International Journal of Grid and Distributed Computing
Vol. 10, No. 1 (2017), pp.111-122
http://dx.doi.org/10.14257/ijgdc.2017.10.1.11

Big Data Compressed Storage Algorithm in Rock Burst Experiment

Yu Zhang[1, 3], Yan-ping Bai*[2], Zhao-yong Lv, Yongzhen Li, Zong-lei Mu

[1] *School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China*
[2] *College of management, Capital Normal University, Beijing, China*
[3] *State Key Laboratory for GeoMechanics and Deep Underground Engineering, China*
yuzhang@bucea.edu.cn, yibaibaby@126.com, lvzhaoyong@bucea.edu.cn, liyongzhen@bucea.edu.cn, muzonglei@bucea.edu.cn

Abstract

State Key Laboratory for GeoMechanics and Deep Underground Engineering (GDLab) has accumulated more than 500 TB data of rock burst experiment. But so far the amount of analyzed data is less than 5%. Data storage dilemma is restricting the study mechanism of rock burst. In this paper, we applied big data technology into analyze of rock burst, and makes deep analysis about characteristic of rock burst data. Basing on this, a big data based data storage systems (BDSS) for rock burst experiment was proposed. BDSS based on Hadoop for rock burst with online data loading and rapid retrieval of data. In Storage node machine cluster in BDSS, Big Data Compressed Storage Algorithm was proposed. The algorithm can provide average compressed ratio about 2.91%, which is as good as WinRAR. And at the same time, considered time for compress data, Big Data Compressed Storage Algorithm is much better than WinRAR. In one word, Experimental analysis shows that the algorithm have excellent performance in rock burst and solve the Data storage dilemma. Research work of this paper laid some foundation of rock burst.

Keywords*: Big Data, Compressed Storage Algorithm, Rock Burst.*

1. Introduction

Rock burst is a kind of geological disaster in underground engineering with excavation, which has become one of worldwide underground engineering problems[1]. In 2006, State Key Laboratory for GeoMechanics and Deep Underground Engineering, GDLab for short, successfully reproduced process of rock burst indoors. Since then, research on the mechanism of rock burst rose to a new level. A lot of research works on the mechanism of rock burst has been done by GDLab[2-6], and make a series of valuable research results[7-11]. But there are some dilemmas during the research progress. The most important one of the dilemmas is data storage dilemma. Data storage dilemma is because of a large number of experimental data is produced in rock burst research. Furthermore, this data is inevitable and determined by the characteristic of rock burst.

For example, one rock burst experiment numbered as "yqsii6#", it generated 33217 txt files in an hour. Hard disk needs 12GB storage space to save all these 33217 files. Along with the further research work, the data obtained also growth with geometric speed. The GDLab has accumulated more than 500 TB of data, but in so far, its analysis ratio is no more than 5 percent.

ISSN: 2005-4262 IJGDC

Facing huge data, only effectively manage these data and analysis these data using modern means, which can establish theoretical basis of rock burst mechanism and forecast. This is the focus of this paper.

2. Related Work

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics [12]. Sagiroglu, S. etc al. presents an overview of big data's content, scope, samples, methods, advantages and challenges in this paper. E Udoh and CH Hsu represent recent progresses in the field, including works on virtualization, big data intelligence, resource management, services computing architectures and modeling, as well as mobile cloud and applications [13]. S. Krishnaveni etc al. had proposed four phases of big data namely data generation, data acquisition, data storage and data analytics [14].

Hadoop has become synonymous with big data. Hadoop derived from Google File System (GFS) [15] and MapReduce (GMR)[16]. The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications [17]. The biggest characteristic of Hadoop is that once it starts running, it wills automatically and rapidly analysis data [18]. Hadoop can analyze datasets of any size [19-22].

Dremel[23] is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google[24].

Catrein etc. had provided a platform for efficiently and flexibly aggregating, storing, and processing big data and make conclusion of sensor networks will automatically collect such data [25]. We also use similar method to collect data as illustrated in section 3.2 of this paper.

3. Our Approach

3.1. Rock Burst Big Data

A series of research results have been achieved by State Key Laboratory for GeoMechanics and Deep Understand Engineering (GDLab for short as below). [26-28]. A great amount of research data been generated[29-30]. For example, one rock burst experiment numbered as “O-1#” generated 29438 txt files. Hard disk needs 1.5GB storage space to save all these 29438 files. Along with the further research work, the data obtained also growth with geometric speed. How to scientifically managing and using these data become the biggest problem which GDLab facing now.

