Big Data Platform Development for Environmental Monitoring in Plant Factory

Daehee Park, Wei Liu, Sangwon Lee

Abstract*: *LED plant factory is a kind of labor-saving production, which is seldom controlled by natural conditions. Based on the mass data of the plant factory, this paper designed a big data management platform which combined with Hadoop distributed cluster and traditional MySQL database .The temperature, humidity, CO2 concentration and illuminance in plant factory can be transferred to the big data platform by wireless transmission network .And then the mass data will be stored, retrieved, and managed by Hadoop framework and the MapReduce programming model . Analysis result can be used to monitor plant growth environment and increase crop yield.

Index Terms*: *Big Data, Environmental Monitoring, Plant Factory, Platform Development, Smart Factory

I. INTRODUCTION

Plant factory is a new type production which using computer to automatic control the environment parameter. In order to ensure the good condition of plant growing, we need to monitor the environment and growth status of the plant in real time and get environmental data in time. On this foundation assessing plant growth environment and obtaining the optimal plan to promote plant growth. According to this purpose, in this paper, a low cost Big Data information acquisition and processing system is constructed by using Hadoop and WSN technology, which provides real-time monitoring, transmission, storage and management of multi-source heterogeneous environment data. Meanwhile, design algorithms to achieve massive data processing.

II. SYSTEM DEVELOPMENT PLATFORM AND KEY TECHNOLOGIES

A. Data characteristics of plant factory

With the continuous development of science and technology, modern technology is widely used in plant factory, so collected data has enormous amounts of information. The data has obvious Big Data characteristics. So, it has difficulties in information collection, transmission, storage and clustering decision making [1, 2].

- The amount of data during plant growth is enormous. There are huge and diverse sensor nodes in modern plant factory. Such as illuminance sensor, temperature and humidity sensor, camera and so on. In a large plant factory, the amount of data developed from TB level to PB level or ZB level.
- Data types are diverse. Data types are characterized by multi-source and heterogeneous features. Including text data, video, picture data, sensor data and so on.
- Fast processing speed. Large amounts of data in a plant factory can be processed by a distributed storage system.

B. Development status of cloud big data

Big Data technology includes data mining, machine learning, virtualization technology and so on. The Big Data system constructed in this paper mainly uses Hadoop platform to build large-scale clusters, which need to use MapReduce programming model, HBase database, HDFS distributed file system and other technologies [3]. Currently relatively mature storage analysis technology is the traditional relational database, such as Oracle database, MySQL database. Although the technology is mature, but it cannot support distributed processing requirements and it does not support concurrent access of users, expandability and sharing is not strong. Based on this, this paper constructs a Big Data storage, analysis and processing system based on Hadoop cloud, which can store large amounts of heterogeneous data in the system, and the data will not be lost. At any time and place, when the user login the system, he can query, analyze and calculate the data.

Big Data platform in this paper uses traditional relational database to store some fixed data and uses the Hadoop cluster to store various types of sensor data. It has the characteristics of high reliability, high fault tolerance and strong expansibility. At the same time, new data mining algorithms are developed to analyze and deal with mass data.

III. CONSTRUCTION AND DEPLOYMENT OF HADOOP PLATFORM

The network architecture of the Big Data processing Hadoop platform is shown in Figure 1.

Revised Manuscript Received on December 22, 2018.
Daehee Park, Department of Information & Communication Engineering, Wonkwang University, Iksan, Republic of Korea
Wei Liu, Department of Mathematics & Computer Science, Hengshui University, Hebei, China
Sangwon Lee, Department of Computer & Software Engineering (Institute of Convergence Creativity), Wonkwang University, Iksan, Republic of Korea, E-mail:sangwonlee@wku.ac.kr

