Research of Construction and Application of Cloud Storage in the Environment of Industry 4.0

Kaifeng Geng(✉) and Li Liu

Software School, Computer Network Center, Nanyang Institute of Technology, Nanyang, 473000 Henan, China
gkf8605@126.com

Abstract. With the geometric growth of data generated from complex system test, experiment and condition monitoring, big data has become the hotspot of Industry 4.0 era. How to meet market demand with efficiency, comprehensiveness and low cost through collection and analysis of data is a problem to be solved. We put forward industrial cloud storage model based on Hadoop on the basis of relevant researches and theories of Industry 4.0, big data Hadoop and so on, and then implement and evaluate each module. It performs well in reliability and expandability through test, providing challenges of big data storage in Industry 4.0 era with effective solution.

Keywords: Industry 4.0 · Hadoop · Cloud storage · Big data

1 Introduction

With the deep merging of informatization and industrialization, information technology has permeated into all aspects of the industrial chain in the field of industry. Bar codes, two-dimensional codes, RFID, Sensors, industrial automation systems, Internet of things and other technologies are widely applied in industry. Especially with the wide application of Internet and Internet of things, industry has entered the new developing stage of internet industry and enterprises have more rich data. In order to make use of data, enterprises need the capacity to support multiple types of information, infrastructures for the store of big data and the capacity of fast and accurate analysis of information after storage.

There is a large amount of data to be collected and processed produced by industrial equipments, the scale of the big data in the data set is from dozens of TB to lots of PB and these data are mostly unstructured in the data type. Moreover, operation of the production line with high speed requires more real-time data, so using traditional data storage schemes to complete the collection and processing of data within an acceptable period of time will be very difficult. With the increasing number of advanced devices and equipments, a large number of operating data goes online, which heralds the coming of Industry 4.0 [1].

To process these high-dimensional data reflecting product information and equipment information, there need prominent computing and storage capacity. Though computers' computing capacity is in improvement, it is still far from the requirement of

J. Wan et al. (Eds.): Industrial IoT 2016, LNICST 173, pp. 104–113, 2016.
DOI: 10.1007/978-3-319-44350-8_11

processing such enormous data. In addition, in a typical big data storage system, the concurrency of the reading and writing of a variety of data is high, so the restriction of the storage and processing capacity of the central server will inevitably lead to a single point of failure, which blocks the development of the fuse of informatization and industrialization as well as the upgrade and transformation of the whole industry. The birth of cloud computing technology brings dawn to problems such as data processing, storage and so on.

2 Related Concepts of Industry 4.0, Big Data and Hadoop

2.1 Industry 4.0

The research project, "Industry 4.0", was firstly put forward in German and has risen as a national strategy. "Industry 4.0" aims at creating an individualized and digital production model of products and services with high flexibility [2]. The core meaning of Industry 4.0 times is information technology, we can realize the effective and rapid docking of two sides based on the demand and supply of data analysis, can reduce the cost of spending and achieve the directional and customized production through the Internet [3].

2.2 Big Data

Big data is a data set whose scale is much larger than the capable scale of traditional database software tools in acquirement, storage, management and analysis of data. Nowadays, sensors, GPS systems, Internet of things and social networks are creating a new flow of data and big data has become a symbol or a characteristic of Internet in current developing stage. Data itself contains value, so it must flow seamlessly and securely to people when making decisions or action and offer basis for decisions at any time. Under the backdrop of technical innovation represented by the cloud computing, these data difficult to be collected and used in the past are getting easier. And with the constant innovation of all industries, big data will create more value for human beings step by step [4].

2.3 Hadoop Technology

Hadoop [5] Framework is an open source project of Apache Foundation, which is a software framework capable of distributed processing of a large amount of data and offers easy programming interface. It is a platform of cloud computing where programmers can easily develop and handle a huge amount of data. MapReduce and HDFS are the core design of the framework. Hadoop cluster is a specific type of cluster designed specifically for the storage and analysis of massive unstructured data. In essence, it is a kind of computing cluster, which will allocate the data analysis work into multiple cluster nodes, so that the data can be processed in parallel.