3.2 Collection of Rock Burst Big Data

3.2.1 True Triaxial Host System

True triaxial host system is designed for rock burst. The system is illustrated in figure 1.

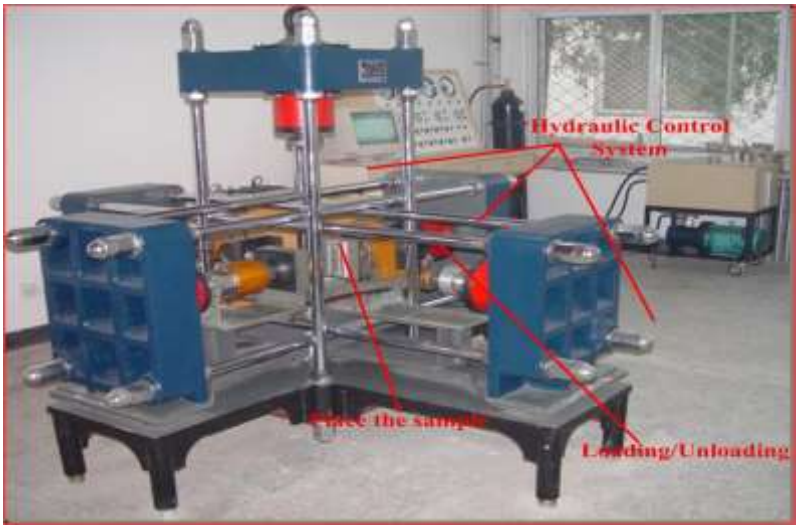


Figure 1. True Triaxial Host System

3.2.2 Data Collection System

System collect rock burst data using two types of sensors from Pengxiang and PAC. Some key parameters of the sensors are shown in figure 3.

Figure 2. Sensors and their Parameters

3.2.3 Massive-data Dilemma

Take one example of rock burst experiment which is numbered "yqsii6#".

i): "yqsii6#" generated 33217 files each hour in average. There needs 12GB storage space to save all these 33217 files.

ii): Each text will be painted into four pictures in JPG format. Each picture needs 300KB storage space to save them on hard disk.

The total disk space can be calculated as followed:

i): One hour disk space can be calculated by formula 1.

$$12GB+4*300KB*33217=51.86GB \quad (1)$$

ii): If the total disk space is 2t~3t, it will be full within 39hours~48hours.

GDLab has accumulated more than 500 TB data of rock burst experiment. But so far the amount of analysed data is less than 5%. Data storage dilemma is restricting the study mechanism of rock burst.

Here we take three typical rock burst experiments as example, studying the characteristics of rock burst experimental data. In the following part of this paper, we will use the same experiments too. Data storage details of three typical rock burst experiments are described in table 1.

Table 1. Data Storage Details Of Three Typical Rock Burst Experiments

Experiment No.	Number of TXT files	Occupied Disk Space (GB)
O-1#	29438	1.50
G-1#	41645	2.11
GO-1#	71351	3.66

From table1 we can see that each experiment has how many txt files and how large the disk space it occupied.

4. Big Data Compressed Storage Algorithm in Rock Burst Experiment

As we mentioned above, also different experiment has different number of txt files, each txt file has accurate 4096 values. Under this principle, each experiment may have hundred millions values as showed in table 2.

Table 2. Data Storage Details of Three Typical Rock Burst Experiments with Values

Experiment No.	Number of TXT files	Number of values	Occupied Disk Space (GB)
O-1#	29438	120578048	1.50
G-1#	41645	170577920	2.11
GO-1#	71351	292253696	3.66

It's impossible to analyse these data in the past. We proposed a big data based algorithm solving this dilemma as below.

4.1 System Structure

We have put forward a big data based data storage systems for rock burst experiment, BDSS for short, which based on Hadoop for rock burst with online data loading and rapid retrieval of data. The structure of BDSS is showed in figure 3.

