# Distributed Storage and Processing Method for Big Data Sensing Information of Machine Operation Condition

Fan Zhang, Zude Zhou, Wenjun Xu
School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China;
Key Lab. of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan 430070, China
Email: fan_dx0202@163.com, zudezhou@whut.edu.cn, xuwenjun@whut.edu.cn

***Abstract*—The traditional relational database cannot satisfy the requirements of the high speed and real-time storage and processing for the distributed Big Data sensing information in the Wide Area Network environment. In this context, the No-SQL database HBase is used to store the big data sensing information of machine operation condition collected by Fiber Bragg Grating sensor network. The distributed storage environment and the optimal database table scheme is built. Moreover, the HBase Rowkey is designed in detail to sharpen the retrieval speed and avoid the server hot point accumulation. Meanwhile, the real-time outlier detection method with the working situational constraint is proposed to monitor the machine working condition. It is implemented by the multi-dimensional histogram statistics method in the Map Reduce distributed environment. Hence, the traditional threshold monitoring method is improved and the false alarm problem is eliminated. Through balancing the performance between HBase and the relational database MySQL, the real-time storage rate of the proposed method can satisfy at least 20 machines running concurrency with 4000 Hz Fiber Bragg Grating sampling frequency by HBase. Also, the effectiveness of real-time outlier detection method is proved by the practical operation data processing.**

***Index Terms*—Machine operation condition, Fiber Bragg Grating sensing, Big Data, distributed storage and processing**

## I. INTRODUCTION

The real-time condition monitoring for mechanical equipment operation safety has become a research focus recently. As the rapid development of manufacturing industry, the manufacturing task is usually no longer confined to single equipment, but extended to run in parallel and cooperative equipment cluster. In this context, the equipment fault not only brings huge economic damage but also result in environmental and social effects. Due to the manufacturing resources diversity, dynamics and distributivity in the WAN (Wide Area Network) environment, it is very important to monitor the machine operation condition and manage the manufacturing task. The aim is to make full use of the real-time sensing information, manage the machine operation condition comprehensively and ensure the manufacturing tasks to be finished on time.

Sensing technology is the information foundation to realize the overall perception of mechanical equipment operation condition. Due to the high performance index, complex equipment structure and harsh working condition, especially for large equipment working at high temperature, high speed and multi coupling field environment, the traditional sensing technology is difficult to achieve real-time dynamic operation condition monitoring in a long term as its defects of electromagnetic interference, big volume or less measurement parameters. Because of the special advantages of FBG (Fiber Bragg Grating) sensor, it is suitable for the overall monitoring of machine operation condition, especially for the mechanical system working in the harsh environment of time-varying, field coupling and high load [1,2]. The FBG sensor network can be deployed on the mechanical equipment to realize real-time and distributed operation condition monitoring in the WAN environment. Hence, the mechanical equipment operation condition can be mastered at anywhere and anytime. In the practical application, the sensing information features of mechanical equipment operation condition based on FBG sensor network are shown as follows:

**i) Real time:** The timeliness is important to the physical meaning of the sensing information. The relation between sensing information and time is close.

**ii) High speed:** The sampling rate is set in millisecond or microsecond usually. Especially for the high speed rotating machine, like aero-engine, stream turbine et al., the FBG sensor network sampling rate is up to 4000Hz to meet the needs of overall perception.

**iii) Massive information:** The large mechanical equipment usually works for many years or even decades. The sensing information storage space will be hundreds of T ($10^{12}$ bytes) or even P ($10^{15}$ bytes).

**iv) Data diversity:** The sensing information includes many types of parameters, such as temperature, stress-strain, mechanical vibration, etc.


doi:10.4304/jsw.9.10.2598-2606

Real time, high speed, massive information, and data diversity are also the basic features of big data. For example, if the FBG sensor network deployed on a kind of machine includes 100 FBGs with 4000Hz sampling rate, the size of sensing information will be 24 million rows in per minute. If there are 50 same machines in the manufacturing WAN environment, the sensing information will be 1728 billion rows one day. Although the traditional relational database has high performance for complex business data storage and management, it is incapable to this kind of high speed big data. Therefore, it puts forward a new requirement of the big data sensing information real-time storage and processing for the machine operation condition [3].

The remainder of the paper is organized as follows. Section 2 describes the related work and key problems about the big data storage and processing technology. For the purpose of massive sensing information distributed real-time storage and processing of FBG sensor network machine operation condition, the big data sensing information distributed storage method is presented in Section 3. Section 4 discusses the real-time outlier detection method with the working situational constraint by Map-Reduce distributed processing mechanism to monitor the machine operation condition from the point of big data mining. The experiment results and analysis of the proposed distributed storage and processing method for FBG big data sensing information are presented in Section 5. Finally, Section 6 concludes the paper.

## II. Related Work

As the development of fiber optic sensing technology, it has been widely used in a number of application fields for safety monitoring, such as aerospace, large-scale bridges, tunnels, dams, mechanical equipments et al., and performed a wide range of advantages. In recent years, researchers keep exploring to use the FBG sensing technology in the area of mechanical equipment online and dynamic condition monitoring. USA military used FBG to monitor the ship working condition [4]. The FBG sensor network is also applied to monitor the aero-engine and spacecraft structural health in Canada, Japan and Israel [5]. In China, it is used for large manufacturing equipment online condition monitoring, such as dumper, large cranes, aero-engine, steam turbines et al., and has achieved a series of research results [6,7]. FBG sensor network with its excellent performance has provided a new method for the equipment overall operation condition monitoring in various environments.