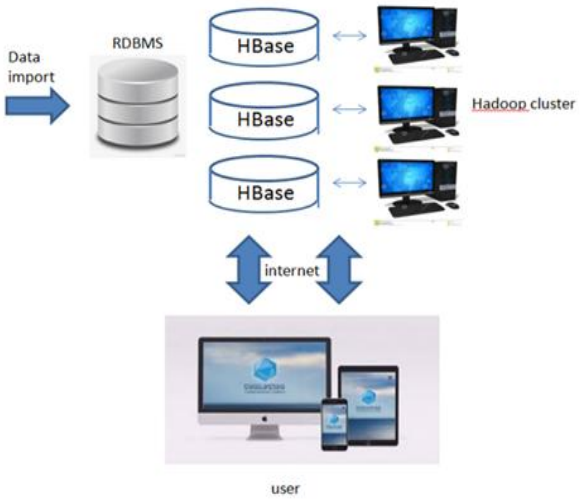

Fig 1 Network architecture of the big data processing Hadoop platform

A. System architecture

The environment automatic control and data acquisition system in the plant factory is constructed through Arduino chip and many sensors.

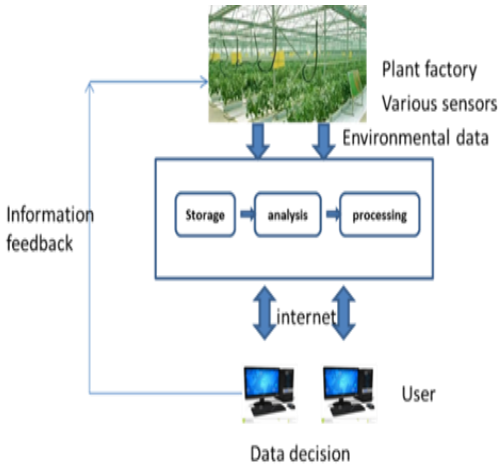

Fig. 2 Model of data acquisition and processing

The structure is shown in Figure 2.

B. Data transmission in sensor networks

Sensor nodes are deployed in the monitoring area to form a monitoring network. When the monitoring data arrives at the sink node, the data will be connected to the network by wire or wireless, and transmitted to the management node, such as temperature, humidity, water, video and so on, these data will be stored simultaneously in the cloud Big Data platform. Managers and users can monitor, query, analyze and process the environment information through the terminal device.

C. Hadoop ecosphere

Hadoop is an open source distributed framework, and its ecological community is shown in Figure 3. It can store and calculate large data sets in a distributed way by deploying them on a number of common computer clusters. It has the advantages of high expansibility, powerful function and high cost performance. The core of Hadoop is MapReduce computing framework, HBase database, and HDFS (Hadoop Distributed File System).

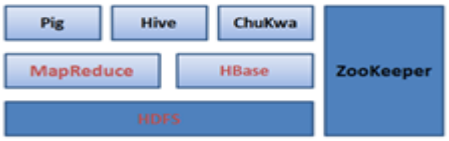

Fig 3 Ecological community of Hadoop

The idea of Hadoop is calculated in data nodes, so MapReduce computing framework, HBase database, and HDFS are running on the same cluster node [4].

D. Deployment of the Hadoop architecture

We use PC to build the server cluster environment, including the main computing node and the secondary computing node. Install Hadoop ecological cluster on all nodes and Select two of these computers as the primary node, when the primary node crashes, another host can continue to work. The rest of the computers act as child nodes and the minimum configuration for each computer is 4G memory, 500G hard disk. Due to the high scalability of Hadoop, the number of nodes can be increased in the late stage, and the storage of large amounts of data can be guaranteed without shutdown. Deploy a Hadoop cluster on a computer cluster that has already installed a Linux environment and a JAVA environment.

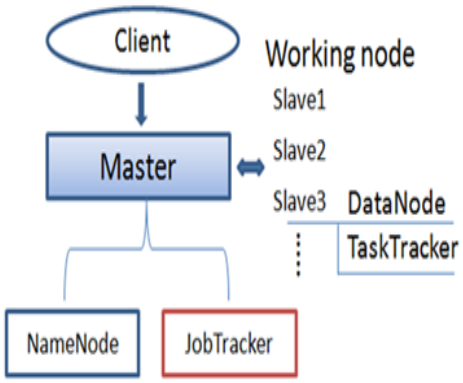

Fig 4 Deployment of Hadoop

The deployment of the architecture is shown in Figure 4.