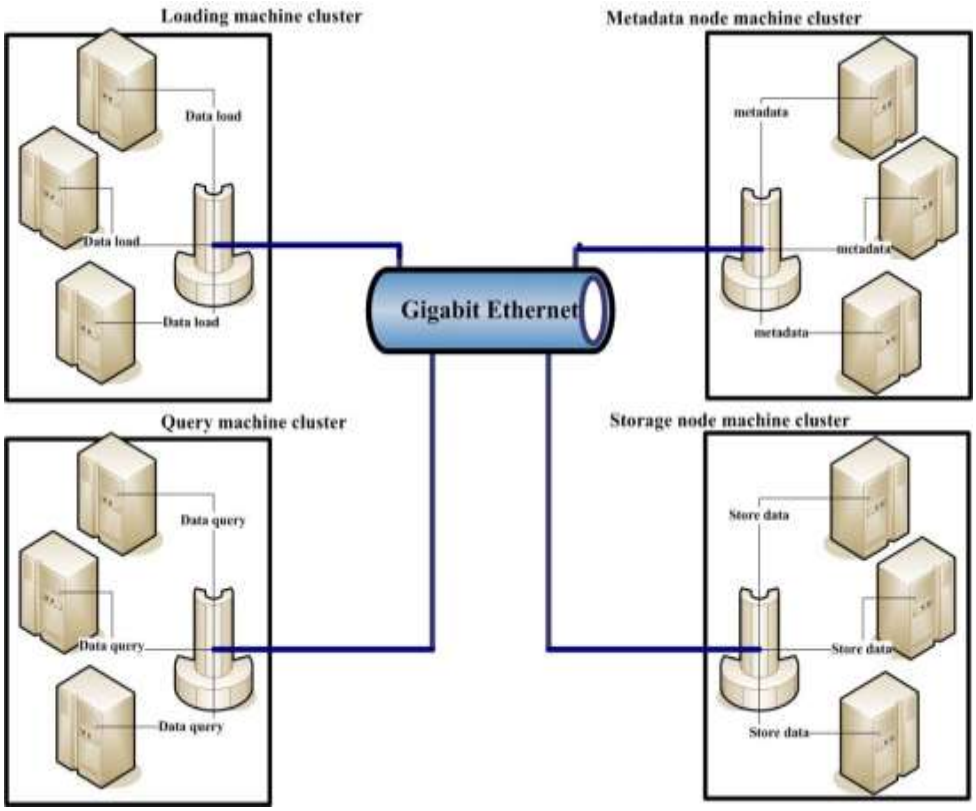


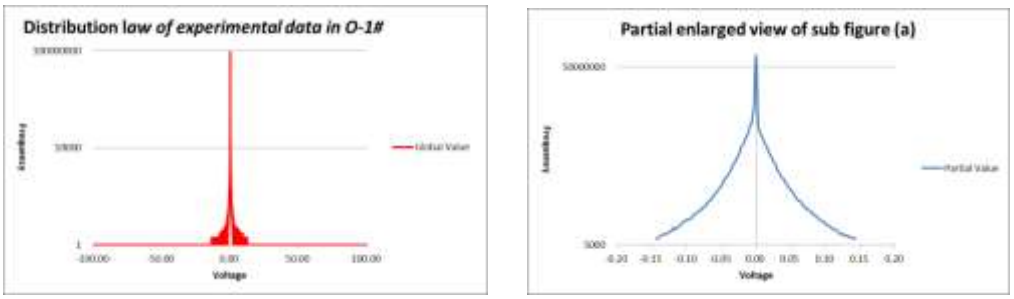
Figure 3. System Structure of BDSS

As illustrated in figure 4, Storage node machine cluster is at the lower right corner. The storage node machine cluster provides persistent storage capacity, long-term preservation of historical data. System storage data sources as small block usually take once or several times from the loading machine cluster as the data block unit.

Under such mechanism, we got some law of experiment data as illustrated in section 4.2.

