International Journal of Engineering & Technology, 7 (4.7) (2018) 204-208

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Data Mining for Information Storage Reliability Assessment by Relative Values

Iskandar N. Nasyrov, Ildar I. Nasyrov, Rustam I. Nasyrov, Bulat A. Khairullin

Kazan Federal University
**Corresponding author E-mail: ecoseti@yandex.ru*

Abstract

The data ambiguity problem for heterogeneous sets of equipment reliability indicators is considered. In fact, the same manufacturers do not always unambiguously fill the SMART parameters with the corresponding values for their different models of hard disk drives. In addition, some of the parameters are sometimes empty, while the other parameters have only zero values.
The scientific task of the research consists in the need to define such a set of parameters that will allow us to obtain a comparative assessment of the reliability of each individual storage device of any model of any manufacturer for its timely replacement.
The following conditions were used to select the parameters suitable for evaluating their relative values:
1) The parameter values for normally operating drives should always be greater or lower than for the failed ones;
2) The monotonicity of changes in the values of parameters in the series should be observed: normally working, withdrawn prematurely, failed;
3) The first two conditions must be fulfilled both in general and in particular, for example, for the drives of each brand separately.
Separate averaging of the values for normally operating, early decommissioned and failed storage media was performed. The maximum of these three values was taken as 100%. The relative distribution of values for each parameter was studied.
Five parameters were selected (5 – "Reallocated sectors count", 7 – "Seek error rate", 184 – "End-to-end error", 196 – "Reallocation event count", 197 – "Current pending sector count", plus another four (1 – "Raw read error rate", 10 – "Spin-up retry counts", 187 – "Reported uncorrectable errors", 198 – "Uncorrectable sector counts"), which require more careful analysis, and one (194 – "Hard disk assembly temperature") for prospective use in solid-state drives, as a result of the relative value study of their suitability for use upon evaluating the reliability of data storage devices.

Keywords: *information, storage, reliability, parameter, estimation.*

1. Introduction

To ensure data security, provided that the effectiveness of the organization performance is maintained, it is necessary to copy information from the unreliable storage device to a new and reliable drive in a timely and complete manner. To this end, SMART technology (self-monitoring, analysis and reporting technology [1]) is used for internal assessment of the hard disk state of a computer, and also as a mechanism for predicting its possible failure. But even the same manufacturers do not always unambiguously fill the SMART parameters for different models of their drives with the corresponding values. Moreover, some of the parameters are sometimes empty, while the other parameters have only zero values. Hence, the scientific task of the research consists in the need to determine such a set of parameters that will allow us to obtain a comparative assessment of the reliability of each individual storage device of any model of any manufacturer for its timely replacement. As a result, five parameters were chosen which fully satisfy the selection criteria, and four more that satisfy in part. One additional parameter is proposed for prospective solid-state drives.

2. Methods

To search for and detect the patterns of data storage device failures, SMART data on hard disk drives from the company Backblaze website were analyzed [2, 3].
We studied 45 SMART parameters of 92530 drives of 93 models of 6 trademarks of HGST (Hitachi Global Storage Technologies), Hitachi (later HGST), Samsung, ST (Seagate), Toshiba, WDC (Western Digital) for the period from 10 April, 2013 to 31 December, 2016 [4].

It was found that 79.58% of the drives continued to function normally at the end of the period under study; 14.74% were decommissioned early, and 5.68% failed. The long operating time for individual drives, reaching a maximum of 31.3 years (274,412 hours for WDC WD10EADS) or 18.7 years (163730 hours for WDC WD800BB) may not be a mistake, but as Backblaze experts suggest [5], is the reality. This is confirmed by the fact that Western Digital hard disk drives were the best among the previously studied.

In total, over 80 SMART parameters are available, but most of them are not used by manufacturers. Therefore, Backblaze specialists recorded only 40 of them in 2013-2014, and starting from 2015 - 45 with numbers 1-5, 7-13, 15, 22, 183, 184, 187-201, 220, 222-226, 240 -242, 250-252, 254, 255 (in 2015 they added 22, 220,

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

222, 224, 226). However, when analyzing these 45 parameters, it was found that not all manufacturers use them. In Table 1, the plus sign indicates those parameters which non-empty values were found in at least one of the drives of any model of the specified manufacturer.

Table 1:. Manufacturers which use the SMART parameters measured by specialists of Backblaze

No.	HGST	Hitachi	Samsung	Seagate	Toshiba	WDC
1	+	+	+	+	+	+
2	+	+		+	+	
3	+	+	+	+	+	+
4	+	+	+	+	+	+
5	+	+	+	+	+	+
7	+	+	+	+	+	+
8	+	+	+	+	+	
9	+	+	+	+	+	+
10	+	+	+	+	+	+
11			+	+		+
12	+	+	+	+	+	+
13			+			
15				+		
22	+					
183			+	+		
184			+	+		+
187			+	+		+
188			+	+		+
189				+		
190			+	+		+
191				+	+	+
192	+	+		+	+	+
193	+	+		+	+	+
194	+	+	+	+	+	+
195			+	+		
196	+	+	+	+	+	+
197	+	+	+	+	+	+
198	+	+	+	+	+	+
199	+	+	+	+	+	+
200			+	+		+
201			+			
220					+	
222					+	
223				+	+	
224					+	
225				+		
226					+	
240				+	+	+
241				+		+
242				+		+
250				+		
251				+		
252				+		
254				+		+
255				+		

Thus, there is a very limited set of parameters that can be used to diagnose and assess the condition of drives of any manufacturer. First of all, it was proposed to use the number of overassigned sectors [6], and not as a separate unit parameter for assessing reliability, but as a collection of data: the current value, the average data accumulation rate from the moment the drive was put into operation, the instantaneous rate of change in the number of overassigned sectors since the last measurement. A similar combination of the mean and instantaneous rate of change of parameter values is used by specialists of Blackbaze [7]. This approach allows:

1) Track drives in which the current value is close to the limit level;

2) Keep under control the hard drives which slowly but steadily break down;

3) Take emergency measures for drives in which a one-time jump in the number of overassigned sectors raises concerns.

The proof of priority as to the number of overassigned sectors in evaluating the state of a hard drive is shown in [8], where the results of a study of 100,000 drives in servers around the world carried out by Google, are presented.

The following conditions can be used to select parameters suitable for estimating their relative values [9]:

1) The parameter values for normally operating drives should always be greater or always lower than for the failed ones;

2) Monotonicity of changes in the values of parameters in the series should be observed: normally operating, withdrawn prematurely, and failed;

3) The first two conditions must be fulfilled both in general and in particular, for example, for the drives of each brand separately.

The latter condition is introduced due to the fact that among all the studied drives, hard drives of the ST brand prevail with a significant margin.

3. Results and Discussion

First, we consider the parameters used by all manufacturers. They are marked in Table 1 with six pluses. These are parameters 1, 3-5, 7, 9 (Figure 1), and also 10, 12, 194, 196-199 (Figure 2).

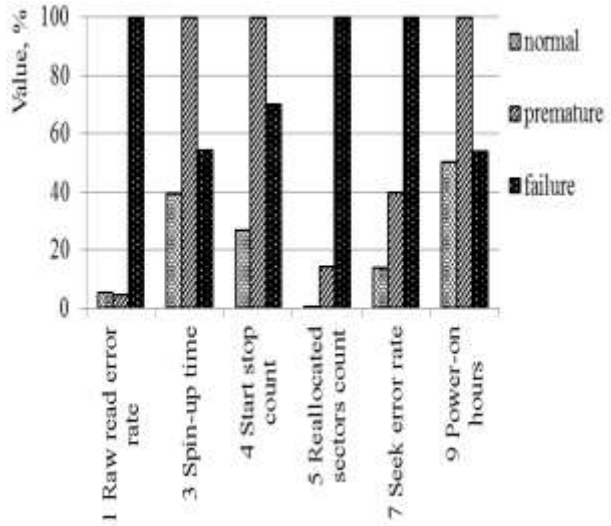


Figure 1:. Averaged values of the first six parameters used by all manufacturers for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices

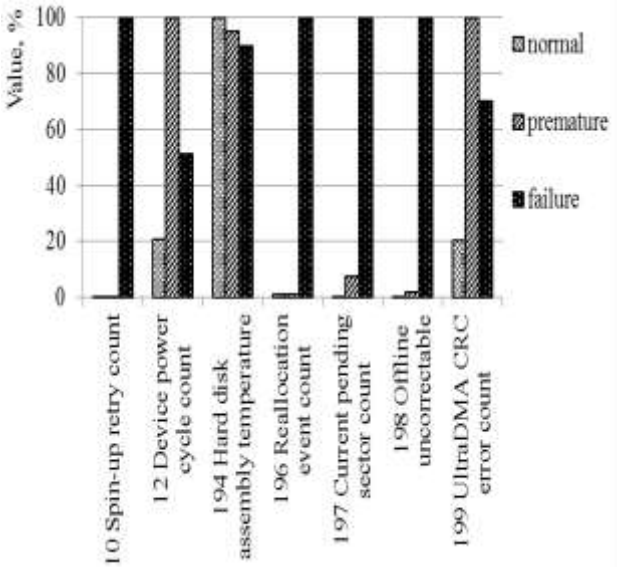


Figure 2:. Averaged values of the second seven parameters used by all manufacturers for the normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices

Averaging was performed for the normally operating, withdrawn prematurely and failed storage devices separately. The maximum of these three values was taken as 100%. The relative distribution of the values for each parameter was studied.

As can be seen from Fig. 1, parameters 5 and 7 satisfy the first two conditions very well, and parameters 3, 4 and 9 satisfy them partially (withdrawn prematurely drives have those parameter values larger than those that failed). Parameter 1 also partially satisfies formally, however, the difference in values between normally operating and withdrawn prematurely drives is small, and it is great in comparison with the failed ones. This circumstance is very useful from the point of view of reliability assessment.

It can be seen from Fig. 2 that parameters 10, 196-198 satisfy the first two conditions very well, and parameters 12 and 199 satisfy them only partially as in Fig. 1. Parameter 194 is also formally satisfied, however, the relative difference in the temperature between the normally operating, withdrawn prematurely and failed drives is small, which can lead to difficulties in practical use.

Secondly, we consider the parameters used by five (8, 192, 193), four (2) and three (11, 184, 187, 188, 190, 191, 200, 240) manufacturers. In Table 1, they are marked with the appropriate number of pluses. Figure 3 shows the first parameters specified above, as well as the parameters recommended by the specialists of Backblaze 187, 188. As can be seen from Figure 3, the only fully satisfying parameter is the recommended one with the number 187. And the second recommended parameter with the number 188 turned out to be satisfying in part, like the parameters 192 and 193.

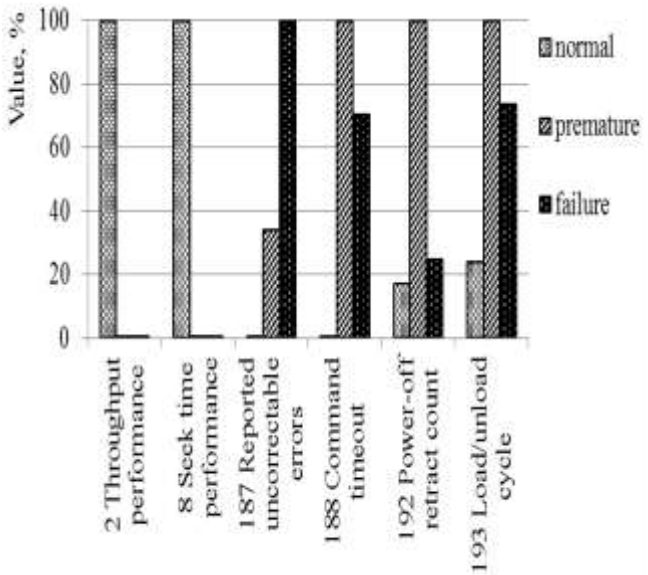


Figure 3:. Averaged values of parameters used by five, four and three manufacturers for normally operating (left in each group), withdrawn prematurely (in the middle), failed (on the right) storage devices

Figure 4 shows the remaining specified parameters used by the three manufacturers. It can be seen that only the parameter 184 satisfies the conditions, and the other parameters obviously do not satisfy them. The parameter 190, like the previously mentioned 194, also formally satisfies those conditions, however, the relative difference in the values of the internal air temperature between normally operating, withdrawn prematurely and failed drives is also small, which can entail the same difficulties in practical use. The parameter with the number 240 satisfies the second condition in part.

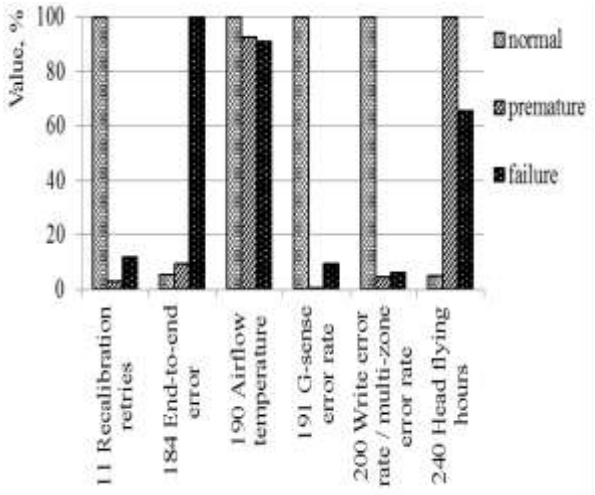


Figure 4:. Average values of the parameters used by the three manufacturers for normally operating (on the left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices

As a result, we have obtained a list of 11 parameters most satisfying the first two conditions. These are the parameters 1, 5, 7, 10, 184, 187, 190, 194, 196-198. The second condition is partially satisfied with parameters 3, 4, 9, 12, 188, 192, 193, 199, 240. Then they were checked for compliance with the third condition. As it turned out, parameters 5, 7, 184, 196, 197 satisfy it very well. Parameters 1, 10, 187 satisfy it partially in view of the small statistics on Samsung drives or the available error near zero of ST storage devices (Figures 5-7).

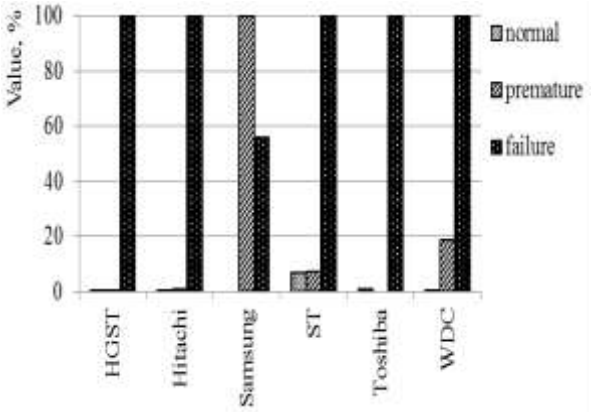


Figure 5:. Average values of the parameter "1 Raw read error rate" for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices of different manufacturers

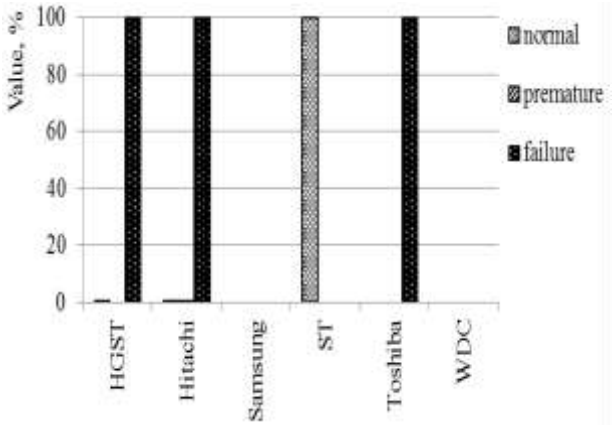


Figure 6:. Averaged values of the parameter "10 Spin-up retry count" for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices of different manufacturers

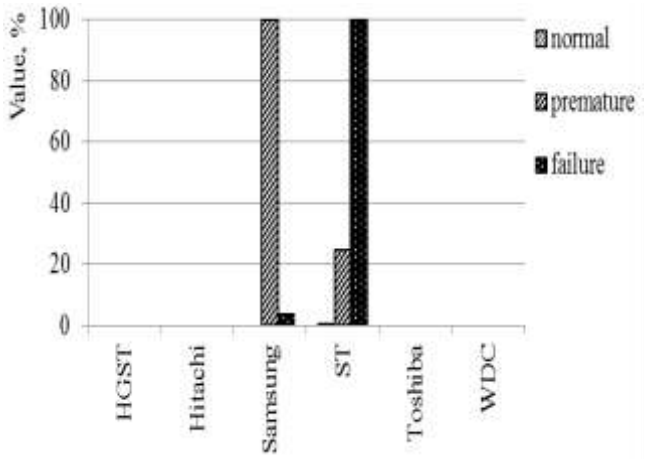


Figure 7:. Averaged values of the parameter "187 Reported UNC errors" for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices of different manufacturers.

The temperature parameters 190 and 194 do not completely satisfy the condition, and the parameter 198 satisfies in part, because of the small statistics volume for WDC products, and therefore we leave it for more detailed consideration (Figures 8-10). However, although the hard disk housing temperature parameter 194 is not suitable for reliability estimation, nevertheless, for solid state drives, its use can be very useful. Therefore, given the perspective, this fact must be borne in mind.