In the manufacturing equipment assemblage environment based on FBG sensor network, the sensing information owns the big data characters of volume, velocity, variety and value. In recent years, many scholars have conducted research on the big data transmission, storage and processing algorithms [8]. Especially, the Hadoop MapReduce framework, GFS (Google File System) distributed data storage mechanism and semi structured data storage and management system have been used to deal with the big data problem and also become the hot research topics [9,10]. In practical, the big data research focus on two aspects mainly: the specific industrial application and big data processing architecture performance improvement [11]. In [12], the disadvantages of traditional relational database and the advantages of the No-SQL database for big data problem are discussed. But it doesn't talk about the solving process in detail. In [13], an implementation of parallel heuristic search algorithm for solving big data combinatorial problems by Map Reduce is described. In [14], the parallel FP (Frequent Pattern) growth algorithm based on Key-Value storage method is used for the massive manufacturing resource in the cloud manufacturing mode. Also, according to the DaaS (Data-as-a-service) concept, the big data analysis service is executed in the data cloud by the Web Service way in [15]. Literature [16] discusses the big data processing and storage key issues in the cloud computing environment, but there is no specific solution. In [17], it presented a performance anomaly detector method which is used to detect performance anomaly and accurately identify the faulty sources in an I/O server of cluster file systems, but the timeliness cannot be guaranteed in the high speed big data context. The big data application in wireless sensor network is introduced in [18]. According to the characters of wireless sensor network and sensor information format, the big data storage method based on HBase is proposed. But the sensor network features and sensing information format are completely different from FBG sensor network. In [19], the K-means method is used to improve the Map Reduce cluster performance. From the current related work, the No-SQL data management and analysis technology based on Hadoop Map-Reduce framework has excellent performance for the big data problem with its scalability, fault tolerance and parallel processing. It is suitable to the big data sensing information storage and processing using the FBG sensor network.

Although there are some achievements about this research, it is only limited to no-manufacturing environment application or the Hadoop open source structure improvement itself. The research about the big data sensing information real-time storage and processing for mechanical equipment based on FBG sensor network in the WAN environment is also relatively few. With the application of big data technology for manufacturing machine operation condition sensing information, it can not only master the equipment working condition and ensure the equipment safety and stability, but also provide data basis for manufacturing task lifecycle management and production decision making.

## III. Sensing Information Storage Method Based On Hadoop HBase

### A. HBase Distributed Storage Architecture

HBase is the Apache Hadoop database which can provide big data real-time and random reading and writing functions with the features of open source, distributed deployment, expandability and column oriented. HBase is developed based on Google Bigtable.

It's the perfect combination of big data storage and parallel processing by Hadoop MapReduce framework.

The HBase distributed storage architecture is shown in Figure 1. From the figure, HBase service system is composed of HRegion server cluster and HBase master server with the master-slave structure. The task of HBase master server is to manage the HRegion server cluster, coordinate each server by ZooKeeper and process the running error. HBase region server is divided by some HRegions. Sensing data is stored in the HRegion cluster. Data mapping relationship between data itself and HRegion server is also stored in HBase Master Server.

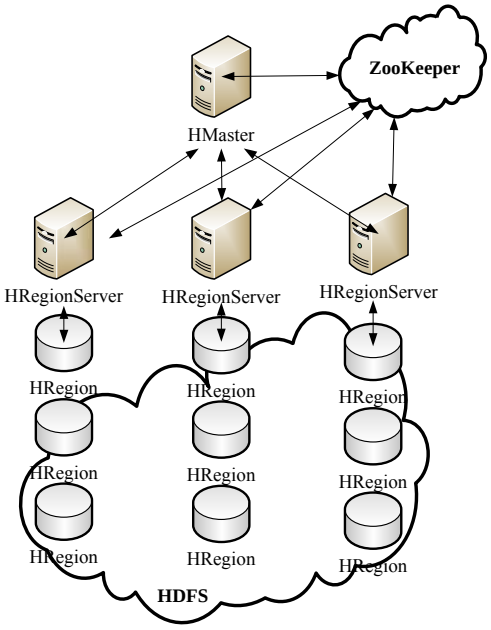


Figure 1. HBase distributed storage architecture

### *B. FBG Sensing Information HBase Table Structure*

By analyzing the FBG sensor network configuration and the FBG sensing information composition, according to the FBG sensor network structure characters, the HBase table structure is shown in Figure 2. The whole network includes several sensing channels. In each channel, it is composed of a series of FBG sensor nodes. In addition to the sensing variants collected by the FBG sensor nodes, such as temperature, stress-strain, vibration et al., the mechanical equipment ID, acquisition time and sensor ID also should be included in the sensing information. Considering the real-time processing requirements and the off-line retrieval effectiveness of the big data sensing information in the distributed environment, the database table concept view structure is shown in Table I. The physical view of the first two rows is shown in Table II. In the table, 'SValue' is the real-time sensing variation and 'CBound' is the working condition information. According to the HBase database character, the NULL columns in the concept view won't be stored in the physical view. So, in order to save the storage space, the value of 'CBound' can be NULL if there is no change about the working condition.

