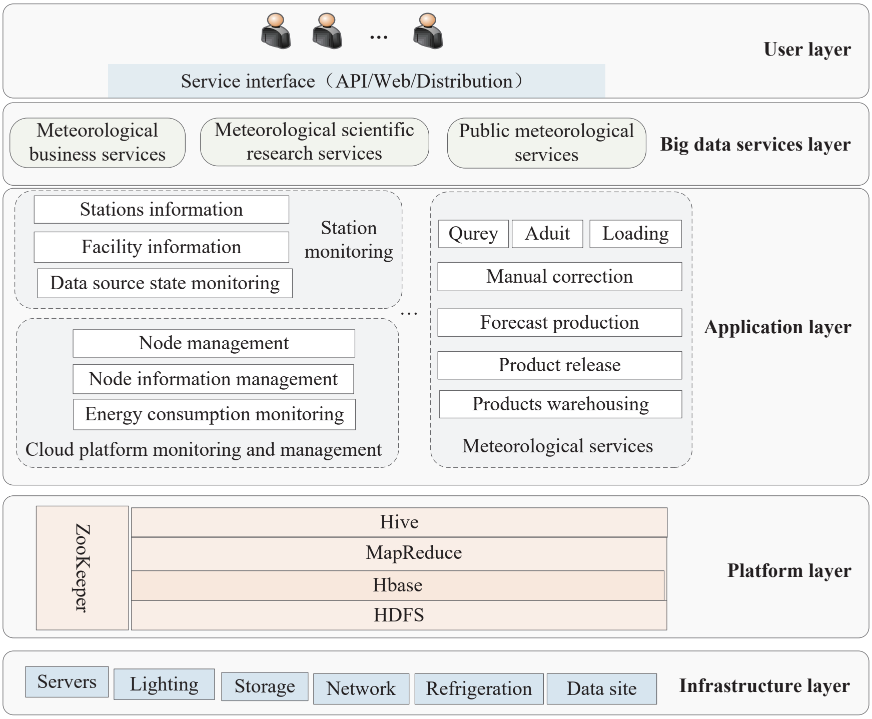


Figure 2 Meteorological big data service framework in cloud environment (see online version for colours)



3 Example of meteorological big data service based on MapReduce

3.1 Meteorological big data processing based on MapReduce

Meteorological big data can be used for parallel processing based on MapReduce, and the processing flow is shown in Figure 3. Before processing the input data, it divides the data into data slices of the same size. The calculation of each data is mapped to a map task. These map tasks will be sent to the nodes where the data is located for calculation. After the calculation is completed, each compute node returns its intermediate calculation result. Finally, multiple reduce tasks are created by the cluster, and the intermediate calculation results returned by all map tasks are combined into the final result.

Map and reduce operation follow the following formats:

$$Combine() : (k2, list(v2)) \rightarrow list(k2, v2) \tag{1}$$

$$Map() : (k1, v1) \rightarrow (k2, list(v2)) \tag{2}$$

$$Reduce() : (k2, list(v2)) \rightarrow list(k3, v3) \tag{3}$$

All of the above operated files are stored to HDFS, and as you see from the above operations, the input of $Map()$ and $Reduce()$ are both the form of $(key, value)$.

Figure 3 Meteorological big data processing based on MapReduce (see online version for colours)

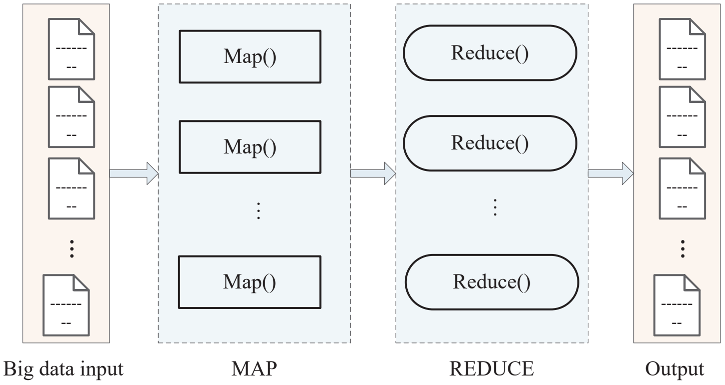
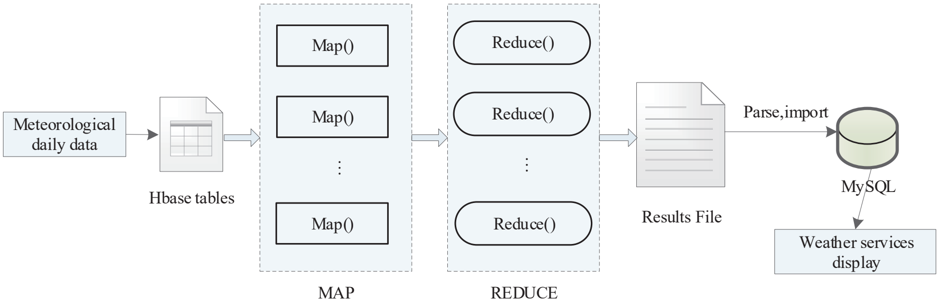


Figure 4 Logical flow of data processing based on MapReduce (see online version for colours)



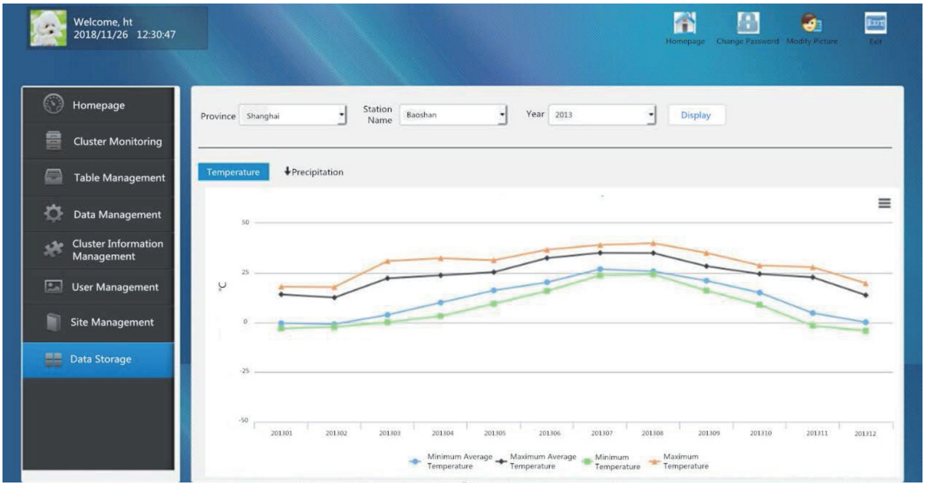
3.2 Examples of temperature and precipitation statistics based on MapReduce

The processing data of this section are the daily surface meteorological data of China International Exchange stations, which is saved in HBase database before processing. In order to provide meteorological display services, MapReduce is used to process and analyse the data. Through the analysis of experimental data, it can be concluded that the statistical analysis of temperature and precipitation has practical significance. When the data of the stations, such as maximum temperature, minimum temperature, maximum average temperature, minimum average temperature and precipitation are counted monthly, we can clearly find the changes of the historical temperature and precipitation can be a reference for the changes of temperature and precipitation in the future.

The logical flow of data processing method is shown in Figure 4. MapReduce distributed process the data stored in HBase. Multiple map tasks read data records in parallel, the input of the data are pairs of $(key, value)$, in which the key is the row key and the value is the row data corresponding to the row key. After the processing of $Map()$, it generates new $(key, value)$ pairs as the input of reduce task. Then after the processing of $Reduce()$, the monthly data is written into the result file and imported into MySQL using the parsing information. Finally, the yearly data is calculated by using the SQL statement and related functions.

Figures 5 and 6 respectively show the statistical analysis of temperature and precipitation in Shanghai Baoshan station from January to December 2013 based on MapReduce. So we can find that MapReduce can effectively support meteorological big data service.

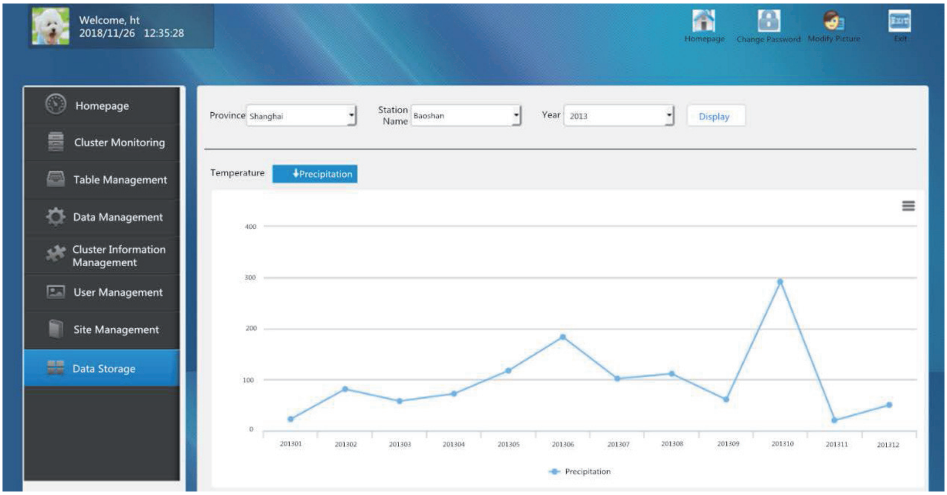
Figure 5 Temperature statistics of Shanghai Baoshan based on MapReduce (January 2013–December 2013) (see online version for colours)



Welcome, ht
2018/11/26 12:30:47

Province Shanghai
Station Name Baoshan
Year 2013

Figure 6 Precipitation statistics of Shanghai baoshan based on MapReduce January 2013–December 2013) (see online version for colours)



Welcome, ht
2018/11/26 12:35:28

Province Shanghai
Station Name Baoshan
Year 2013

4 Conclusions and future work

The meteorological big data service framework designed in this paper lays a foundation for meteorological big data services and provides a technical reference for the meteorological industry. The service and application of meteorological big data in the cloud environment can accelerate the informationisation process of the meteorological industry, promote the information sharing of the meteorological department, improve the remote disaster backup plan, and improve people's daily life.