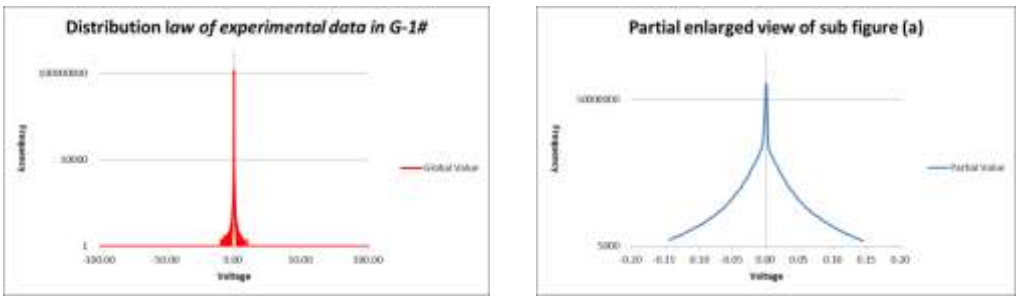
4.2 Law of Experimental Data

Firstly, we separately got the distribution law of experimental data in three experiments as showed in figure 4, 5 and 6. Each figure has two sub figures. Sub figure (a) is the distribution law of experimental data, and sub figure (b) is the partial enlarged view of sub figure (a).



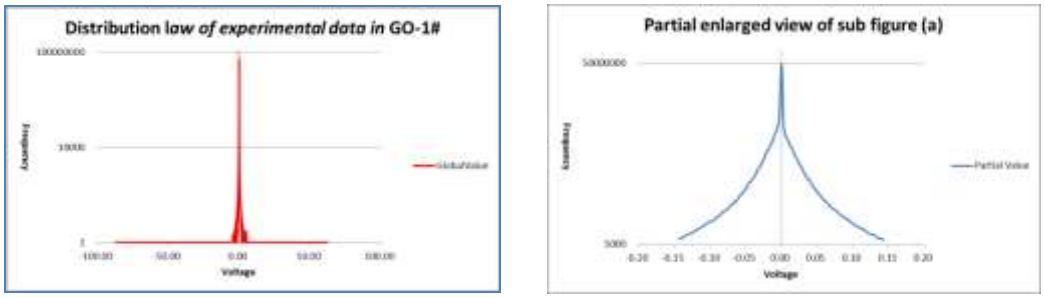
(a) (b)

Figure 4. Distribution Law of Experimental Data in O-1#



(a) (b)

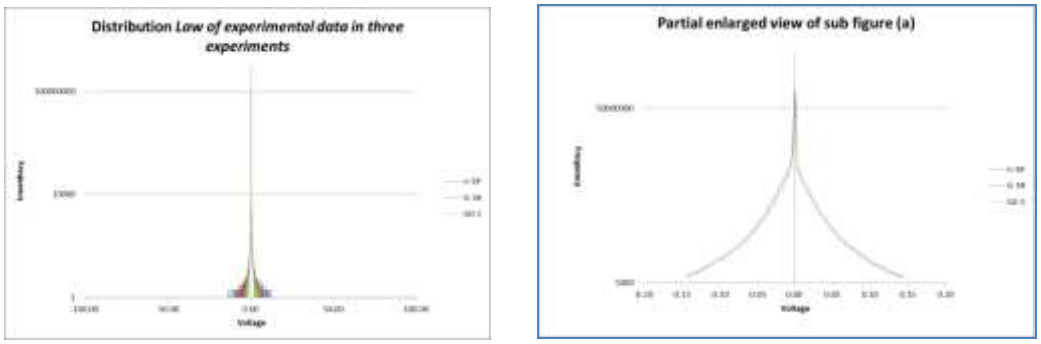
Figure 5. Distribution Law of Experimental Data in G-1#



(a) (b)

Figure 6. Distribution Law of Experimental Data in GO-1#

Then we combine figure 4, 5 and 6 in figure7 in order to make the comparison of the three experiments is more obvious.



(a) (b)

Figure 7. Distribution Law of Experimental Data in Three Experiments

From all above we found that the distribution law of experimental data is a regular distribution. Then we made some further research on it, and proposed the big data compressed storage algorithm in section 4.3.

4.3 Big Data Compressed Storage Algorithm

We proposed and implemented the big data compressed storage algorithm. Here we illustrated some key principles of the algorithm.

i) Combined all txt files into a new one txt file.

ii) New txt file is combined by triples {X, Y1Y2Y3Y4, Data}.

X represents the txt file's sequence number is X.

Y1Y2Y3Y4 is a number with 4 bits. It represents the line number in a txt file.

Data is the specific data in original txt files.

So, {X, Y1Y2Y3Y4, Data} is represents the specific data in line Y1Y2Y3Y4 in Xth txt in one experiment.

iii) If the data is zero, we simply discard the value.