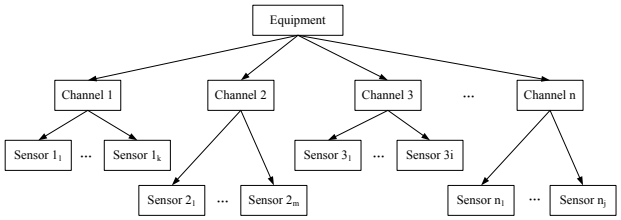


Figure 2. FBG sensor network structure

TABLE I
HBASE DATABASE TABLE CONCEPT VIEW STRUCTURE

| Row key | Time stamp | Column Family: SValue | | Column Family: CBound | |
|---|---|---|---|---|---|
| | | Column | Value | Column | Value |
| r1 | t4 | SV:1 | value1 | CB:1 | value1 |
| r2 | t3 | SV:2 | value2 | NULL | NULL |
| r3 | t2 | SV:3 | value3 | CB:2 | value2 |
| r4 | t1 | SV:4 | value4 | NULL | NULL |

TABLE II
HBASE DATABASE TABLE PHYSICAL VIEW

| Row Key | Time Stamp | Column Family: SValue | |
|---|---|---|---|
| | | Column | Value |
| r1 | t4 | SV:1 | value1 |

| Row Key | Time Stamp | Column Family: CBound | |
|---|---|---|---|
| | | Column | Column |
| r1 | t4 | CB:1 | value1 |

| Row Key | Time Stamp | Column Family: SValue | |
|---|---|---|---|
| | | Column | Value |
| r2 | t3 | SV:2 | value2 |

The Rowkey design of the table is the key issue which concerned with the storage size and the retrieval effectiveness. So, it should be considered from the Rowkey code length, hash code and uniqueness field, three aspects. Because the data persistence file HFile is stored by the Key-Value mechanism, greatly storage space will be occupied if the Rowkey is too long. For example, if the Rowkey length is 100 bytes, 10 million rows of data will occupy 100*10 million=10 billion bytes, nearly 1 G bytes. In this situation, the storage efficiency will be influenced greatly. In addition, partial data will be cached to the memory when HBase is working, the memory utilization also will be reduced if the Rowkey is too long. Meanwhile, for HBase is a distributed B tree in fact, if there is no hash field, all the new data will be stored in one Regionserver and the hot point accumulation phenomenon will be happened. In this case, the data retrieval load will be focused on a specific Regionserver and the query efficiency will be reduced greatly. Also, as the increasing number of client, the single storage node is easy to be down. Above all, the Rowkey constituent components are:

| Hbase Rowkey | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1th byte | 2th byte | 3th byte | 4th byte | 5th byte | 6th byte | 7th byte | 8th byte | 9th byte | 10th byte |
| Hash field | | Device ID | Sensor ID | Time field (Reverse chronological) | | | | | |
| 0~65535 | | 0~255 | 0~255 | Year (0~99) | Month (1~12) | Day (1~31) | Hour (0~23) | Minute (0~59) | Second (0~59) |

Figure 3. Rowkey fields design

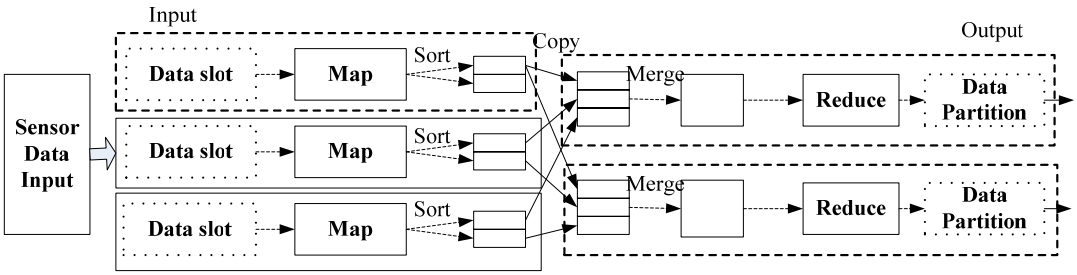


Figure 4. Map-Reduce work mechanism

*Rowkey=Hash Field+Device ID+Sensor ID+Time field.*

The design of sensing information Rowkey is shown in Figure 3.The generation method of hash field is shown as follows: H = H (M), 'H ()' is a one-way hash function, 'M' is an any length plaintext and 'h' is the fixed length hash value. The hash function is shown in Equation (1). Here, it is simplified as Equation (2) to generate the hash field. This function was presented by Professor Bernstein in the comp.lang.C news group. Also, it is one of the most efficient functions for hash code.

The latest real-time sensing information owns the maximum value usually. In order to speed up the retrieval rate of the latest sensing information, the time field generation method is shown in Equation (3). Due to the features of HBase that the data will be stored in the order of smallest to the largest according to the time code, so the latest sensing information should be arranged to the front section in HBase by using the difference between the maximum time value and the data acquisition real time as the time code.

$$h(m)=\sum_{i=0}^{|m|}(m_i << p_i)\otimes(m_i >> q_i) \quad (1)$$

$$H(M)=((h<<5)+h)+M[i] \quad (2)$$

$$T=T_{\max}-T_{realtime} \quad (3)$$

## IV. Outlier Detection Of Machine Operation Condition Based On Map-Reduce

### A. Map-Reduce Work Mechanism

Map-Reduce is a kind of programming model for big data parallel computing. The algorithm can be run in the distributed system and the efficiency can be improved by this model [20]. As Equation (4), the [k1, v1] Key-Value pairs is converted to the [k2, v2] Key-Value pairs by Map processing. Then, the Value list with the same Key is merged by Reduce operation. The whole M-R implementation process is simplified as Equation (5). In (5), the data set is *D* and the Map result is *I*. The work mechanism is shown in Figure 4. In the figure, the input sensing information is divided into some sections and mapped to Key-Value pairs by Map function. Then the Key-Value pairs are processed by the parallel Reduce function and the results are merged and outputted finally.

