Managing Sensor Data Using Mapreduce

KUNG WU PARK[1], KUNG O. KIM[2], ROHIT GANGPA[3]
[1, 2] *Informatics and Engineering Department, Chonnam National University, South Korea*
[3]*Information Technology, Amrapali Institue of Technology, India*

Abstract—Recently, as the construction of a large-scale sensor network increases, a system for efficiently managing large-scale sensor data is required. In this paper, we propose a cloud-based sensor data management system with low cost, high scalability, and high efficiency. In the proposed system, sensor data is transmitted to the cloud through the cloud gateway, and abnormal situation detection and event processing are performed at this time. The sensor data sent to the cloud is stored in Hadoop HBase, a distributed column-oriented database, and processed in parallel through the MapReduce model-based query processing module. As the processed result is provided through REST-based web service, it can be linked with application programs of various platforms.

Indexed Terms — sensor data management, cloud computing, hadoop, hbase, mapreduce.

I. INTRODUCTION

Recently, with the development of wireless communication devices and sensor technology, a large-scale (Ubiquitous Sensor Network) can be built, and methods to utilize it in various application fields are being studied [1][2][3][4]. For example, in order to build an urban environment monitoring system, a large-scale sensor network that collects various environmental data such as temperature, humidity, wind speed, carbon dioxide concentration, and noise from various places such as roads, rivers, parks, and waste treatment plants is required. It is necessary to build a sensor data management system that manages the collected large-scale sensor data. Existing research uses a distributed database system on a single server or multiple servers built as a grid to manage sensor data, so system expansion is not easy and system construction and management costs are high.

Cloud computing has recently been in the spotlight as a method to store and manage large-scale data. Cloud computing is computing that provides 'virtualized IT resources as a service using Internet technology. Users borrow and use IT resources (software, storage, server, network) as much as they need, and pay as much as they use. says [5][6][7]. Management using cloud computing has two major advantages. First, it is highly scalable. Through the cloud, users can receive as much storage space as needed and computing power for data processing in real-time. The second is the cost part. Since you only have to pay for the resources you use, you can reduce the cost for construction and server operation. With these advantages, research on applications using the cloud is being actively conducted, and in particular, after Google's MapReduce parallel processing model was announced, the field has been expanded to the field of scientific computation [8].

Recently, research using cloud computing as a way to store and process large-scale sensor data is being conducted. In [11], the integration of the Internet and the sensor network was proposed, and in [12], a framework for combining the sensor network and the web-based application program using the cloud was presented. [13] presented a medical system through a cloud server. Existing studies focus on the combination of the sensor network and the cloud, so there is a lack of research on how to store large-scale sensor data or parallel processing. In addition, since outlier detection and event processing that requires a quick response are performed in the cloud, the response time is delayed.

In this paper, we propose a sensor data management system that distributes and stores sensor data collected from a large-scale sensor network and performs parallel processing. The proposed system transmits the sensor data collected from the sensor network to the cloud through the cloud gateway. At this time, the cloud gateway converts the data collected from heterogeneous sensors into standardized messages and transmits them to the cloud.

Sensor data sent to the cloud is stored in Hadoop HBase, a distributed column-oriented database. Distributed column-oriented database has good scalability and parallel processing of MapReduce because it is distributed and redundantly stored in multiple nodes [15]. In this paper, we present a data schema for efficient search and design and implement a MapReduce-based sensor data processing module. The user transmits a query through the REST (Representational State Transfer)-based web interface [16] and is designed to receive the results so that it can be linked with application programs of various platforms.

The structure of this paper is as follows. Section 2 examines Hadoop HBase and MapReduce parallel processing models as related studies. Section 3 describes the cloud-based sensor data management system proposed in this paper, and Section 4 describes the conclusion and future research directions.

II. PREVIOUS WORKS

This section describes the cloud technologies for the sensor data management system proposed in this paper. First, we describe Hadoop HBase, a distributed column-oriented database for storing sensor data, and a MapReduce model for parallel processing of sensor data.

A. HBase

HBase is a distributed column-oriented database implemented in Hadoop HDFS (Hadoop Distributed File System), modeled after Google's BigTable [14][15]. HBase provides real-time random access to large data sets and supports a range of searches in recent versions. In addition, it supports both the batch processing method calculation using MapReduce and the point query method that enables random access.
In HBase, data is stored in tables. A table is composed of rows and columns, and a table cell is the intersection of a row and a column and stores data in the form of a raw byte array.

A row in a table consists of a row key and a column, and the columns are grouped into a column family. Rows are identified by the row key, which is the primary key of the table.

It is sorted and basically sorted in byte order, and all table access is done through the row key of the table. A column family has one or more members. existing HBase has various restrictions on large-scale scalability and distributed processing, HBase can be expanded linearly by adding only nodes. In addition, parallel processing of large data sets is possible using MapReduce, enabling fast data processing.

