The application of big data in production and environment

Ya Chuan YAO, Qiang HAN, Hong Ying YANG, Xu Yao ZHANG
School of Physics & Electronic Engineering, Sichuan University of Science & Engineering, Zigong, SiChuan, 643000, China
610851229@qq.com

Abstract. Big data technology is applied to the industry, production process parameters, input cost, energy saving and water saving, environmental impact and other large amounts of data involved in the production of industry are analyzed, it established a comprehensive data analysis platform. This paper designs the micro cloud hierarchical distributed processing framework based on the research, and studies a data-intensive storage scheme based on cloud computing. Research on big data storage scheme of industrial production based on cloud computing. Industrial production data continues to grow rapidly, the pressure of data storage is more and more large, to make full use of storage resources, to meet the actual needs of industrial production data storage and efficient processing of data.

1. Introduction
Big data has the characteristics of 5V, that is, Volume (large amount), Variety (diversity), Value (value), Velocity (high-speed), Veracity (authenticity). Industrial environment data is mostly unstructured data, this data is obviously greater than the difficulty of data structure, need to consider the implementation of non structure data while retaining the feature data into structured data. The system is mainly constructed from the key technologies of big data collection, transmission, storage, real-time data processing, heterogeneous data fusion and big data visualization. Data acquisition part of the system will be all kinds of sensors (differential pressure transmitter, flowmeter, liquid level sensor, temperature transmitter, etc.) RFID sensing device and a plurality of wireless receiver deployment and salt chemical industry equipment, consisting of a local wireless LAN as the perception layer of the system through the RDF algorithm deployment of heterogeneous nodes. The wireless receiver receives data through the automatic sensing device and networking gateway connected to the Internet network layer, the data is transmitted to the data center, in the application layer through data mining and other big data analysis and storage of data, while interacting with support 3G/4G or 5G smart mobile devices and data center and application platform.

2. Big data Overall Structure
The hierarchical module of the big data processing process is divided into four levels, namely, big data acquisition, storage, analysis, display and application modules, which are briefly explained below.
2.1. Data Acquisition

The main system sensing layer is the data collection, the statistical analysis of the collected data analysis feedback, the main types of sensor and RFID camera, sensing equipment deployed in the geological survey and mining, salt system, two alkali basic chemical and fine chemical industry equipment system, in order to obtain massive data. Due to the existence of sensing devices and wide range of distribution. Due to the existence of many sensing devices and a wide range of distribution, it is proposed to use the asynchronous node algorithm to form part of the local area network. RFID asynchronous algorithm used to solve the shortcomings such as the short life of the network and the redundancy of the collected data. The collected information including the salt chemical industry in pH, conductivity, dissolved oxygen, redox potential, BOD and ammonia and so on

In the random distribution network, it is assumed that all sensor nodes and sink nodes are homogeneous Poisson distributions consistent with the sum of the parameters. For a selected heterogeneous node, its network lifetime can be expressed as:

\[ L = \frac{E_i}{P} = \frac{E_i}{\lambda_0 - n_s (\epsilon_i + \epsilon_r) + \epsilon_i} \]  

(1)

\( E_i \) is a key node of the initial energy consumption; \( P \) represents the energy consumed by the key node in a data acquisition process; \( \epsilon_i \) represents the energy consumed by the sensor node to send a packet; \( \epsilon_r \) indicates the energy consumed by the sensor node to receive a packet; \( n_s \) is the number of key nodes; \( \epsilon_i \) indicates node consumption in the first between the packet and second packets of idle listening channel energy. The energy consumption of the network can be calculated by formula (1).

2.2. Data Storage and Processing Platform Based on Hadoop

Cloud platform is the ideal carrier for big data storage. Cloud computing technology is an important part of big data storage and processing technology. Its core is the parallel processing of massive data storage and data, namely distributed file system DFS (distributed file system) and MapReduce technology. Hadoop provides an open source HDFS technology that not only inherits the high fault tolerance of DFS and the advantages of deploying on low hardware, but also has the ability of high
reliability, high scalability and load balancing. The Hadoop contains a parallel programming model for open source implementation of MapReduce. Based on the above, the research group uses Hadoop to build the data storage and processing platform. At the same time, in order to meet the real-time requirements of information, the storage system is designed from the following four aspects: (1) According to performance and analysis requirements; (2) Real-time database system with real-time data with high performance requirements; (3) Using traditional parallel data warehouse system for core business data; (4) Adopting distributed file system for a large number of historical and unstructured data.

This paper proposes a multi-level storage system based on Hadoop for big data in salt chemical information, as shown in Figure 2:

At the same time, in order to solve the big data because of the large scale of data, analysis and processing time can not meet the real-time problem, research group intends to use the memory database (MMDB: main memory database) technology, which is placed directly on the data in memory database.