$$\begin{aligned} &Map: k_1, v_1 \rightarrow list(k_2, v_2); \\ &\mathrm{Re}\,duce: k_2, list(v_2) \rightarrow list(v_2) \end{aligned} \quad (4)$$

$$\mathrm{MR(D)} = \mathrm{R}(\mathrm{M(D)}) = \mathrm{list(I)} \quad (5)$$

### B. Big Data Sensing Information Outlier Detection

It is difficult to establish precise equipment fault monitoring and diagnosis model before the machine structure and work principle are known very well for the variety manufacturing equipment and complex working conditions. For example, the abnormal sensing information will be acquired if there are something wrong with the sensors or the sensing system, the fault alarm also will be generated by threshold based monitoring mechanism. But it's the false alarm and the users will get confused. Thus, the big data sensing information provides a new idea for this problem. From the point of data miner, the real-time sensing information is processed for outlier detection to monitor the mechanical equipment operation condition.

The purpose of outlier detection is to detect the abnormal behavior which is different from the expected object action. The criterion model for estimating the sensing information is established by data clustering analysis. Through the outlier detection mechanism, the abnormal condition which is deviated from the standard model will be captured. Also, the normal behavior model

is constrained by the corresponding working situation, so, the model can be used on the machine which working in different environment. If the behavior attribute values deviate from the predicted value, they will be judged as the situational constraint outlier. The outlier detection process is shown as Equation (6) and (7) which the situation attribute learning model is U and the behavior attribute model is V. It is used to judge whether the probability of detection object "o" which belongs to the behavior attribute V under the constraint of situation attribute U is greater than the threshold $\xi$.

$$S(o)=\sum_{U_j} p(o\in U_j)\sum_{V_i} p(o\in V_i)\,p(V_i\,|\,U_j) \quad (6)$$

$$S(o)>\xi \quad (7)$$

The sensing information outlier detection method under special working situational constraint consists of three steps, as shown in Figure 5.

**i) Detection model construction:** By historical data analysis and histogram statistics, the behavior model can be built under certain situation constraint. Meanwhile, the historical outlier in database also should be removed.

**ii) Real-time data stream detection:** The real-time data stream in each time window will be processed by the detection model. If the sensing information is in the interval statistics, it is regarded as the normal behavior. Otherwise, it's the outlier behavior and will be further analyzed.

**iii) Data update:** The new information will be updated to the database to improve the detection model.

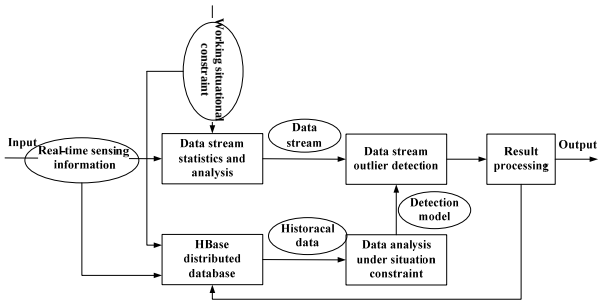


Figure 5. Sensing information outlier detection implementation steps

### C. Histogram Processing

**i) Multivariate histogram historical data analysis:** Histogram is a no-parameter model which is used frequently in statistics. The multivariate histogram is built by historical sensing data under specific situation constraint. The measurement section is decided by the maximum and minimum of the sensing information. Then, the whole section is divided into a number of small cells. The sensing information probability in each cell is M. The abscissa is defined by the measurement data and the ordinate is defined by M. So, in each cell, the statistics histogram includes the measurement data and the probability under the situational constraint. The collection $M$ of statistical probability divided by $S_k$ in situation $Q_j$ is shown as Equation (8).

$$M=\{x_i|x_i\in P(S_k|Q_j)\} \quad (8)$$

**ii) Pseudo outlier information removal:** If the statistics probability threshold is K, for the historical sensing information, the historical histogram data which is less than K would be removed. The rest is used as the detection basis.

**iii) Histogram simplified:** According to the different situational constraint granularity, the adjacent histogram which the probability is greater than a certain threshold can be merged.

**iv) Sensing information outlier detection:** The real-time data stream is acquired by the time window and inspected by the histogram model. If the sensing data belongs to one of the histogram container, it's normal. Otherwise, it is considered as outlier.

**v) Off-line analysis:** According to the working condition and other related information, the outlier should be further analyzed to decide whether there is some fault in the equipment.

**vi) Data update:** The new normal sensing information will be updated into the database. It is the basis of the new histogram model.

### D. Situational Constraint Description

According to the situation granularity, the data cube is built as the situation judgment basis. The situation set can be described as $Q=\{Q_1,Q_2,...,Q_n\}$ which means the situation attributes of various working conditions. For example, a kind of port machine situational cube under the conditions of temperature T, humidity H, wind force W and load L is shown in Figure 6. The multidimensional granularity can be combined with each other to form a large situational cube.