IV. RESULTS AND DISCUSSIONS

Design of database and data storage layer

A. Database design

Data collected from a plant factory is diverse, multi-source. Considering the needs of cloud management, the system can combine the distributed HBase and MySQL database to support the real-time reading and writing.

RDBMS is suitable for continuously updated data sets. When the data set is indexed, it can retrieve data quickly and update data for small data sets. So, we can combine relational databases MySQL and HBase to process some structured data.

Most of the massive data and a large number of unstructured data stored in the distributed system, and a small amount of real-time data stored in the relational database. In the concrete implementation process, we can use Sqoop and other tools to import the sensor information of MySQL database into HDFS and HBase periodically.

B. Design of sensor data sheet in Hbase database

HBase is a column oriented distributed database. It can design the storage structure of the table according to the user's needs and complete the distributed storage of mass data in the plant factory. When collecting data, we need to use HBase JAVA API to complete the data collection and storage. Table 1 and table 2 show the humidity sensor information storage logic diagram and the humidity data storage logic diagram respectively.

C. Storage of other types of data

In addition to temperature, humidity, illumination and other data, video and images are non table structured data. Video and image storage technologies of HBase enable it to have good video and image storage capabilities [5].

Table 1 Humidity sensor information storage logic diagram

Rowkey	Timestamp	Column Family: Sensor Information	
ID	TS1	Column	Value
ID	TS2	Location	Factory1
ID	TS3	SensorNum	1
ID	TS4	SensorType	Humidity
ID	TS5	Time	0:00

Table 2 Humidity data storage logic diagram

Row key	Timestamp	Column Family: Data	
		Column	Value
INfoID	TS1	Data:hum	60%RH

Through the API structure provided by Hadoop, the accepted video and picture stream files are uploaded from the local to the HDFS. Underlying storage system of HBase is HDFS, so it has a good docking with HDFS.

Design of visual management function

Managers and users can implement specific data information management functions through visual interfaces. The management system mainly includes user management module, data management module and cluster management module. The data management module mainly implements monitoring data uploading, downloading and deleting data into HDFS, HBase and MySQL. Based on this, the distributed management and service of heterogeneous multi-source plant environment information are realized.

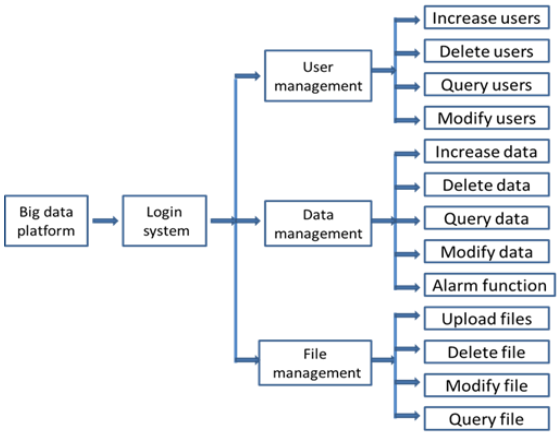

Fig 5 System flow chart

The structure is shown in Figure 5.

Execution process and test

The specific execution flow is the query module, which judges the query value according to the user's request connects the HBase cluster according to the query value using different query methods. The process needs to be implemented through the API interface of the HBase.

HBase mainly uses the scan method for queries. According to the query condition, the number of records is returned, and the returned result set is put into the ResultScanner () object by the getScanner () method and specify a page to display the results, so the query interface is realized. For the returned data, JavaWeb technology is used to extract the data from the database, then the data is stored into the JavaBean entity. Finally, the system uses Servlet for data acquisition. If the sensor environment parameter exceeds the system pre-set threshold, the system will alert the prompts. At the same time, we can also get data analysis and prediction results by Big Data platform.