MapReduce is a programming model for processing and generating big data sets. The working process of MapReduce is divided into two stages: map stage and reduce stage, among which, "map" is to divide a task into multiple tasks, while "reduce" is to summarize the processing result of multiple divided tasks and then get the final result [6]. The user-defined Map function is used to process a data set based on key/value pair and output the intermediate data set based on key/value, while Reduce function is used to merge all intermediate value with the same intermediate key.

Using the ideas of functional programming language Lisp, MapReduce defines the following two abstract programming interfaces, Map and Reduce, which will be implementated by programming [7]:

map: (k1; v1) → [(k2; v2)]

Input: data presented by key/value pair (k1, v1)

Processing: Document data record (such as a line in a text file, or data in the table rows) will be transfered into the map function in the form of "key/value pair; The map function will deal with these key/value pairs and output intermediate results in the form of another key/value pair [(k2, v2)].

Output: intermediate data presented by [(k2, v2)]

reduce: (k2; [v2]) → [(k3; v3)]

Input: the key/value pair [(k2, v2)] will be merging processed which is output by map, different value under the same primary key will be combinated to a list (v2), so the input of reduce is (k2; [v2]);

Processing: the incoming intermediate results will be further processed or listed in some sort, and generate the final output (k3; v3).

Output: the final output [(k3; v3)]

HDFS is a distributed file system designed for storing large files in the mode of streaming data access which can not only customize the block size of store files (the default is 64M), but also customize the number of copies and the security level with high fault-tolerance and security [8].

3 The Influence of the Cloud Computing to Industry 4.0

With the cloud storage getting cheaper and the cloud processing getting more powerful, the cloud becomes the best choice for the storage and analysis of data collected by enterprises. The innate characteristics of the cloud computing, that is, cheap storage and good performance computing make it better than other traditional technologies serving the industry.

3.1 Low Requirements of the Cloud Computing for the Configuration of the Client and the Server, and Low Cost

Hadoop cluster is relatively cheap, there are two main reasons. Its required software is open source, so it can reduce the cost. In fact, you can free download Apache Hadoop distribution. At the same time, the Hadoop cluster controls the costs by supporting commercial hardware. So you don't have to buy the server hardware to build a powerful

Hadoop cluster [9]. One of the core concepts of the cloud computing is to reduce the processing load of the user terminal through constant improvement of the processing capacity of the "cloud". Clients only need to input and output, and all other functions such as computing, storage and processing are managed by the "cloud". Users only need to order relevant services of the "cloud" according to their own needs. In addition, the storage equipments of the "cloud" can be cheap PCs, even old computers. Compared with the single professional storage equipments with large volume, the "cloud" has larger storage capacity and lower storage cost, and can realize dynamic upgrade and extensions according to the demand [10].

3.2 Cloud Computing Can Offer Massive Computing and Storage Capacity, Has a Great Deal of Extensibility

Like any other type of data, an important problem faced by big data analysis is also the increasing amount of data. And the biggest advantage of big data is that it can realize real-time or near real-time analysis and process. Hadoop cluster parallel processing capabilities can significantly improve the analysis speed, but with the increase of the amount of data to analysis, the cluster's capacity is likely to be affected. But thankfully, by adding additional cluster nodes can effectively extend the cluster.

The cloud computing can gather resources such as memories, hard drives and CPUs of all nodes into a giant virtual cooperative working pool of resources, providing storage and computing services for the outside together. With the increase of nodes, the capacity of storage and computing can unlimitedly increase.

3.3 Cloud Computing Can Provide Storage with High Reliability and Security

Data collected in all parts of the industry such as production and sales will be stored multiple service nodes of the cloud with multiple copies. Data stored in the cloud will not be affected even if accidentally deleted. And there is no need of fearing virus invasion and the data loss caused by hardware damage.

4 Industry Cloud Storage Model Construction Based on Hadoop Technology

Hadoop is a tool which can realize data storage expansion by using standard hardware and can distribute data among many low-cost computers. After data distribution follows the difficulties of data location and handling which can be solved by Map Reduce. Map Reduce provides a framework, and data in a cluster are parallel processed among many nodes. It is allowed to map the processing to many location data and cut similar data elements to a single result [11].