Based on the work done in this paper, we still find some shortcomings of the architecture, especially in the storage and processing of meteorological data, the efficiency is low, and the calculation speed is slow when dealing with complex computing tasks. So we will adjust And optimise the storage and processing architecture of meteorological data to enable faster storage and computational efficiency. Finally, we constantly update our methods based on actual performance.

References

Abdelbaky, M., Kim, H., Rodero, I. and Parashar, M. (2012) 'Accelerating MapReduce analytics using cometcloud', in *IEEE Fifth International Conference on Cloud Computing*.

Abdullahi, A.U., Ahmad, R. and Zakaria, N.M. (2016) 'Big data: performance profiling of meteorological and oceanographic data on hive', in *International Conference on Computer & Information Sciences*.

Alarabi, L., Mokbel, M.F. and Musleh, M. (2018) 'St-Hadoop: a MapReduce framework for spatio-temporal data', *GeoInformatica*, Vol. 22, No. 4, pp.785–813.

Chen, D., Wang, L., Holger, R. and Chen, J. (2013) 'G-Hadoop: MapReduce across distributed data centers for data-intensive; computing', *Future Generation Computer Systems*, Vol. 29, No. 3, pp.739–750.

Chen, D., Zeng, L., Liang, Z. and Xiao, W. (2014) 'HBase-based distributed storage system for meteorological gound minute data', *Journal of Computer Applications*, Vol. 34, No. 9, pp.2617–2621.

Cordeiro, R.L.F., Traina, C., Traina, A.J.M., López, J., Kang, U. and Faloutsos, C. (2011) 'Clustering very large multi-dimensional datasets with MapReduce', in *ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*.

Hu, C., Li, W., Cheng, X., Yu, J., Wang, S. and Bie, R. (2018) 'A secure and verifiable access control scheme for big data storage in clouds', *IEEE Transactions on Big data*, Vol. 4, No. 3, pp.341–355.

Ismail, K.A., Majid, M.A., Zain, J.M. and Abu Bakar, N.A. (2017) 'Big data prediction framework for weather temperature based on MapReduce algorithm', in *Open Systems*.

Lakshmanan, V. and Humphrey, T.W. (2014) 'A MapReduce technique to mosaic continental-scale weather radar data in real-time', *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, Vol. 7, No. 2, pp.721–732.

Li, Z. and Shen, H. (2017) 'Measuring scale-up and scale-out hadoop with remote and local file systems and selecting the best platform', *IEEE Transactions on Parallel & Distributed Systems*, No. 99, p.1.

Li, B., Mazur, E., Diao, Y., Mcgregor, A. and Shenoy, P. (2012) 'Scalla: a platform for scalable one-pass analytics using MapReduce', *ACM Transactions on Database Systems*, Vol. 37, No. 4, pp.1–43.

Li, J., Chen, X., Li, M., Li, J., Lee, P.P.C. and Lou, W. (2014) 'Secure deduplication with efficient and reliable convergent key management', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 25, No. 6, pp.1615–1625.

Li, J., Wang, J., Lyu, B., Wu, J. and Yang, X. (2018) 'An improved algorithm for optimizing MapReduce based on locality and overlapping', *Tsinghua Science and Technology*, Vol. 23, No. 6, pp.112–121.

Ma, X., Fan, X., Liu, J., Jiang, H. and Kai, P. (2017) 'Vlocality: revisiting data locality for MapReduce in virtualized clouds', *IEEE Network*, Vol. 31, No. 1, pp.28–35.

Rui, Z., Hildebrand, D. and Tewari, R. (2014) 'In unity there is strength: showcasing a unified big data platform with MapReduce over both object and file storage', in *IEEE International Conference on Big Data*.

Shuai, Z. (2017) 'Application-aware network design for Hadoop MapReduce optimization using software-defined networking', *IEEE Transactions on Network & Service Management*, No. 99, p.1.

Wang, X., Gai, Z. and Qi, S. (2017) 'An approach for extracting big micro-scale severe weather region trajectories automatically from meteorological radar data', in *IEEE International Conference on Big Data*.

Xu, X., Liu, Q., Luo, Y., Peng, K., Zhang, X., Meng, S. and Qi, L. (2018) 'A computation offloading method over big data for IoT-enabled cloud-edge computing', *Future Generation Computer Systems*, Vol. 95, No. 2, pp.522–533.