2.3. Data Analysis

The focus of the research is to analyze the massive data obtained in the salt chemical industry, so as to find the optimal salt production process, the minimum cost investment, the minimum environmental impact in order to maximize the economic and social benefits. The parameters of salt chemical production involves complicated, involving site location, technical scheme, equipment scheme and engineering project, the main raw materials, fuel supply, environmental impact assessment, labor safety, industrial hygiene and fire control, organization and human resources etc.. These factors from the economic, ecological environment, social benefits point of view there are interrelated, mutual restraint relationship, and the classification of these information itself is a vague concept. In view of the uncertainty, redundancy and fuzziness of big data, the research group adopts the combination of neural network and D-S evidence theory two-level fusion algorithm to optimize the output data. The design process of the two level fusion algorithm is shown in figure 3.
2.4. Realization of Big data Visualization Technology

The data from acquisition and storage to the statistical analysis finally is to serve the staff, in order to make big data of salt chemical analysis can meet the simple operation, reading, this paper uses the RIA (Rich Internet Application) technology to realize the visual characteristics on the client. Operators don’t have to care about data, how to carry out complicated data mining in the background, just need to operate easily on PC or other handheld devices. RIA has the features of desktop applications (interactive user interface, quick interface, response time, drag and drop characteristics, online and offline operations) and Web application features (immediate deployment, cross platform, etc.). Its communication features include real-time interactive voice and images, and data can be cached on the client, which can achieve a response speed based on HTML data from the server is faster and less user interface.

The client system uses one-stop portal technology and design management, the use of Microsoft's main browser can provide high video and audio, low-cost delivery of Silverlight technology to achieve resource consumption services portal. The front-end UI communicates with the "cloud" computing service and data service through HTTP, SOAP and other protocols. The service center interacts with other services through TCP (WCF), HTTP, SOAP and other protocols, and returns the result to the front end. The client processes and formats the results returned by the service call and displays the results in common text, tables, pictures, and other forms.

3. Big data Processing Framework Based on Micro Cloud

3.1. Overall Structure

There are many options for big data processing strategies, which are divided into two types: centralized processing and distributed processing. The cloud solution is based on the distributed processing, and a three-layer cloud structure is proposed for the distribution of the mixed region of the geographical region and the functional region. Each of the three hierarchical clouds runs the same function, corresponding to the three tier analysis framework for big data cloud processing, as shown in figure 4.
The outer layer of the cloud platform is preprocessed by data fusion, every cloud platform is in the face of massive data in a certain field, but these data may be abnormal, the contaminated data. Of course some abnormal data is true and valuable data, and need to be retained for further analysis. In this way, the cloud platform based on large amounts of data, excluding invalid data, while finishing data, based on local learning and model analysis, the sparse data fusion, the identification of uncertain data, incomplete data supplement, the initial mining of big data.

The middle layer of cloud platform plays an important role in sharing information and protecting privacy. The middle layer is connected to the outer and inner layers of the cloud platform, which only stores the output data of the outer layer and shares the big data. In addition, based on the privacy protection policy, the data is marked with privacy so that the inner layer can be used for reference.

As the big data mining platform, the kernel of cloud platform is computing and data access. This part is the key to big data processing. It is necessary to integrate the current data based on historical data, give effective results of big data calculation processing, and conclude the data input to the next level cloud platform.

3.2. Cloud based on Hadoop

Big data processing can’t be separated from cloud computing technology, cloud platform can have a variety of forms, which is a cloud to achieve the open source platform, including parallel computing model Map Reduce, distributed file system HDFS, and Distributed database Hbase, making Hadoop become a complete ecological chain system, Hadoop in big data and cloud computing between the built a solid and reliable bridge. In other words, large amounts of data is super computer processing object, and a large number of computers how to build a super computer like a system, its operating function system is Hadoop, micro-cloud system, the overall block diagram shown in Figure 5.
3.3. Performance Analysis
The data link features were retained with the massive data based on the main feature and Secondary feature, of course, the way for extremely small probability data will execute the cleaning process, with the characteristics of 5V data, simple statistics merging, big data, and obtained the information value of big data. Compared with several typical big data mining algorithms, the performance is shown in table 1.

| Algorithms          | Time complexity | Space complexity | Accuracy |
|---------------------|-----------------|------------------|----------|
| Genetic algorithm   | High            | High             | High     |
| Cloud computing     | Low             | Low              | Medium   |
| Feature simplicity  | Low             | Low              | High     |

3.4. Big data Processing Based on Micro Cloud
The cloud is located in the local area, which is used to deal with the big data of the region. In any region, because of the similarity of the user environment, the big data corresponding to the micro
cloud has strong temporal and spatial correlation. The passage of time, the vast majority of the time, are the same data, and the system of disturbance or abnormal events, will lead to data changes.