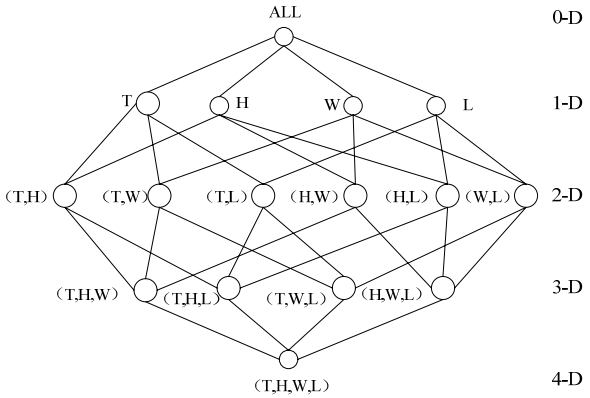


Figure 6. Multidimensional situation cube description

### E. Outlier Detection Based on Map-Reduce

By the Map-Reduce pattern, the outlier detection efficiency can be improved greatly in the distributed processing environment. First, the histogram statistics model will be built based on Map-Reduce historical sensing information analysis.

**i) Sensing information segmentation:** The sensing information within a period of time under certain situation is divided into uniform segmentations.

**ii) Map processing:** Under the constraint of situation attribute granularity K, the behavior attribute, namely the occurrence numbers in each certain histogram range are calculated. The container serial number in certain

statistical interval is considered as the Key 'K' and the sensing information occurrence number is considered as the Value 'V', like <key, 1>. So, after Map processing, the Key-Value pairs will be got as is shown in Equation (9). $D_I$ is the statistic interval, $D_i$ is the sensing data in situation $Q_{ik}$.

$$M((D_i, D_I) \mid D_i \in D(Q_{ik})) = List(k_{D_I}, 1) \quad (9)$$

**iii) Reduce processing:** The Key-Value pairs with the same key will be merged and the statistical percentage will be calculated as Equation (10).

$$M((D_i, D_I) \mid D_i \in D(Q_{ik})) = List(k_{D_I}, 1) \quad (10)$$

Through the above process, the histogram outlier detection model can be built based on historical big data sensing information. The real-time data stream outlier detection mechanism is processed according to this model. The whole Map-Reduce processing model is shown in Equation (11).

$$MP((D_i, Q_i)) = List(Q_{ik}, MP((D_i, D_I) \mid D_i \in Q_{ik})) \quad (11)$$

The Map-Reduce processing of the real-time sensing information is shown as follows.

**i) Map processing:** According to the first step, the data stream in certain time window will be divided into some segments. The container serial number in the histogram is considered as Key and the occurrence time in each container are considered as Value "1". If the real-time sensing information does not belong to any one container, the actual sensing data is considered as Key and the occurrence times of this data is considered as Value "1", like <key, 1>.

**ii) Reduce processing:** The statistics result with the same key will be merged and the result will be updated to database as new histogram model. The sensing information which doesn't fall into the certain container is considered as outlier. The outlier data set will be further processed to decide whether there is something wrong with the mechanical equipment.

## V. Experimental Verification

### A. Big Data Sensing Information Storage

The FBG sensing information analysis system is built to test the HBase storage capability for big data sensing information. Also, it is compared with the traditional relationship database MySQL. The testing environment is one Master server and three Region Servers in the LAN (Local Area Network) environment. The sampling rete of FBG wavelength demodulator is 4000Hz and the maximum capacity is 256 FBGs in the FBG sensor network. That means, besides the static information, the dynamic sensor data size is 4000 rows and each row is 1K 256 double precision floating point numbers in per second, about 4M bytes. The testing data and time consumption result is described in the following table.

TABLE III
TESTING DATA OF BIG DATA SENSING INFORMATION

| No. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Rows | 4K | 40K | 400K | 4M | 40M | 400M |
| Time(ms) | 33 | 289 | 4619 | 53216 | 408273 | 3501642 |

The insertion efficiency, namely the database writing number of rows in each unit time, is shown in Figure 7. From the Figure, 70 rows can be stored in the HBase every millisecond. That means, without considering the influence of network load in the WAN environment, HBase can meet the data storage requirements of 20 machines working together. The comparing result between HBase and MySQL is shown in Figure 8. In the MySQL database, batch processing insertion and producer-consumer memory control mechanism are also used to enhance the storage capability. From the Figure, for the single machine, both of MySQL and HBase can meet the database writing requirement, but the performance of HBase is much better than MySQL.

Meanwhile, the working situation of machine cluster in the WAN environment is also simulated. The execution efficiency result with 20 data writing threads is shown in Figure 9 and Figure 10. Due to the long time consumption for testing the 400M rows, there is no testing result about it. From Figure 9, the insertion speed of HBase is at least 10 rows/ms for each thread, which means 200 rows/ms about 200Mbytes/s for the total 20 threads. From the comparing result in Figure 10, the concurrent performance of MySQL is declined to a great extent, only about 2rows/ms. So, the HBase is much more suitable for the real-time big data sensing information storage in the WAN environment for machine cluster condition monitoring based on FBG sensor network.