5. Experiment and Analysis

5.1 Experiments Compressed Ratio of Big Data Compressed Storage Algorithm

We compared occupied disk space before compressed and after compressed with Big Data Compressed Storage Algorithm. The results are listed in table 3.

Table 3. Data Storage Details before and after Compressed Under the Algorithm

Experiment No.	Occupied Disk Space (MB)-before compressed BC	Occupied Disk Space (MB)-after compressed AC	Compressed Ratio =AC/BC
O-1#	1536	44	2.86%
G-1#	2161	52	2.41%
GO-1#	3748	130	3.47%

From table 3 we can get the conclusion that big data compressed storage algorithm has significant efficient, which means the average compressed ratio is 2.91%. In figure 8, we compare occupied disk space between original data and after compressed by our algorithm.

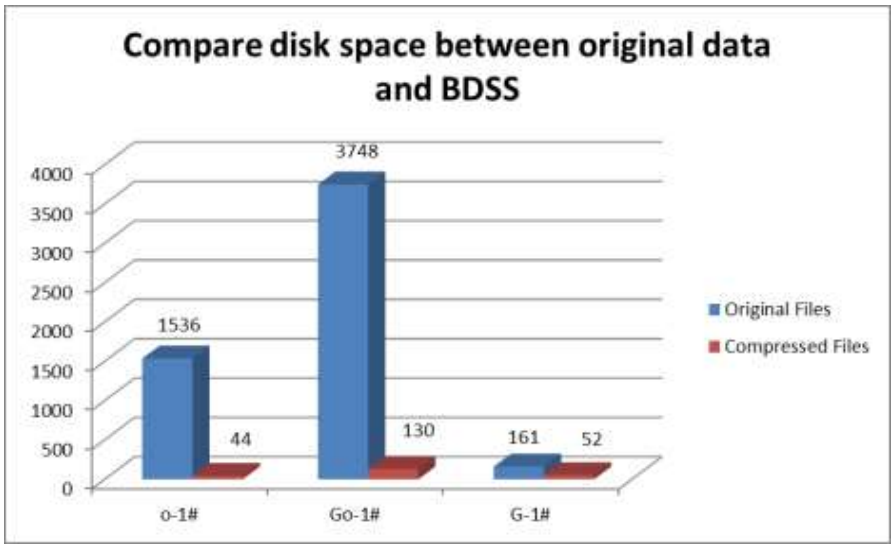


Figure 8. Compare Disk Space between Original Data and BDSS

5.2 Compressed Ratio between winrar and Big Data Compressed Storage Algorithm

As we all know, WinRAR is the most excellent compress software around the world. So we compare the compressed ratio between WinRAR and Big Data Compressed Storage Algorithm as showed in figure 9.

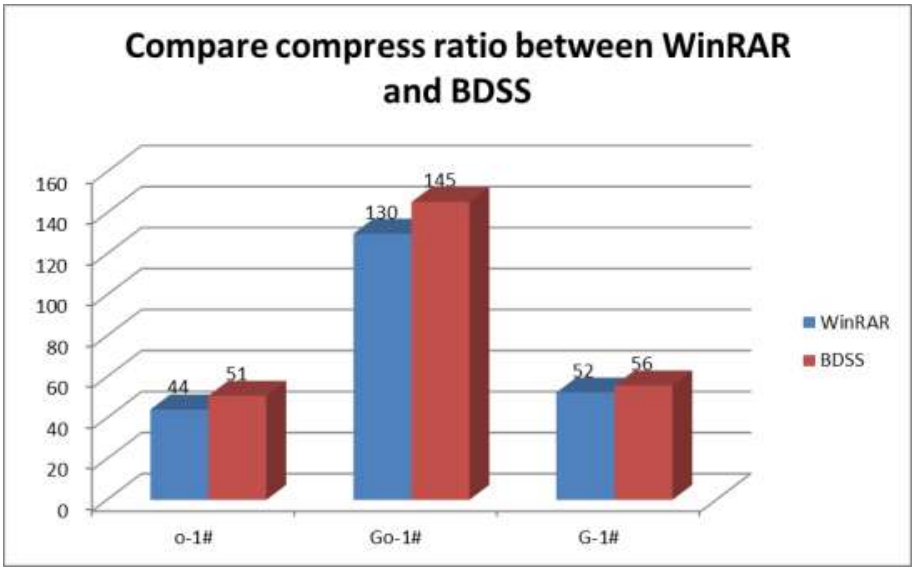


Figure 9. Compare Compress Ratio between WinRAR and BDSS

From figure 10 we can see that Big Data Compressed Storage Algorithm and WinRAR have nearly the same compress ratio. Also there are slight differences between them, but if we consider the slight difference under the originally data, it's nearly null. Take experiment O-1# as example, we can be calculated by formula 2.

$$(51-44)/1536=0.0046 \qquad (2)$$

There is one most important thing that WinRAR needs time to rar and unrar the data.