B. MapReduce

Parallel processing technology, such as, is focused on applications that require high-performance computing, so it is difficult to apply when there is a large amount of data to be processed. In addition, for large-scale parallel processing of data, scalability should be easy according to data increase, and job distribution should be made to minimize network traffic due to data movement between nodes. MapReduce is a parallel processing model announced by Google that considers these issues. MapReduce is a model that parallelizes key-value-based data. It generates intermediate results by executing each Map task in parallel on input data that is divided and stored in multiple nodes to generate intermediate results. It consists of two steps to calculate the final result by executing the Reduce task.

The Apache Group's Hadoop system developed this MapReduce model as an open-source. Hadoop is currently used by several companies such as IBM, Adobe, Powerset, and Facebook.

III. MANAGING SENSOR DATA USING MAPREDUCE

This section describes the cloud-based management system for large-scale sensor data proposed in this paper.

A. Cloud Gateway

The cloud gateway server plays a role in transmitting a large amount of sensor data generated from multiple sensor networks to the cloud. The cloud gateway converts heterogeneous sensor data into a standardized message (XML) through the sensor network common interface, transmits it to the cloud, and stores the sensor data for a certain period in the local storage. And, through the abnormal situation detection and intelligent event processing module, control signals and event messages for emergency situations are transmitted to sensors or managers. At this time, if

sensor data of a previous point in time such as a time series query is required, abnormal situation detection or intelligent event processing is performed within the gateway by utilizing the sensor data of the previous point in time stored in local storage. Cloud server has high data processing capacity, but it is difficult to guarantee fast response. Therefore, as the cloud gateway takes charge of this part, quick response processing is possible.

A. Sensor Data Schema

Our schema utilizes a different column-based data storage system for scalability and distributed storage. A table consists of row keys and columns, and columns are grouped into column families. The data in the table is sorted by row key, and data access is done through the row key. Therefore, for efficient storage and fast retrieval, a schema design different from that of the existing RDBMS is required. The sensor_id table is a table that stores sensor information. The Sensors belong to specific groups according to the user, location, and sensor type. A low key is composed by combining group_id and <sensor_id to facilitate sensor search and sensor search by the group. And, in the column family, there are sensor_info, which stores sensor information, and location, which stores sensor location information, and stores information in the sensor.

IV. SYSTEM INTERFACE

The system proposed in this paper provides a REST-based web interface to enable interworking with application programs of various platforms. Figure 1 shows the sensor data management client. The interface was devised so that sensor data could be searched under various conditions such as sensor ID, group ID, sensor type, time, and value size. Transmits the nut in the form of an XML document. Weekly transmits the received data to Java. The sensor list and information about them are output in a table format. When a sensor is selected, detailed information about the corresponding sensor and the sensor data values at the time of search are visualized and output in a chart format.

The sensor_data table is a table that stores the collected sensor data, and the row key is composed by combining the recently collected data search and range quality timestamp values.

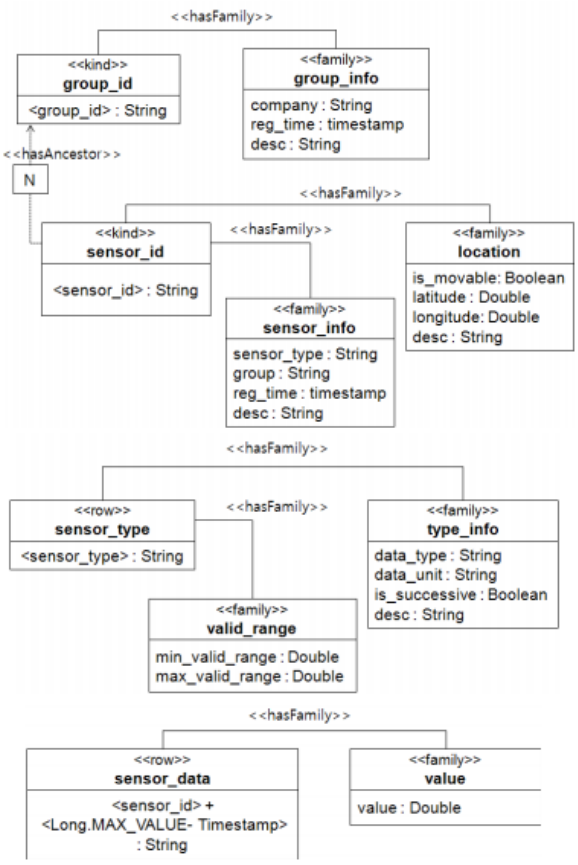

Figure 1. Sensor data schema

HBase distributes and stores sensor data across multiple nodes in the cloud. The sensor data stored in the HTable of each node is processed in parallel by the MapReduce model as shown in Figure 2. The sensor data is distributed and stored in the HTable of each node in the cloud, and user queries are executed in parallel at each node by TableMapper. The key-value list results obtained through the table mapper of each node are sorted and replicated and transmitted to the TableReducer. The table reducer merges these results and stores the final result in a table.