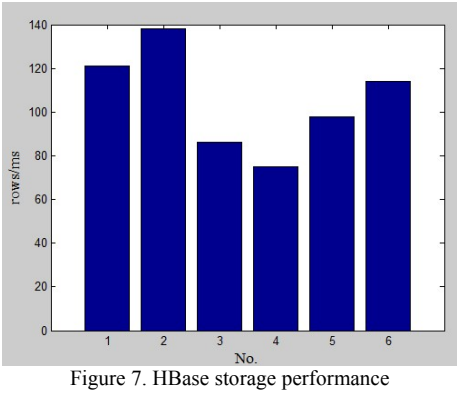

Figure 7. HBase storage performance

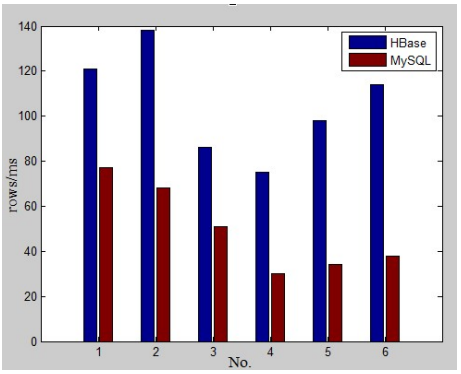

Figure 8. Comparing of Hbase and MySQL

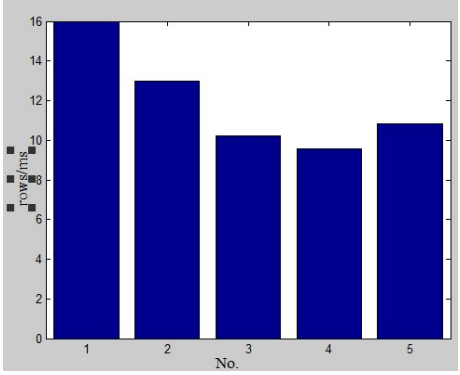

Figure 9. HBase storage performance of 20 threads

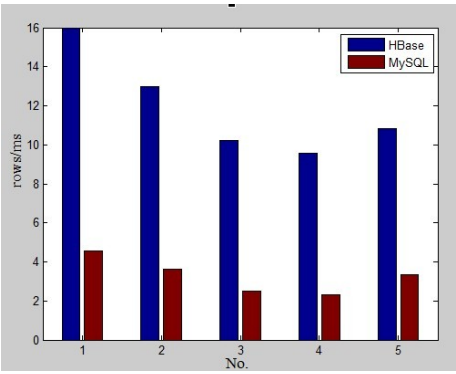

Figure 10. Storage performance comparing of 20 threads

### B. *Outlier Detection of Big Data Sensing Information*

The outlier detection method is used in the port equipment and the involved 3D models are shown in Figure 11. The real-time sensing information based on FBG sensor network is shown in Figure 12. Here, the monitoring object is gantry crane which is the common and key machine in port industry. The practical working situation during a certain period is: the load is 160 tones, the environment temperature is 20-30 Celsius degrees and the level running distance of the upper car is 0-60 miters. The histogram statistics for the fifth FBG stress-strain on the alar plate of 1/4 main beam is shown in Figure 13. The span of gantry crane is 100 meters. The measured stress-strain interval is defined as the abscissa, such as [0, 40] [40, 80] [80, 120]. It is increasing at the interval of 40 meters until it is greater than 400. The ordinate is defined as the sensing information ration between numbers of this interval and the total numbers.


Figure 11. The port machinery models

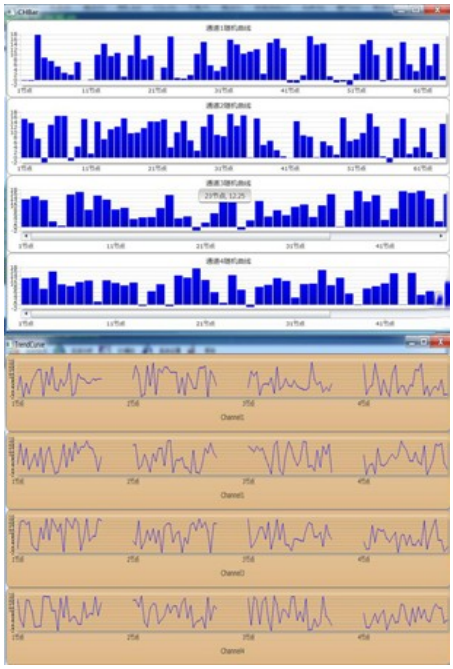
Figure 12. Sensing information based on FBG sensor network

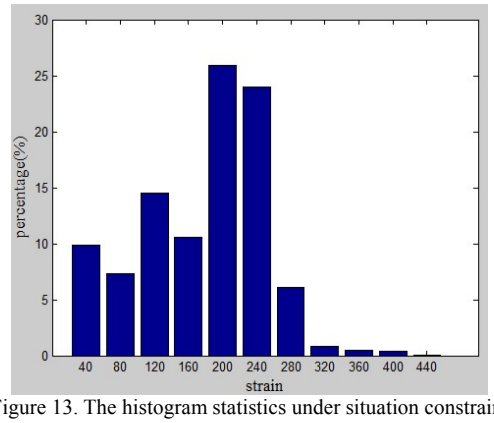

Figure 13. The histogram statistics under situation constraint

The situational constraint model is built according to the measurable working situation parameters, such as the load, environment temperature, running distance from the reference zero point in the practical working condition et al. Further, the model is simplified to describe the situation attributes of the temperature range [20, 30] Celsius degrees and the load [100, 180] tons. Under this situation, the upper car running distance from the reference zero point is considered as the dynamic situational constraint attribute. The situational constraint serial number is K1-K6, which means the car of the machine running distance is [0, 60] meters and the interval is 10 meters which are shown as Figures 14-19 respectively. In the figures, the FBG stress-strain interval (micro strain) is defined as the abscissa and the statistical probability (percentage) in the corresponding interval is defined as the ordinate.