5.3 Analysis of time performance

As we discussed in the end of section 5.2, WinRAR has high compress ratio at the expense of time. When we use data every time, we needs time to rar or unrar the data. But rock burst experiment has high real-time requirement. So WinRAR is not suitable in rock burst experiment.

WinRAR and Big Data Compressed Storage Algorithm both need time to compress data before storage them. The compressed data by Big Data Compressed Storage Algorithm can be used directly but compressed data by WinRAR should be unrar before used. It's obviously that Big Data Compressed Storage Algorithm has good performance than WinRAR in the field of rock burst.

Since then, we still compare time for compress data between WinRAR and Big Data Compressed Storage Algorithm. The result is showed in figure 10.

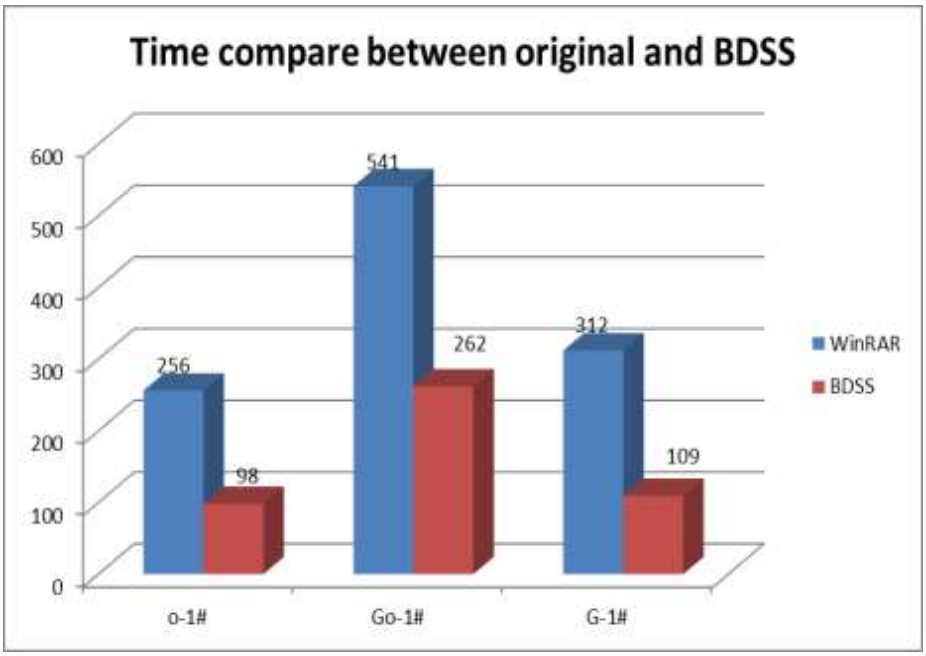


Figure 10. Time Compare between Original and BDSS

From figure 10 we can see that, from time for compress data view, Big Data Compressed Storage Algorithm is still better than WinRAR.

6. Conclusion

From all above, we can draw the conclusions that Big Data Compressed Storage Algorithm is suitable for rock burst, and the algorithm has excellent performance.

In the future, we will go on analysis the distribution law of experimental data and improve our algorithm to get a better compress ratio than ever.

Acknowledgements

This paper is supported by National Natural Science Foundation of China (51444003 and 61540003), and also Foundation of State Key Laboratory for GeoMechanics and Deep Underground Engineering (SKLGDUEK1523), and foundation of Beijing University of Civil Engineering and Architecture (00331615029, 00331613033 and 00331613028), and also Excellent Teachers Development Foundation of BUCEA (responded by Zhang Yu.). And partly supported by 2014 Technology Projects of Beijing Education Committee under Grant No. KM201410028011, and Beijing higher school teachers training of young teachers in the 2014 general foreign school training project（No.067145301400）. And also be supported by the science research project from Beijing Municipal Commission of Education (No.SM201410017003) and the outstanding academic leaders cultivate plan project (No.BIPT-BPOAL-2013).