Aimed at challenges faced by the construction of big data storage system at the present, on the basis of the advantages of the Hadoop technology, we put forward the industrial cloud storage model based on the Hadoop platform, which can effectively solve problems such as processing capacity limitation, storage capacity limitation and

single point failure. In the cloud computing technology, data collected by sensing equipments, bar codes, two-dimensional codes, RFID and so on are provided in the form of saas (software-as-a-service) in the cloud, terminals are responsible for collecting data and sending data to cloud applications, and present massive intuitive data as well as statistic results in all parts of the industrial production, These data are generally widely distributed and unstructured, but Hadoop is very suitable for this kind of data because the work principle of Hadoop is that the data is split into slices, each "slice" will be analyzed by assigning to a specific cluster nodes. The data distribution is not have to uniform, for each shard is processed separately on each independence cluster nodes.

which can not only provide the independency of the platform, but also reduce the possibility of problems caused by loading central server and the single point fault. Meanwhile, using MapReduce model can also avoid single fault in the calculation of lots of high dimensional data. This framework showed in Fig. 1 is divided into the front end and the back end from top to bottom.

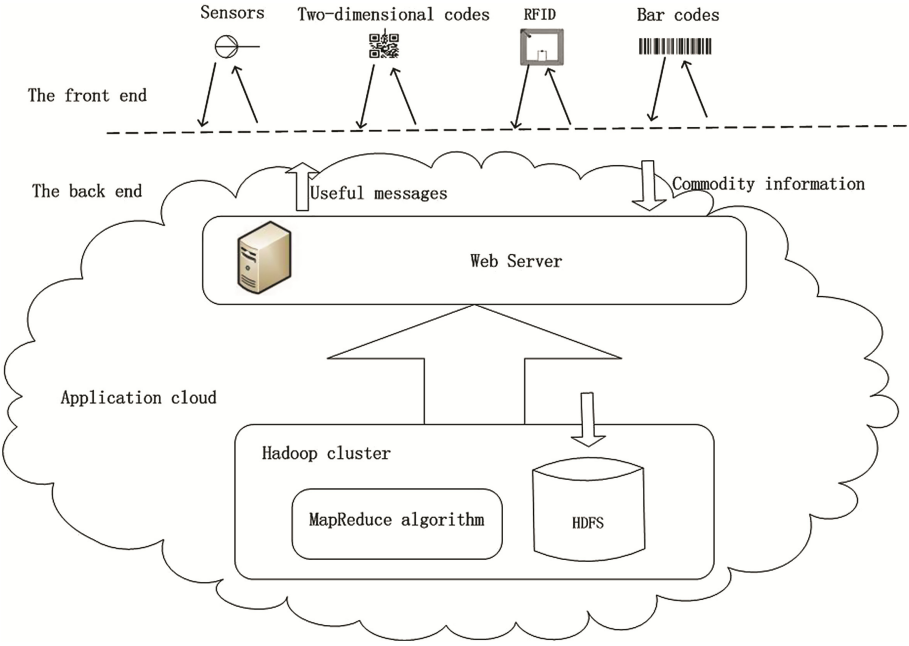


Fig. 1. Industry cloud storage model based on Hadoop technology

4.1 The Front End

The front end consists of the sensor, the bar code, the QR code, RFID and other mobile devices, which is used for collecting data and presenting optimized information. Signals collected from all parts of the industrial production are uploaded to the back end through this application for the further processing, and are presented in the appropriate form.

This information can provides basis for decision making, bringing deep knowledge of the industry and competitive advantages.

4.2 The Back End

The back end is the core of this model, mainly including three modules such as the WEB server, MapReduce algorithm and HDFS cloud storage. WEB server is responsible for the communication between the Hadoop cluster and the WEB interface as well as receiving sensor signals from the front end and displaying all kinds of optimized data. A Hadoop cluster with multiple nodes is responsible for solving parallel processing tasks. Each node has a copy of MapReduce java program and MapReduce is responsible for the large-scale parallel computing of industrial big data according to a certain algorithm. As a distributed file system, HDFS is used to store big data generated in all parts and provides high throughput for accessing application data, ensuring seamless transfer of data among servers and improving the reliability of the whole system.