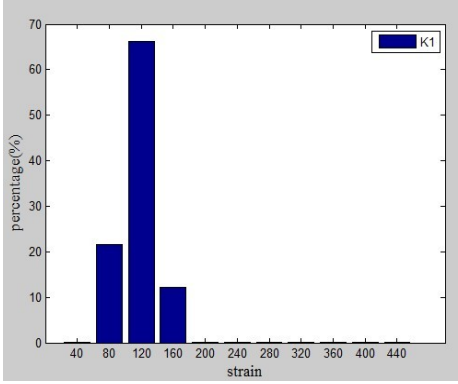


Figure 14. K1 Situation attribute 0-10 meters

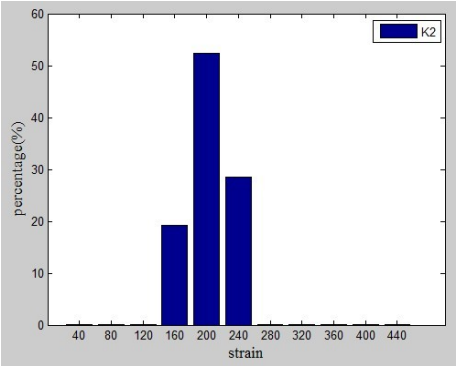


Figure 15. K2 Situation attribute10-20 meters

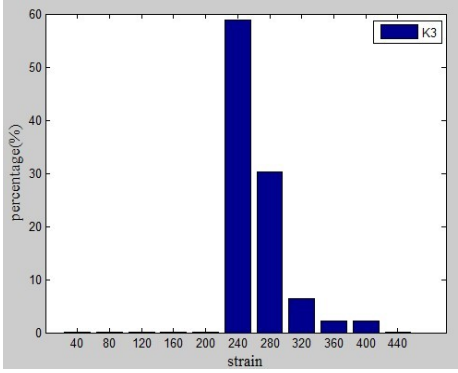


Figure 16. K3 Situation attribute20-30 meters

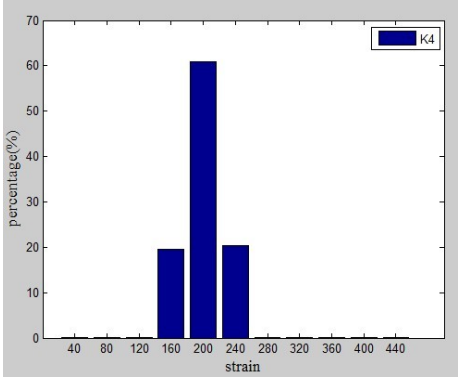


Figure 17. K4 Situation attribute30-40 meters

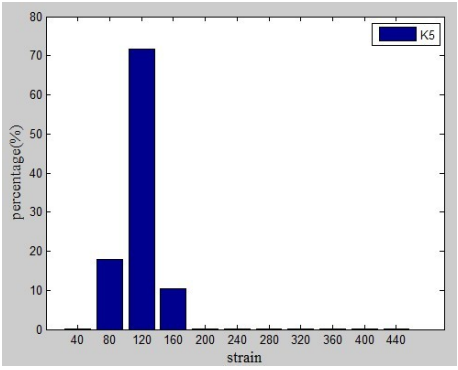


Figure 18. K5 Situation attribute40-50 meters

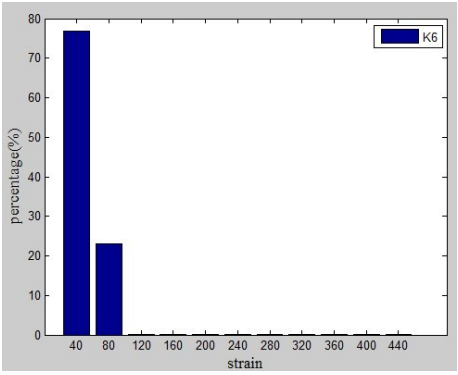


Figure 19. K6 Situation attribute50-60 meters

The real-time data stream is processed by the detection model. The testing data is 40000 points in 10 second from the fifth FBG sensor node of the FBG sensor network in the gantry crane. It is processed by Map-Reduce and the outlier detection result is shown in Figure 20. From the Figure, in addition to the normal sensing information, there is also a small amount of outlier data. The outlier data is shown in Figure 21. In the practical working condition, some abnormal sensing information is caused by the measurement system error or FBG sensor inaccuracy. So, it is not enough to prove that there is something wrong with the equipment. Through the sensing information outlier detection mechanism, the false alarm problem caused by the traditional threshold method can be avoided. If the accumulation of outlier probability is greater than the threshold after a certain time, the outlier information needs to be processed further to determine whether there are some faults in the mechanical equipment.