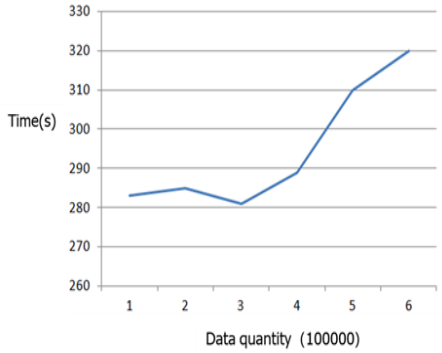

Fig 6 Time consumption of data import

We use client test programs to record and display the time consumed by importing data into the HBase. If each data inserted into HBase is less than 1MB and each time about one hundred thousand data volumes are inserted, the consumption time is relatively stable, about 250s-340s, as shown in Figure 6.

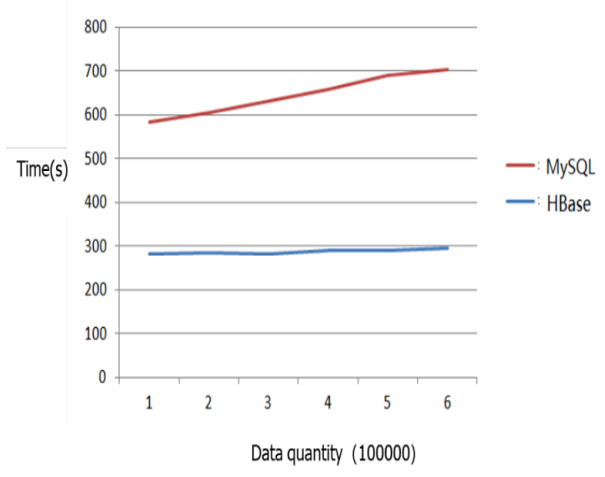

Fig 7 Comparison of time consumption

Comparing the performance of HBase and MySQL in Figure 7, we can see that when the amount of data is relatively large, the overall performance of MySQL database is relatively low compared with the cluster system. With the increasing number of imported data records, its stability and scalability are also gradually reduced. However, the query time of HBase database has been relatively stable, and the efficiency is much higher.

Data mining algorithm

For specific data processing, the K-means clustering algorithm can be used.

A. K-means algorithm

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining [6]. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given an initial set of k means $m1^{(1)}$, …, $mk^{(1)}$ (see below), the algorithm proceeds by alternating between two steps [7]:

(1) Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean (where each xp is assigned to exactly one $s^{(t)}$, even if it could be assigned to two or more of them).

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\},$$

(2) Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

B. Testing and comparison of algorithms

We apply the K-means algorithm to the Big Data platform, and collect the different data samples from the plant factory. So we can get the time consumption of the K-means clustering algorithm.