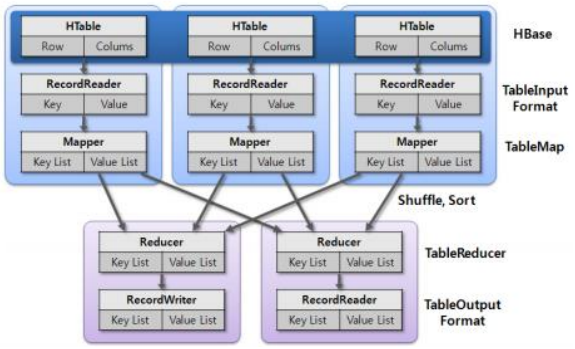

Figure 2: MapReduce execution

CONCLUSION

Since the sensor network continuously generates data from the moment it is built, the server that manages sensor data must have high scalability and data processing capability.

In this paper, a cloud-based sensor data management system that satisfies these requirements at a low cost is presented. The proposed system uses Hadoop HBase, a column-oriented database, to distribute and store large-scale sensor data in the cloud. A data schema for efficient search and management was proposed, and a sensor data-parallel processing module using the MapReduce model was implemented. In addition, it is designed to send a query through a REST-based web interface and receive the result.

REFERENCES

[1] K. Aberer, G. Alonso, and D. Kossmann, "Data Management for a Smart Earth - The Swiss NCCR-MICS initiative", SIGMOD Record, Vol. 35, No. 4, pp. 40-45.

[2] K. Aberer, M. Hauswirth, and A. Salehi, "Infrastructure for Data Processing in Large-Scale Interconnected Sensor Networks", 2007 International Conference on Mobile Data Management, pp. 10-11. 198-205, May.

[3] C. Jardak, J. Riihijärvi, and P. Mähönen, "Extremely Large-Scale Sensing Applications for Planetary WSNs," In Proceedings of the 2nd ACM International Workshop on Hot Topics in Planetary-Scale Measurement, pp. 10-11. 1-6, June.

[4] K. Aberer, G. Alonso, and D. Kossmann, "Data Management for a Smart Earth", SIGMOD Record, Vol. 35, No. 4, pp. 40-45.

[5] B. Hayes, "Cloud Computing", Communications of The ACM, Vol. 51, No. 7, pp. 9-11, July.

[6] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and Grid Computing 360-Degree Compared", Grid Computing Environment Workshop(GCE'08), pp. 1-10, Nov.

[7] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities", Proceedings of the 2008 10th IEEE International Conference on High Performance Computing and Communications, pp. 5-13.

[8] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters", In OSDI'04: 6th Symposium on Operating System Design and Implementation, December.

[9] S. Manakkadu, S. P. Joshi, T. Halverson, and S. Dutta. "Top-k User-Based Collaborative Recommendation System Using MapReduce." In 2021 IEEE International Conference on Big Data (Big Data), pp. 4021-4025. IEEE, 2021.

[10] I. Hwang, K. Jung, . Im, and J. Lee, "Improving the Map/Reduce Model through Data Distribution and Task Progress Scheduling," Journal of the Korea Contents Association, Vol.10, No.10, pp.78-85.

[11] H.-C. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, "Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters," Proc. of the ACM SIGMOD International Conference on Management of Data.

[12] S. Ghemawat, H. Gobioff, and S. Leung. "The Google File System," Proc. of ACM Symposium on Operating Systems Principles, pp.29-43.

[13] H. Zhao, S. Yang, Z. Chen, S. Jin, H. Yin, and L. Li, "MapReduce Model-Based Optimization of Range Queries," Proc. of the International Conference on Fuzzy Systems and Knowledge Discovery(FSKD '12), pp.2487-2492.

[14] H. Liu and D. Orban, "GridBatch: Cloud Computing for Large-Scale Data-Intensive Batch Applications", Proc. 8th IEEE International

Symposium on Cluster Computing and the Grid (CCGRID'08), pp. 295 305, May.

[15] C. Vecchiola, S. Pandey, and R. Buyya, "High-Performance Cloud Computing: A View of Scientific Applications", 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks, pp. 4-16.

[16] V. Rajesh, J. M. Gnanasekar, R. S. Ponmagal, and P. Anbalagan, "Integration of Wireless Sensor Network with Cloud", 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 321-323.

[17] M. Hassan, B. Song, and E. Huh, "A Framework of Sensor-cloud Integration Opportunities and Challenges", Conference On Ubiquitous Information Management And Communication, pp. 618-626.

[18] X. Lee, S. Lee, PT True, LT Vinh, AM Khattak, M. Han, DV Hung, MM Hassan, M. Kim, K. Koo, Y. Lee, and E. Huh, "Secured WSN - integrated cloud computing for u-life care", In Proceedings of the 7th IEEE Conference on Consumer Communications and Networking Conference, pp. 702-703, Jan.

[19] Hadoop: http:/ /hadoop.apache.org

[20] HBase: http://hadoop.apache.org/hbase