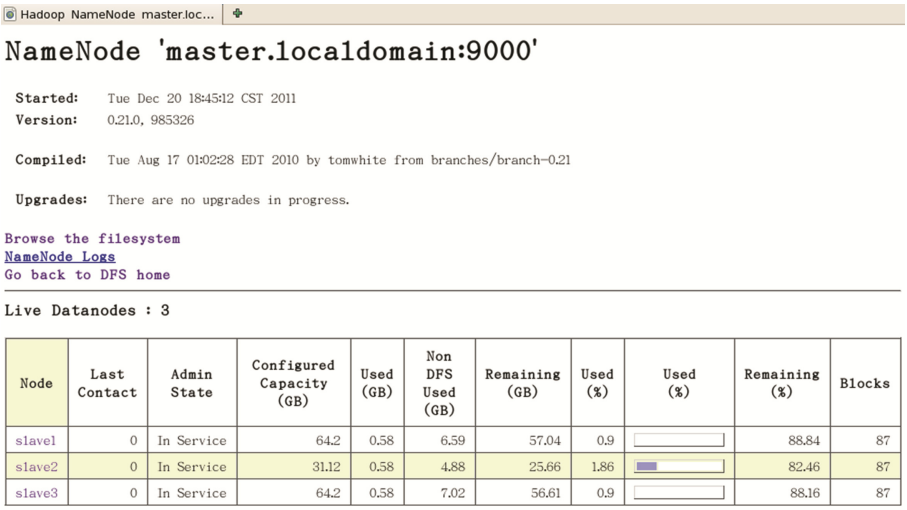
5 System Implementation

Due to the complex structure and powerful function of a whole industrial cloud storage system, we only implement the cloud environment, MapReduce and HDFS cloud storage. In order to make users' operation convenient, a WEB interface is needed to be developed to upload collected data and display processing results. There need two java scripts for the interaction between the front-end and the back-end: one is for uploading data files of sensing signals to HDFS; the other is for accessing the output files of Reducer, and getting the maximum key/value pair through comparing there key/value pairs whose value can reflect the real condition of the products. Key codes of the uploaded Java script are as follows:

```
String src = "e://degree.txt";
String dst = "hdfs://210.42.241.66:9000/user/Hadoop/degree.txt";
InputStream in = new BufferedInputStream(new FileInputStream(src));
FileSystem fs = FileSystem.get(URI.create(dst), new Configuration(););
OutputStream out = fs.create(new Path(dst), new Progressable() { });
IOUtils.copyBytes(in, out, 4096, true)
```

5.1 Cloud Environment

Cloud environment plays a vital role in the system architecture, providing the whole system with massive storage and computing capacity. In this experiment, we use Hadoop cluster built by 3 PCs as the cloud, among which one serves as the master node, the other three serve as slave nodes. Hadoop version is 0.21.0, JDK version is 1.6 and the operating system is Red Hat Linux5.4. Details of the installation and configuration process need not be repeated here, and nodes' state information after load balance is show in Fig. 2.


Hadoop NameNode master.loc...

NameNode 'master.localdomain:9000'

Started: Tue Dec 20 18:45:12 CST 2011
Version: 0.21.0, 985326

Compiled: Tue Aug 17 01:02:28 EDT 2010 by tomwhite from branches/branch-0.21

Upgrades: There are no upgrades in progress.

Browse the filesystem
NameNode Logs
Go back to DFS home

Live Datanodes : 3

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)	Remaining (%)	Blocks
slave1	0	In Service	64.2	0.58	6.59	57.04	0.9		88.84	87
slave2	0	In Service	31.12	0.58	4.88	25.66	1.86		82.46	87
slave3	0	In Service	64.2	0.58	7.02	56.61	0.9		88.16	87

Fig. 2. State information of Hadoop cluster nodes

5.2 MapReduce Implementation

The difficulty of industrial data management is not limited to the quantity of information. As data have different formats and sources, there exist problems of diversity and complexity of the data. There often exists processes information "island" which must be merged, stored and analyzed to obtain meaningful values.

In the MapReduce model, each master has a JobTracker (JobTracker is a master service. After the startup of software, JobTracker receives job, and is responsible for scheduling each sub-task of the job to run on TaskTracker, monitoring them and restarting the task if discovering failed task. In general, JobTracker should be deployed in a single machine). Each salve has a TaskTracker (TaskTracker is a slaver service running at multiple nodes. TaskTracker initiatively communicates with JobTracker, receives tasks and is responsible for the direct implementation of each task) [12].