References

[1] J. Fanzhi, X. Xiaodong and Z. Dongsheng, **(2003)**.
[2] H. Manchao, X. Jia, M. Coli, E. Livi and L. Sousa, **(2012)**.
[3] M. C. He, J. L. Miao and J. L. Feng, **(2010)**.
[4] M. C. He, H. P. Xie, S. P. Peng and Y. D. Jiang, **(2005)**.
[5] M. C. He, W. Nie, Z. Y. Zhao and W. Guo, **(2012)**.
[6] M. C. He, J. L. Miao and J. L. Feng, **(2007)**.
[7] M. He, J. Miao, D. Li and C. Wang, **(2007)**.

[8] M. He, **(2007)**.
[9] M. C. He, G. X. Yang, J. L. Miao, X. N. Jia and T. T. Jiang, **(2009)**.
[10] J. L. Miao, M. C. He, D. J. Li, F. J. Zeng and X. Zhang, **(2009)**.
[11] D. Li, X. Jia, J. Miao, M. He and D. Li, **(2010)**.
[13] E. Udoh and C. H. Hsu, **(2013)**.
[14] S. Krishnaveni, A. Satheesh and E. Kannan, **(2015)**.
[15] J. Dean and S. Ghemawat, **(2008)**.
[20] S. Landset, T. M. Khoshgoftaar, A. N. Richter and T. Hasanin, **(2015)**.
[21] H. Chennamsetty, S. Chalasani and D. Riley, **(2015)**.
[26] D. Catrein and C. QSC AG, **(2013)**.
[27] H. Manchao, X. Jia, M. Coli, E. Livi and L. Sousa, **(2012)**.
[28] M. C. He, J. L. Miao and J. L. Feng, **(2010)**.
[29] M. C. He, W. Nie, Z. Y. Zhao and W. Guo, **(2012)**.
[30] G. Weili, P. Yanyan, W. Hu and H. Manchao, **(2010)**.
[31] P. Zikopoulos and C. Eaton, **(2011)**.
[32] T. White, **(2012)**.
[33] S. Sagiroglu and D. Sinanc, **(2013)**.
[34] C. Wang, J. Wang, X. Lin, W. Wang, H. Wang, H. Li and R. Li, **(2010)**.
[35] K. Shvachko, H. Kuang, S. Radia and R. Chansler, **(2010)**.
[36] In Electrical, "Computer and Communication Technologies", **(2015)**.
[37] M. Kornacker, A. Behm, V. Bittorf, T. Bobrovytsky, C. Ching, A. Choi and M. Yoder, **(2015)**.
[38] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton and T. Vassilakis, **(2010)**.
[39] F. N. Afrati, D. Delorey, M. Pasumansky and J. D. Ullman, **(2014)**.
[40] M. He, X. Jia and D. Liu, **(2015)**.

Authors

Yu Zhang, he received his doctor's degree of Computer Science & Engineering in July 2009 at Beijing Institute of Technology. He is currently a teacher at School of Computer Science, Beijing University of Civil Engineering and Architecture. His research interests include Big Data, Distributed Network and Cloud Computing.

Yan-ping Bai, She received her doctor's degree of Economics in July 2007 at University of International Business and Economics. She is currently a teacher at college of management, Capital Normal University. Her research interests include Environmental Economics and Public Goods.

Zhao-yong Lv, He received his master's degree of control theory & engineering in March 2011 at Beijing University of Civil Engineering and Architecture. He is currently an experiment teacher at Computing Center. His research interests include Big Data and Control algorithm.

Yong-zhen Li, He received his master's degree of Software Engineering in July 2009 at Beijing Institute of Technology, now studying for his doctor's degree at Beijing University of Technology. He is currently a teacher at Beijing University of Civil Engineering and Architecture. His research interests include Cloud Computing and Code Confusion.

Zong-lei Mu, He obtained his master's degree of Computer Application Technology in January 2010 at Beijing University of Chemical Technology. He is currently as director of network information center at Beijing University of Civil Engineering and Architecture. His research interests include Network Security, Data Mining and Database Application.