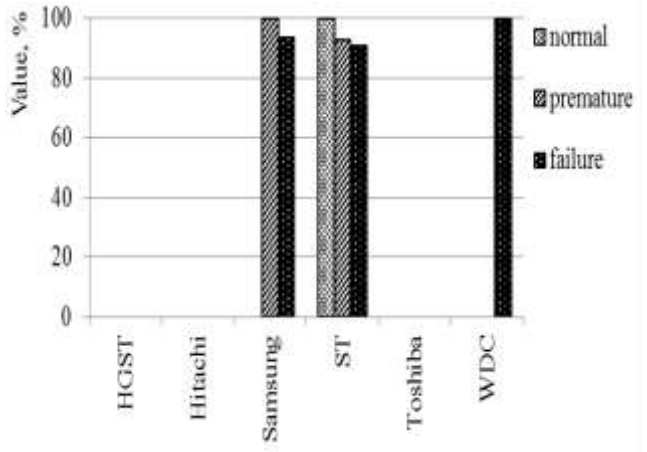


Figure 8:. Averaged values of the parameter "190 Airflow temperature" for normally operating (left in each group), withdrawn prematurely (in the middle), failed (right) data drives of different manufacturers

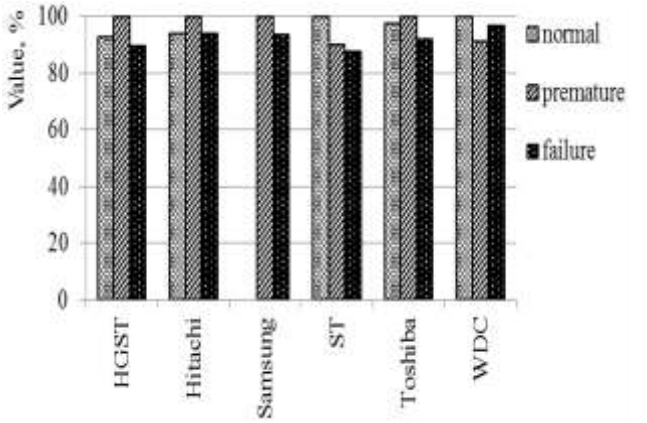


Figure 9:. Averaged values of the parameter "194 Hard disk assembly temperature" for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices of different manufacturers

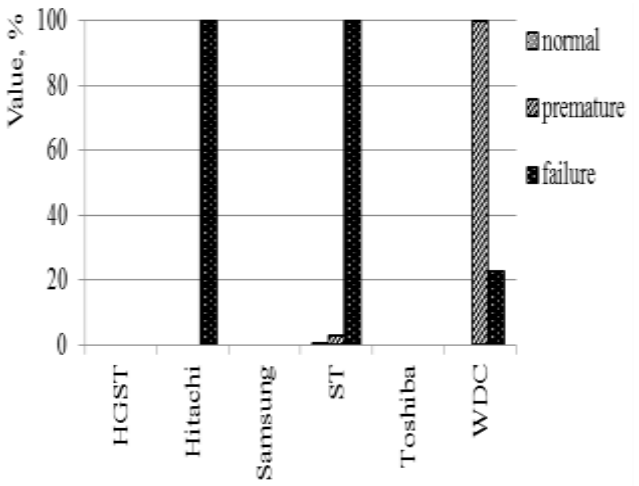


Figure 10:. Averaged values of the parameter "198 Uncorrectable sector count" for normally operating (left in each group), withdrawn prematurely (in the middle), and failed (on the right) storage devices of different manufacturers

4. Summary

Thus, after considering the relative values for further study on the suitability for use in assessment of the reliability of data storage devices, we still have five parameters (5 Reallocated sectors count, 7 Seek error rate, 184 End-to-end error, 196 Reallocation event count, 197 Current pending sector count) plus four more (1 Raw read error rate, 10 Spin-up retry count, 187 Reported uncorrectable errors, 198 Uncorrectable sector count), which required more thorough analysis, and one (194 Hard disk assembly temperature) for further perspective use for solid state drives.

5. Conclusions

Similar studies on the same data with disparate groups of disks were conducted in [10], where a search for universal predictors of disk failures that could be applied to disks of all brands and models was carried out. The main problem was also a significant number of SMART-parameters, which were absent for most brands and models of disks of the specified data set. As a result, the authors were forced to discard parameters that were absent in at least 90% of the disks, after which 21 parameters remained.