The implementation of MapReduce includes the map stage and the reduce stage: the map stage extracts the characteristics of each line's data in the matrix (the matrix represents the collected data) and gets the evaluation results. Among them, each line in the matrix represents the signal s from an electrode. The inputting of Mapper is a key/value pair. Key is a matrix file and the inputting of Mapper is to generate a new key/value pair, for example, (Good, 1), (Neutral, 2), or (Bad, 4). The reduce stage processes the unique key from the map and then obtains three key/value pairs which represent the amount of Good, Neutral and Bad keys among the totality. For example, key/value pairs, (Good, 2), (Neutral, 2), (Bad, 1), (Bad, 4) and (Good, 1) generate new key/value pairs (Good, 3), (Neutral, 2) and (Bad, 5) after Reducer's processing.

5.3 Implementation of HDFS Cloud Storage

In the framework of HDFS cloud storage, master has a NameNode and slave has a DataNode. NameNode is a central server, responsible for managing the namespace of the file system and the access of the client to files. DataNote is generally responsible for managing the storage of its nodes. HDFS exposes the name space of the file system and users can store data in it in the form of file. Seeing from the inside, a file is actually split into one or more data blocks which are stored in a set of DataNote. NameNode implements the operation of the file system's name space, as Fig. 3 shows.

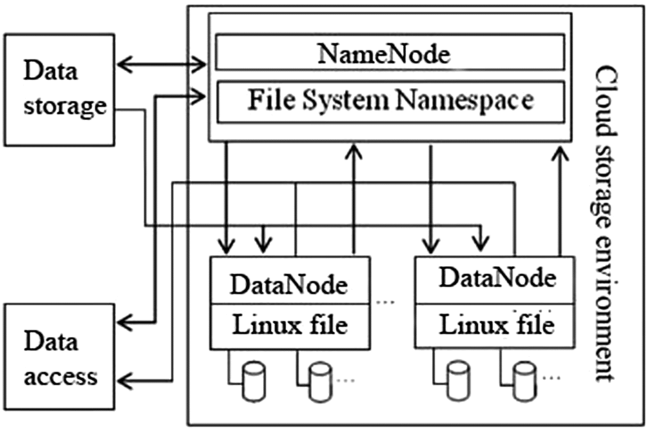


Fig. 3. HDFS cloud storage architecture diagram

6 Evaluation and Testing

We use a large data set to evaluate the system and design two different tests for the objective test performance: load testing and performance testing.

6.1 Specific Environment Configuration

First set of the test:

The configuration of a single computer is as follows:

Platform: Red Hat Enterprise Linux 5 update 4; Processor: 32 Bits, 3.0 GHz Intel Dual-core; memory: 4 GB; drive: 500G. Network card: 100M full duplex network connection.

Second set of the test:

The configuration of the four-node Dadoop computer cluster is as follows (configuration is much lower than the first set of the test):

Platform: Red Hat Enterprise Linux 5 update 4; Processor: 32 Bits, 1.8 GHz Intel Dual-core; memory: 2 GB; Drive: 500G; Network card: 100M full duplex network connection.

6.2 Load Testing

Load testing is to test the scalability and reliability of the system. In order to simulate the loading environment of the system, we use Neoload to record user context and simulate 500 virtual users. Ganglia monitor is used to monitor the average utilization of the node CPU. As Fig. 4 shows, after the test, when there are 500 concurrent users, the utilization of the system CPU is less than 5 %, which shows this system is very energy-efficient.