After the introduction of micro-cloud, for industrial production and environmental systems, micro-cloud corresponds to the system's functional subsystem, the formation of micro-cloud-based hierarchical distribution management structure. A certain micro-cloud, the management of a region of the data; a micro-cloud, the management of a number of sensors to run and maintain data; a micro-cloud, corresponding to the line monitoring and detection data.

In the hierarchical structure layer of cloud, big data in the region following the completion of treatment, while processing the relevant information to complete the transfer to the upper cloud, this layer treatment, comprehensive treatment and application form of the industrial production system.

3.5. Cloud performance for hierarchical distribution processing

Based on the system control, the three-level cloud structure of the total control room, the workshop organization and the field instrument is designed. The data processing flow of the whole local data is completed, the local power grid data is small, the information characteristic is obvious, Reflect the local information, but need a larger range of information to extract the data delivered to the upper processing, the formation of micro-cloud delivery to the next level of convergence cloud.

If a big data cloud portion corresponding to the data processing capacity of N, the complexity is O(N), and the processing time f (N) is equal to T. M micro-cloud classification distribution processing, and the processing time there is

\[ F(NM) \approx f(N) = T \]  \hspace{1cm} (2)

The magnitude is still T. If the system uses a large cloud of centralized processing, the large cloud corresponding to the big data for the NM, the processing complexity of O (NM), the processing time in the linear conditions for the NT, and more likely to increase nonlinear, there will be

\[ F'(NM) \geq NT \]  \hspace{1cm} (3)

The corresponding large cloud data processing algorithm is more complex and more resource consumption. Performance comparison shown in Figure 6.

![Figure 6 Comparison of hierarchical distribution and concentrated cloud performance](image)

4. Conclusion

Real-time dynamic monitoring of system traffic can provide a more comprehensive, complete and sophisticated system running state diagram to run and manage personnel through extensive analysis, diagnostics and optimization. If you give the appropriate auxiliary decision-making, control the implementation of programs and response plans, will be able to run the system and management to do more sophisticated, more accurate and more timely.

It is noted that the use of grid power has the characteristics of regional, based on electricity, the
design of the total control room, workshop agencies, field equipment, three micro-cloud structure, to solve the data generated by the large amount of data, mass data convergence Data block correlation extraction, processing delay and other issues, micro-cloud processing specific area data, the characteristics of a more clear, hierarchical distribution processing delay can be better control.

Acknowledgements
This work was financially supported by Enterprise information and Internet of things measurement and control technology of Key Laboratory of Sichuan province open project (2014WYJ05), Zigong science and Technology Bureau project(2014DZ10),project of Liquor making biological technology and application of key laboratory of Sichuan province(NJ2014-14), Major training project of Sichuan University of Science and Engineering(2014PY15)

References
[1] Wenpeng Mao, Yixi Yu, Bin Wang. AMI data analysis method[J]. Proceedings of The Chinese Society for Electrical Engineering ,2015,35(1):29-36.
[2] Hui Jiang, Kun Wang, Yihui Wang. Energy big data: A survey[J]. IEEE Access,2016,4:3844-3861
[3] Vaibhav Fanibhare, Vijay Dahake. Smart Grids: Map Reduce framework using Hadoop[C]. 2016 3rd International Conference on Signal Processing and Integrated Networks(SPIN), 2016:400-405.
[4] Yuanzhuo Wang. Status and Prospect of network big data[J]. Journal of Computer Science, 2013(36):1126-1130.
[5] Jonathan.V. Big Data Technology Literature Review[J]. Computer Science, 2015(12):7.
[6] Rajeshwari. State of the Art of Big Data Analytics: A Survey[J]. International Journal of Computer Applications ,2015(22):39.
[7] Chun-Wei Tsai1, Tsai1. Big data analytics: a survey[J]. Journal of Big Data, 2015(2):32.
[8] Vivek Kumar Singh. Scientometric mapping of research on “Big Data”[J]. Scientometrics, 2015(105):728.
[9] Zhang Haiyang, Li Pengju. Literature review of marketing theory based on Big Data[J]. Journal of Zhouyi research, 2014(5):44-48.
[10] Benjamin. Mining Big Data: Current Status, and Forecast to the Future[J]. SIGKDD exploration, 2015(14):2-3.
[11] Kim, Ibrar. Information fusion in government big data: Foundations, state-of-the-art, applications, challenges, and future research directions[J]. Information Management, 2016(04):3.
[12] J. Fiosina, M. Fiosins. Big data processing and mining for next generation intelligent transportation systems[J]. Sciences and Engineering, 2013(63):23-38.
[13] L. Jake. Big Data application in Biomedical research and Health care: a literature review[J]. Biomedical informatics insights, 2016(08):1-5.
[14] M. Mancini. Exploiting big data for improving health care services[J]. Journal of E-Learning and Knowledge Society, 2014(02):23-33.