In [11-15], SMART parameters of the specified data set were also used to determine the intensity and prediction of disk drive failures. Therefore, the choice of parameters for assessing the reliability of information storage devices based on the values of SMART parameters is really important for ensuring data security in any organization.

Acknowledgement

The work is carried out according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

[1] S.M.A.R.T. From Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org/wiki/S.M.A.R.T. Checked on 10/03/2018.
[2] Hard Drive Data and Stats / Backblaze. URL: https://www.backblaze.com/b2/hard-drive-test-data.html. Checked on 10/03/2018.
[3] Beach B. Reliability Data Set For 41,000 Hard Drives Now Open Source. URL: https://www.backblaze.com/blog/hard-drive-data-feb2015/. Checked on 10/03/2018.
[4] Nasyrov R.I., Nasyrov I.N., Timergaliev S.N. Cluster analysis of information storage devices that failed during operation in a large

data center // Information technology. Automation. Updating and solving the problems of training highly qualified personnel (ITAP-2017): materials of the international scientific-practical conference on 19 May, 2017. - Naberezhnye Chelny: KFU, 2017. - Pз. 95-102. URL: https://cloud.mail.ru/public/LBcn/phxM8D1S5.

[5] Beach B. How long do disk drives last? URL: https://www.backblaze.com/blog/how-long-do-disk-drives-last/. Checked on 10/03/2018.

[6] Nasyrov R.I. Grading indicators for information storage devices by reliability degree // VIII Kama Readings: a collection of reports of the All-Russian Scientific and Practical Conference on 22 April, 2016. - In 2 parts - Part 1. - Naberezhnye Chelny: CPI NCHI KFU, 2016. - 124.URL: https://cloud.mail.ru/public/JPMr/8qr2jXKFC.

[7] Klein A. Hard Drive Reliability Stats for Q1 2015. URL: https://www.backblaze.com/blog/hard-drive-reliability-q1-2015/. Checked on 10/03/2018.

[8] Pinheiro E., Weber W.-D., Barroso L.A. Failure Trends in a Large Disk Drive Population // The Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07). San Jose, California, USA, February 13-16, 2007. URL: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/ru//archive/disk_failures.pdf.

[9] Nasyrov R.I., Nasyrov I.N. Choice of parameters for the method of forecasting the reliability of data storage devices in large data centers. // Quality. Innovation. Education. - 2017. - No. 5 (144). - Pp. 40-48. URL: https://library.ru/item.asp?id=29869743.

[10] Rincón C.A.C., Paris J.-F., Vilalta R., Cheng A.M.K., Long D.D.E. Disk failure prediction in heterogeneous environments // Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2017. Seattle, WA, USA, July 9-12, 2017. URL: http://ieeexplore.ieee.org/document/8046776/.

[11] Qian J., Skelton S., Moore J., Jiang H. P3: Priority based proactive prediction for soon-to-fail disks // Proceedings of the 10th IEEE International Conference on Networking, Architecture and Storage, NAS 2015. Boston, MA, USA, August 6-7, 2015. – 7255224. – p. 81-86. URL: http://ieeexplore.ieee.org/document/7255224/.

[12] Botezatu M.M., Giurgiu I., Bogojeska J., Wiesmann D. Predicting disk replacement towards reliable data centers // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. San Francisco, California, USA, August 13-17, 2016. – p. 39-48. URL: https://dl.acm.org/citation.cfm?doid=2939672.2939699.

[13] Chaves I.C., de Paula M.R.P., Leite L.G.M., Queiroz L., Pordeus J.P., Machado J.C. BaNHFaP: A Bayesian Network Based Failure Prediction Approach for Hard Disk Drives // Proceedings of the 5th Brazilian Conference on Intelligent Systems, BRACIS 2016. Recife, Pernambuco, BR, October 9-12, 2016. – 7839624. – p. 427-432. URL: http://ieeexplore.ieee.org/document/7839624/.

[14] Gaber S., Ben-Harush O., Savir A. Predicting HDD failures from compound SMART attributes // Proceedings of the 10th ACM International Systems and Storage Conference, SYSTOR '17. Haifa, Israel, May 22-24, 2017. – Article No. 31. URL: https://dl.acm.org/citation.cfm?doid=3078468.3081875.

[15] Gopalakrishnan P.K., Behdad S. Usage of product lifecycle data to detect hard disk drives failure factors // Proceedings of the ASME International Design Engineering Technical Conference. Cleveland, Ohio, USA, August 6–9, 2017. URL: http://proceedings.asmedigitalcollection.asme.org/proceeding.aspx?articleid=2662132.