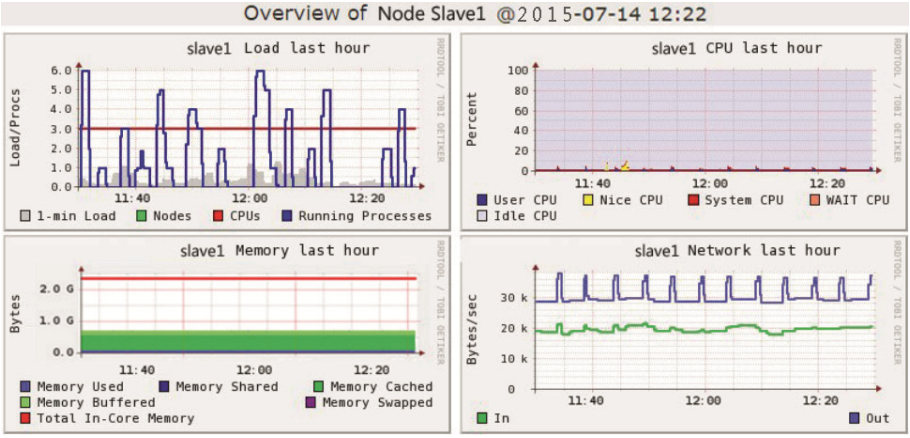


Fig. 4. Ganglia monitors the status of Slave1 node

6.3 Performance Test

Comparison method is used to evaluate and test the system's overall performance. The first test: write a normal Java application as a skill base for reference in the test and evaluation, which is running on a local machine to calculate the baseline performance; the second test: the same algorithm runs on the Hadoop cluster with 4 nodes to implement Java MapReduce.

When running MapReduce algorithm program, we choose two different users with different amount to concurrently access the system. The first test: simulate 5 virtual users, after the test, estimate the baseline performance of the first test is 0.8 s and the baseline performance of the second test is 5 s. The experiment result shows: Local Java applications are much faster than applications running on the Hadoop cluster. The second test: simulate 500 virtual users, after the test, estimate the baseline performance of the first test is 61 s and the baseline performance of the second test is 22 s. The experiment result shows: Java applications running on the Hadoop cluster are much faster than local applications.

Through above two experiments we can conclude that: in the case of small-scale users, local ordinary Java program responses obviously faster, and compared with traditional Java programs, using HDFS and MapReduce has no advantage. However, in the case of large-scale users, because MapReduce is much potential in processing large files (TB level), it distributes the large-scale operation of the data set to each node on the

network for implementation. In general, data amount generated in the environment of Industry 4.0 is relatively large, therefore the proposed system has reasonable structure and good performance [13].

7 Conclusion

Faced with the increasing large amount of data under the environment of Industry 4.0, only using appropriate tools for prediction and analysis can the large amount of disorderly data be processed into usable information. The proposed industrial cloud storage model based on Hadoop technology has good performance through the test which can provide the data analysis and processing of the industry with efficient help and explain some certain uncertainty as well as predict problems that may occur so as to help companies to make more "wise" decisions.

References

1. Wang, X.: Industry 4.0: intelligent industry. Technol. Internet Things (12), 1–3 (2013). (in Chinese)
2. Wahlster, W.: From industry 1.0 to industry 4.0: towards the 4th industrial revolution. Forum Business meets Research (2012). (in Chinese)
3. Yen, C.T., Liu, Y.C., Lin, C.C., et al.: Advanced manufacturing solution to industry 4.0 trend through sensing network and cloud computing technologies. In: 2014 IEEE International Conference on Automation Science and Engineering (CASE), pp. 1150–1152. IEEE (2014). (in Chinese)
4. Peng, W.: The Key Technology and Application of Cloud Computing, pp. 73–75. The People's Posts and Telecommunications Publishing House, Beijing (2010). (in Chinese)
5. Liu, K., Li, A.: Research and implementation of cloud storage based on Hadoop. Micro Comput. Inf. **27**(7) (2011). (in Chinese)
6. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008). (in Chinese)
7. Lin, J., Dyer, C.: Data-intensive text processing with MapReduce. Synth. Lect. Hum. Lang. Technol. **3**(1), 1–177 (2010). (in Chinese)
8. Gao, H., Zhai, Y.: Research of mobile learning model based on hadoop. China Audiovisual Education **2011**(288). (in Chinese)
9. Borthakur, D.: The hadoop distributed file system: architecture and design. Hadoop Proj. Website **11**(2007), 21 (2007). (in Chinese)
10. Cheng, X.: Demands, environment and service of industrial big data under the structure of industrial 4.0. J. Chifeng Uni. (Nat. Sci. Ed.) **2015**(4), 14–15. (in Chinese)
11. Shafer, J., Rixner, S., Cox, A.L.: The hadoop distributed filesystem: balancing portability and performance. In: 2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS), pp. 122–133. IEEE (2010). (in Chinese)
12. Shvachko, K., Kuang, H., Radia, S., et al.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. IEEE (2010). (in Chinese)
13. Lee, J., Kao, H.A., Yang, S.: Service innovation and smart analytics for Industry 4.0 and big data environment. Procedia CIRP **16**, 3–8 (2014). (in Chinese)