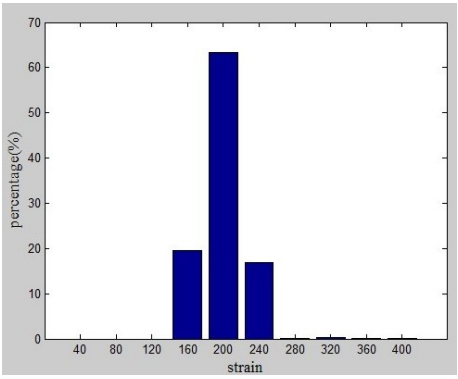


Figure 20. The histogram outlier detection result

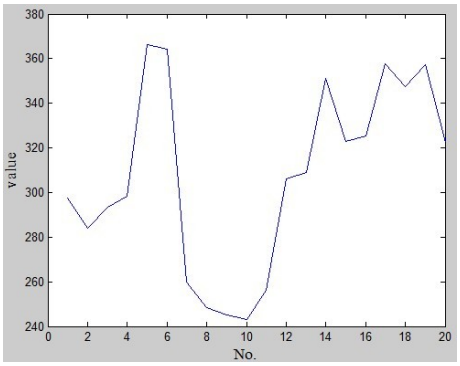


Figure 21. The outlier data

## VI. CONCLUSION

FBG sensor network provides a reliable, comprehensive and accurate way for the machine operation condition monitoring with its advantages of anti-electromagnetic interference, corrosion resistance, easy to reuse and long life. The sensing information which owns the characters of big data, high speed, real time and diversity is the valuable data resources to monitor the machine operation condition. For better use of the data resource, the No-SQL distributed database HBase is used to store the big data sensing information. The storage optimal database tables are designed. In order to avoid the hotspot accumulation problem and improve the retrieve efficiency, the table RowKey is also designed in detail. The correctness of the scheme is verified by comparing HBase and the traditional relationship database MySQL in various data size. Both the storage speed and expandability can meet the requirements of big data sensing information. Meanwhile, the real-time outlier detection with the working situational constraint method is proposed by Hadoop Map-Reduce distributed processing. Through the statistical property of outlier detection and the Map-Reduce distributed processing mechanism, the fault alarm problem cased by the system error or sensor inaccuracy can be avoided. Hence, the precision of FBG sensor network monitoring system for machine operation condition is improved by this way.

## ACKNOWLEDGMENTS

This research is supported by the National High Technology Research and Development Program of China (863 Program) (Grant No. 2012AA041203).

## REFERENCES

[1] Zude Zhou, Quan Liu, Qingsong Ai, et al., "Intelligent monitoring and diagnosis for modern mechanical equipment based on the integration of embedded technology and FBGS technology," Measurement, 2010, 44(9):1499-1511.

[2] Yumiao Wang, Jianmin. Gong, D.Y. Wang, "A Quasi-Distributed Sensing network with time-division-multiplexed fiber Bragg gratings," Photonics Technology Letters, 2011, 23(2):70-72.

[3] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics," //International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2012:19-20.

[4] G. Sagvolden, K. Pran, L. Vines, et al, " Fiber Optic System for ship hull monitoring," 15th International Conference on Optical Fiber Sensors, 2002, 1:435-438.

[5] T. Ogisu, M. Shimanuki, H. Yoneda, et al, "Damage growth monitoring for a bonding layer of the aircraft bonding structure," Smart structures and materials 2006: Industrial and Commercial Applications of Smart Structures Technologies, 2006:6171.

[6] H. Takeda, K. Sekine, M. Kume, et al, "Development of highly reliable advanced grid structure demonstrator using FBG sensors," Proc.SPIE, 2008, 6930:230-237.

[7] Zude Zhou, Desheng Jiang, Dongsheng Zhang, "Digital monitoring for heavy duty mechanical equipment based on fiber Bragg grating sensor," Science in China, 2009, 52(2):285-293.

[8] Lidong Zhai, Li Guo, Xiang Cui, et al, "research on real-time publish/subscribe system supported by data-integration," Journal of Software, 2011, 6(6):1133-1139.

[9] Dongsheng Zhang, Dan Guo, Wei Li, et al, "Study on weigh-in-motion system based on chirped fiber gratings," Advanced sensor systems and applications III, 2007, 6830: V8300-V8305.

[10] K. Shvachko, H.R. Kuang and S. Radiaet , "The Hadoop Distributed File System", IEEE 26thSymposium on Mass Storage Systems and Technologies (MSST), 2010, pp. 1-10,

[11] Fay Chang, Jeffrey Dean and Sanjay Ghemawat et al., "Bigtable: a distributed storage system for structured data", OSDI f06,Berkeley, USA, 2006, pp. 205-218.

[12] Xiongpai Qin, Huiju Wang, Xiaoyong Du et al., "Big Data Analysis-competition and symbiosis of RDBMS and Mapreduce," Journal of Software, 2012, 23(1):32-45.

[13] Rohit P. Kondekar, "Implementation and analysis of iterative MapReduce based heuristic algorithm for solving N-Puzzle," Journal of Computers, 2014, 9(2):420-424.

[14] Jie Wang, Qinghan. Dai, Yu Zeng, et al, "Parallel frequent pattern growth algorithm optimization in cloud manufacturing environment," Computer Integrated Manufacturing Systems, 2012, 18(9):2124-2129.

[15] Haluk Demirkan, Dursun Delen, "Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud," Decision Support System, 2013, 55(1):412-421.

[16] Jiehui Ju, "Survey on cloud storage," Journal of Computers, 2011, 6(8):1764-1771.

[17] Xin Chen, XuBin He, He Guo, et al, "Design and evaluation of an online anomaly detector for distributed storage systems," Journal of Software, 2011, 6(12): 2379-2390.

[18] QingKui Chen, Lizhen Zhou, "HBase-based storage system for large- scale data in wireless sensor network," Journal of Computer Applications, 2012, 32(7):1920-1923.

[19] Yang Yang, Xiang Long and Bo Jiang, "K-Means method for grouping in hybrid MapReduce cluster," Journal of Computers, 2013, 8(10):2648-2655.

[20] Jeff Dean, Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, 2008, 51:107-113.