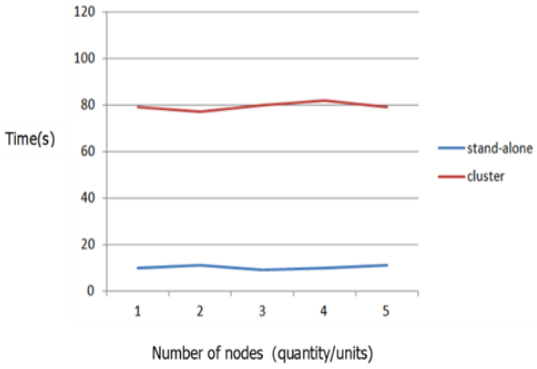

Fig 8 Consumption time of 10^6 records

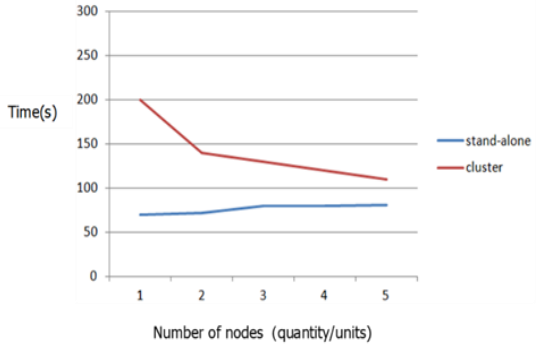

Fig 9 Consumption time of 10^7 records

Figure 8, Figure 9, and Figure 10 shows the elapsed time of running K algorithms for different samples. We can see that with the increase of computing nodes, both Hadoop cluster and single machine computing time will increase. When the amount of data is small, the initialization of MapReduce computing takes a certain amount of time, so the cluster computing time will be more time-consuming than the single machine. But with the increase of calculation, the advantages of clustering are gradually displayed, and the consumption time is greatly reduced, the speed is obviously improved.

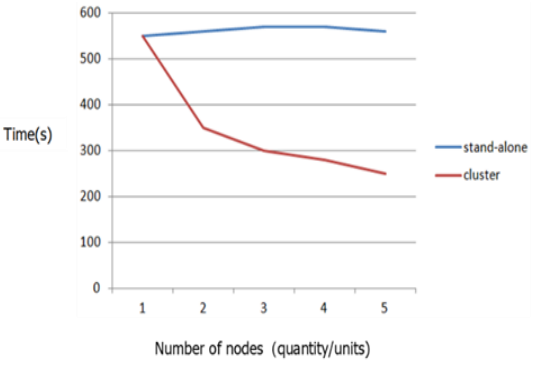

Fig 10 Consumption time of 10^8 records

Thus, the advantages of Big Data platforms and algorithms used in dealing with large amounts of data are obvious. With the increase of nodes, the speedup ratio increases, and the computing performance gradually improves.

V. CONCLUSIONS

In this paper, a monitoring system of plant environment based on Big Data is constructed. The system uses sensors to collect environmental parameters such as temperature, humidity, light intensity, environment and so on in the plant environment. The collected data can be analyzed and processed on the Hadoop platform using an improved data mining algorithm, So, as to realize the real-time monitoring of crop growth environment, early warning of crop growth environment, and find the most suitable environment for plant growth.

ACKNOWLEDGMENT

This paper was supported by Wonkwang University in 2016.

REFERENCES

1. L. Li, X. Jing, and S. Li, “Research and design of SASS platform for agricultural information integration service”, Agricultural Chemical Journal, Vol. 34, No. 6, 2013, pp. 244-248.
2. L. Chen, Y. Hu, and F. Zhang, “Performance improvement of cloud computing technology in agricultural product safety system”, Journal of Agricultural Engineering, Vol. 29, No. 24, 2013, pp. 268-274.
3. S. Kumar, “Evolution of Spark Framework for simplifying Big Data Analytics”, Proceeding of the 2016 International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 3597-3602.
4. Ghorbaniparsa, Fatemeh, and Hamideh Ofoghi. "Comparing Patatin Class I and Camv 35s Promoters in Expression of Human Calcitonin Gene in Potato (Solanum Tuberosum Cvs. Kardal And Marfona)." The International Journal of Biotechnology 5. 4 (2016): 52-61.
5. J Yoon, D. W. Jung, C. H. Kang, and S. Lee, “Forensic Investigation Framework for The Document Store NOSQL DBMS: MongoDB as a Case Study”, Digital Investigation, 2016, pp. 18-20.
6. Haseeb, M., Abidin, I. S. Z., Hye, Q. M. A., & Hartani, N. H. (2018). The Impact of Renewable Energy on Economic Well-Being of Malaysia: Fresh Evidence from Auto Regressive Distributed Lag Bound Testing Approach. *International Journal of Energy Economics and Policy, 9*(1), 269-275.
7. Nayyar, Anand, and Vikram Puri. "Comprehensive Analysis & Performance Comparison of Clustering Algorithms for Big Data." Review of Computer Engineering Research 4. 2 (2017): 54-80.
8. P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data”, Procedia Computer Science, Vol. 78, 2016.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication