Parameters Selection for Information Storage Reliability Assessment and Prediction by Absolute Values

*Iskandar N. Nasyrov, Ildar I. Nasyrov, Rustam I. Nasyrov, Bulat A. Khairullin
Kazan Federal University
e-mail: ecoseti@yandex.ru
*Corresponding Author

Abstract- The problem of choosing parameters for estimating and predicting the reliability of an information storage device is considered. It is that manufacturers of hard disk drives do not always unambiguously fill SMART parameters with corresponding values for different models. 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 estimating and predicting the reliability of each individual storage device of any model of any manufacturer for its timely replacement.
For this purpose, a separate grouping of normally operating, early-decommissioned and failed drives was performed .The scale of the values for each parameter was divided into ranges. A number of storage devices that fall within a certain range of values, was counted. The distribution of storage devices was studied in absolute values for each parameter under consideration.
The following conditions were used to select suitable parameters for estimating and predicting the reliability of the parameters based on their values:
1) The number of normally operating drives that have a reliability parameter value within the range of large values should always be less than those that failed;
2) The monotonicity of the increase in the number of drives in the series should be observed for large values of reliability parameters: normally operating, early removed, and failed;
3) The first two conditions must be fulfilled both in general and in particular, for example, for the drives of each manufacturer separately.
Nine parameters were selected as a result of studying absolute values for the suitability to use in evaluating and predicting the reliability of data storage devices: 1 Raw read error rate, 5 Reallocated sectors count, 7 Seek error rate, 10 Spin-up retry count, 184 End-to-end error, 187 Reported uncorrectable errors, 196 Reallocation event count, 197 Current pending sector count, 198 Uncorrectable sector count.
***Key words*:** information, storage device, reliability, parameter, predicting.

1. Introduction

To ensure data security and timely copying of information from an unreliable to a new and reliable drive, SMART technology (self-monitoring, analysis and reporting technology [1]) is used. It is used both for internal evaluation of a computer's hard disk, and for predicting its possible failure. But the problem is that the manufacturers of different models of their hard disk drives do not always unambiguously fill the SMART parameters with the appropriate 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 evaluate and predict the reliability of information for each individual drive of any model and any manufacturer for its timely replacement. As a result, nine parameters were selected, fully or partially satisfying the selection conditions.

2. Methods

The objects of the study were information storage devices in one of the world's largest commercial data-center of the company Backblaze. The subject of the study is the reliability of the hard disk drives used there. To find the patterns of failure of the drives, SMART-data given on the company's website have been analyzed [2, 3].
In total, over 80 SMART parameters are theoretically available, but in the course of practical analysis it was found that not all manufacturers use all of them. Taking into account this circumstance, 45 SMART parameters were studied for the period from 10 April, 2013 to 31 December, 2016, for 92530 drives of 93 models of 6 brands: HGST (Hitachi Global Storage Technologies, 17497 pcs.), Hitachi (later HGST, 13246 pcs.), Samsung (18 pcs.), ST (Seagate, 57438 pcs.), Toshiba (355 pcs.), WDC (Western Digital , 3976 pcs.) [4]. It was revealed that 79.58% of the drives continued to function normally at the end of the period under study, 14.74% were early withdrawn from use, and 5.68% failed.
Each of these types of drives was separately grouped by the values of the parameters which were equal to 0 or 1 or 2, or were within the ranges from 3 to 4 inclusive and similarly below: 5-9, 10-99, 100-999, 1000-9999,

ISSN 1943-023X
Received: 5 Jul 2018/Accepted: 10 Aug 2018

10000 -99999, 100000-999999, 1000000-9999999, 10000000-99999999, 100000000-999999999, from 1,000,000,000 and above. The number of drives of each type was calculated separately, the absolute values of the parameters of which fell within the corresponding range.

The following conditions were used to select the appropriate parameters for estimating the reliability based on their values [5]:

1) The number of normally operating drives that have a reliability parameter value within the large values range should always be less than those that failed;

2) For large values of reliability parameters, the monotonicity of increase in the number of drives within the series should be observed: normally operating, early withdrawn from use, and failed;

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

The latter condition was introduced due to the fact that in the company Backblaze, hard drives ST prevail with a significant margin among all the used drives. And with any averaging, the final result will be determined by the values of the drive parameters of the specified brand.

3. Results And Discussion

Figure 1 shows the distribution of the number of hard drives that have a value of SMART parameter "1 Raw read error rate" within the above ranges. Normally operating drives are located on the left in each group within the range, early withdrawn from use in the middle, and failed on the right. The number of drives with zero values of the parameter is given by the figures located at the top for the normally operating drives, early withdrawn from use - in the middle, and failed - at the bottom.

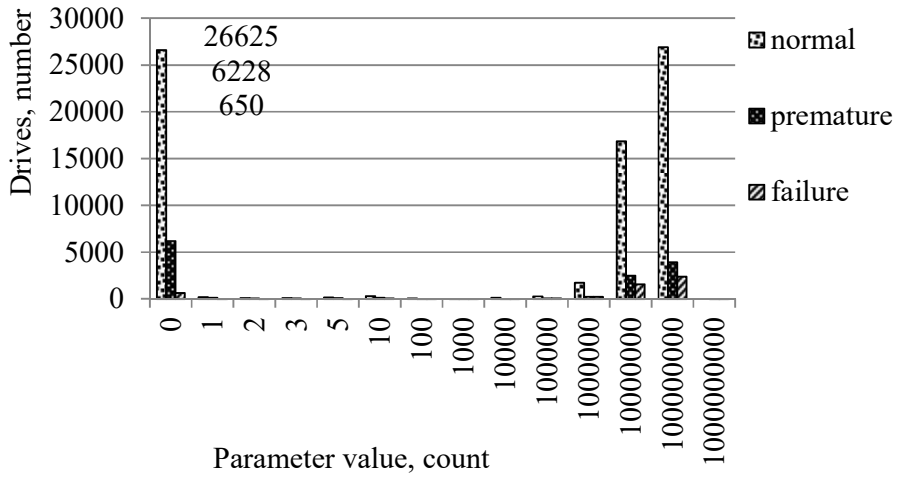

Figure 1 - The number of drives that have the SMART parameter value "1 Raw read error rate" within a certain range: normally operating (left in each group), early withdrawn from use (in the middle), and failed (right)

As can be seen on Fig. 1, parameter 1 is not satisfied with the conditions imposed to absolute values in this case. Therefore, we must additionally consider its behavior separately for each manufacturer, i.e. check for compliance with the third condition.

The parameter "5 Reallocated sectors count" completely corresponds to the conditions presented for reliability assessment (Figure 2). It is necessary to pay attention that here and further the vertical axis is truncated.

ISSN 1943-023X
Received: 5 Jul 2018/Accepted: 10 Aug 2018

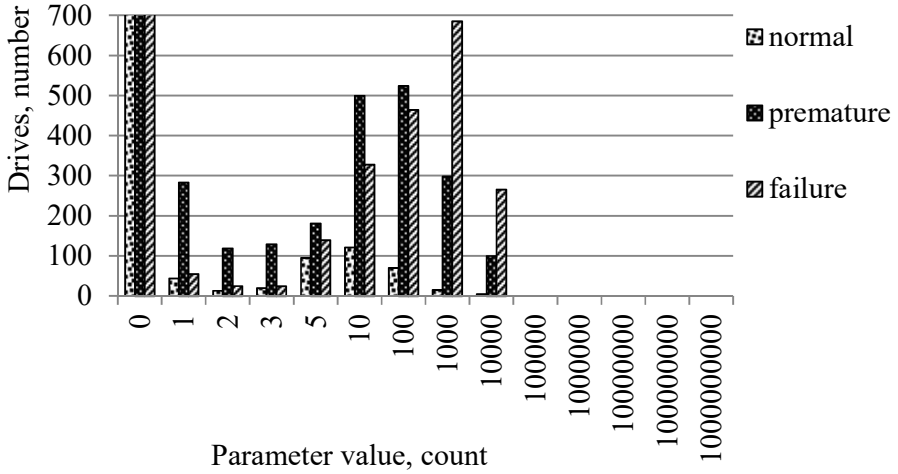

Figure 2 - The number of drives that have the SMART parameter value "5 Reallocated sectors count" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (on the right)

Like parameter 1, parameter "7 The Seek error rate" does not satisfy with the conditions set to absolute values (Figure 3). Therefore, in the same way, it is also necessary to consider its behavior separately for each manufacturer, i.e. check for compliance with the third condition.

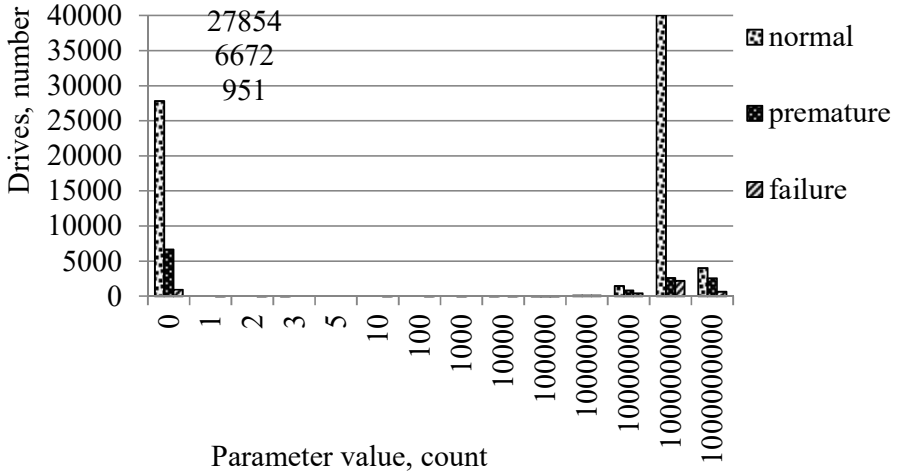

Figure 3 - Number of drives that have the SMART parameter value "7 Seek error rate" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (on the right)

Parameters "10 Spin-up retry count" (the number of retries to spin-up disks to the operating speed in case if the first attempt was unsuccessful) and "184 End-to-end error" (the parity of the data between the host and the hard disk) fully correspond to the set conditions. However, the number of values that are different from zero are small, so the graphic images are not very expressive (Figures 4, 5).

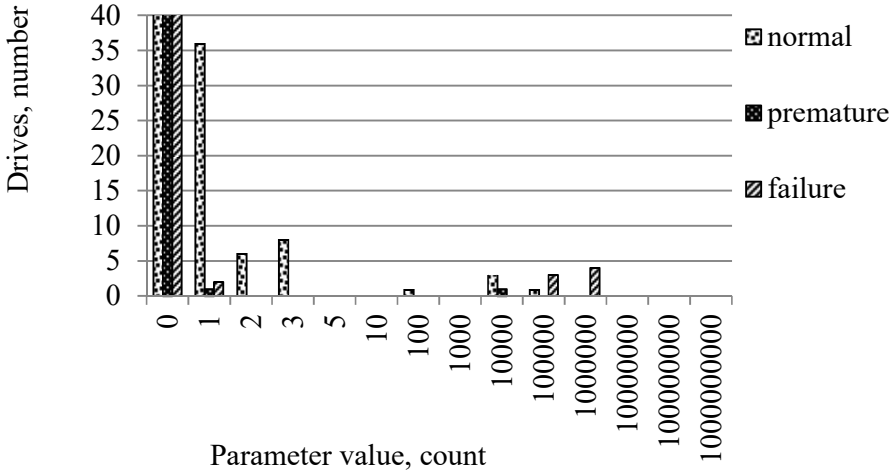

Figure 4 - The number of drives that have a value of SMART parameter "10 Spin-up retry count" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (right)

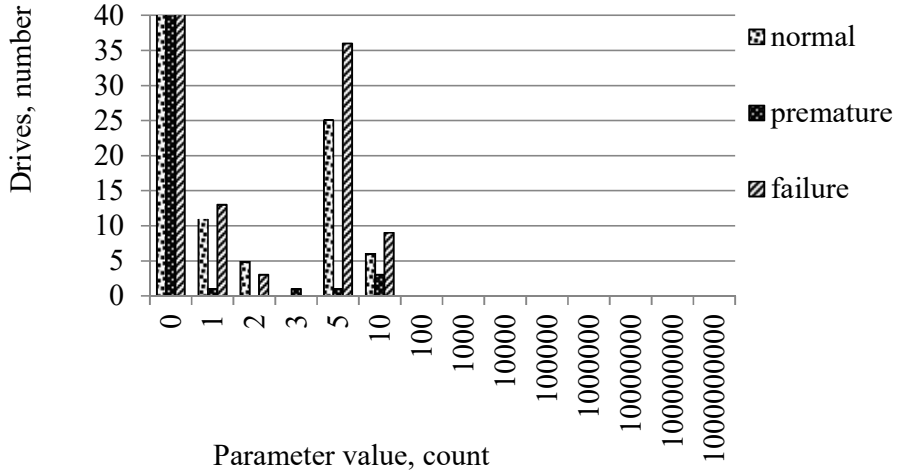

Figure 5 - The number of drives that have the value of the SMART parameter "184 End-to-end error" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (on the right)

Parameters "187 Reported uncorrectable errors" (errors that could not be recovered using hardware removal methods), "196 Reallocation event count", "197 Current pending sector count", and "198 Uncorrectable sector count" (the number of sectors not corrected by disk means) completely correspond to the set conditions (Figures 6-9).

ISSN 1943-023X
Received: 5 Jul 2018/Accepted: 10 Aug 2018

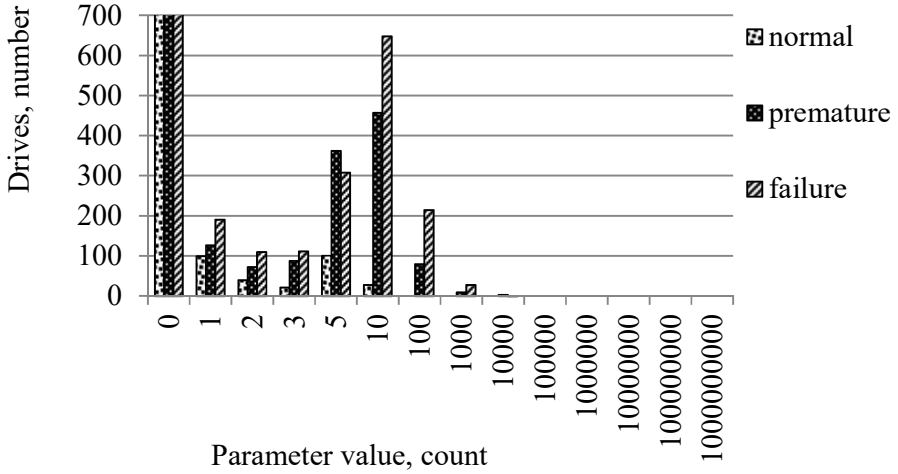

Figure 6 - The number of drives that have the value of the SMART parameter "187 Reported uncorrectable errors" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (right)

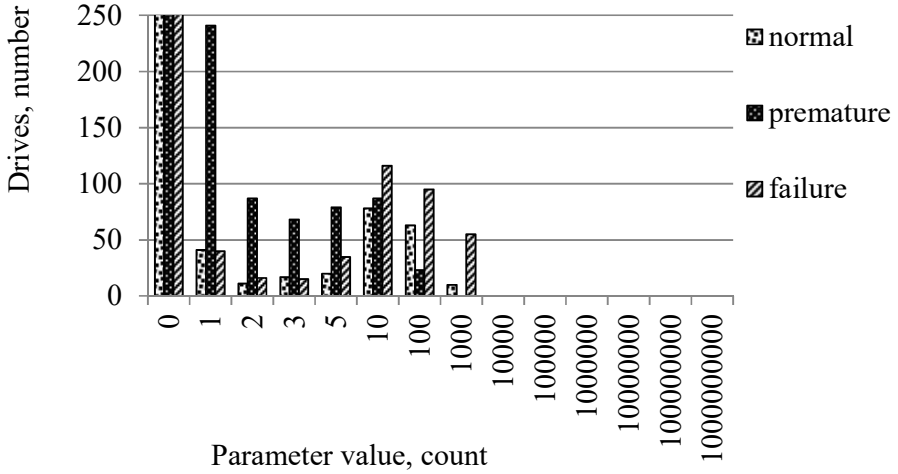

Figure 7 - The number of drives that have the value of the SMART parameter "196 Reallocation event count" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (right)

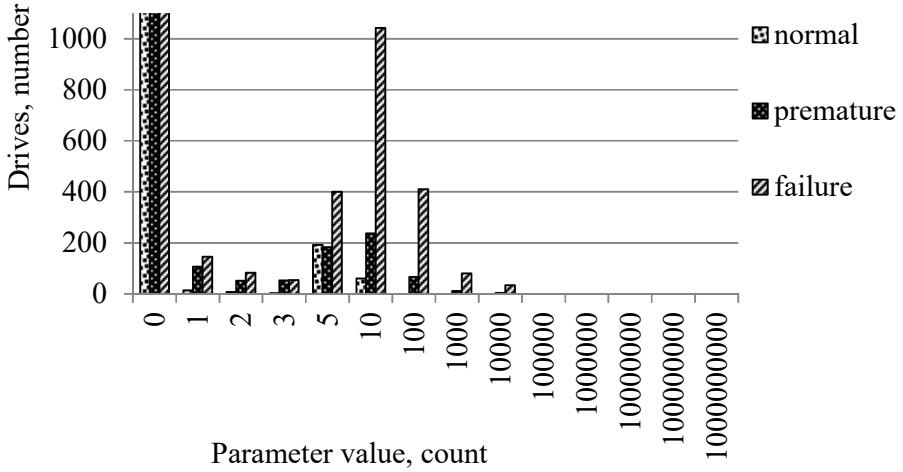

Figure 8 - The number of drives that have the value of the SMART parameter "197 Current pending sector count" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (right)

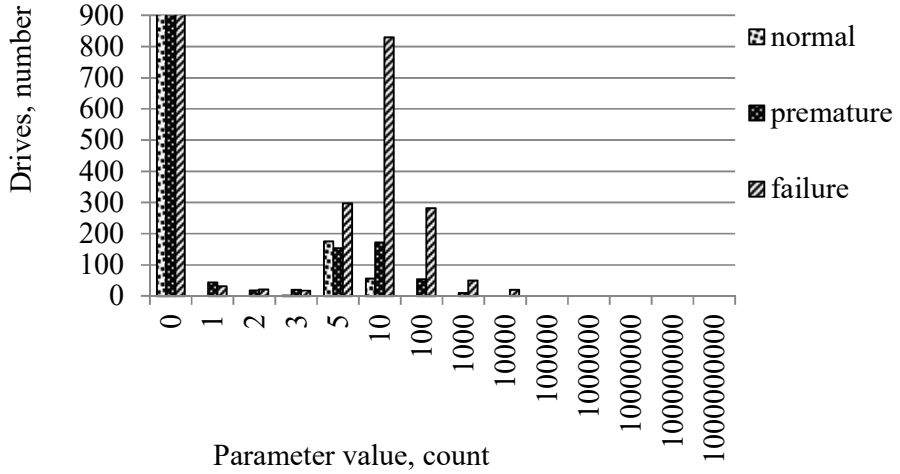

Figure 9 - The number of drives that have the value of the SMART parameter "198 Uncorrectable sector count" in a certain range: normally operating (left in each group), early withdrawn from use (in the middle), failed (on the right)

Verification by the third condition applied to the specified parameters separately for each manufacturer revealed that all the drives are sufficiently adequate for it. The exception is ST branded drives with parameters 1, 7 and 10. And as it was noted earlier, their inconsistencies contributed too much to distorting the overall results, because of their numerousness. In view of this circumstance, these parameters can also be used in future to evaluate and predict reliability.

4. Summary

Thus, the study of the absolute values of the parameters "1 Raw read error rate", 5 "Reallocated sectors count", "7 Seek error rate", "10 Spin-up retry count", "184 End-to-end error", "187 Reported uncorrectable errors", "196 Reallocation event count", "197 Current pending sector count", "198 Uncorrectable sector count" showed their suitability for use in assessing and predicting the reliability of data storage devices.

5. Conclusions

Analogous studies using the same data with disparate groups of disks were carried out in [6], where a search was conducted for universal predictors of disk failures that could be applied to disks of all brands and models. 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 for at least 90% of the disks, after which 21 parameters remained for consideration.

In [7-12], SMART parameters of the specified data set were used to determine the intensity and prediction of disk drive failures, too. Therefore, the choice of parameters for estimating and predicting the reliability of information storage devices based on SMART parameter values is important for ensuring data security in any organization, indeed.

6. Acknowledgments

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

[5] Nasyrov, R.I. Parameters of mathematical models for predicting the reliability of information storage devices in large data centers / R.I. Nasyrov, I.N. Nasyrov // Intellect. Innovation. Investments. - 2017. - No. 8. - Pp. 56-61. - Access mode: https://elibrary.ru/item.asp?id=30541824.

ISSN 1943-023X
Received: 5 Jul 2018/Accepted: 10 Aug 2018

[6] 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/.

[7] 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/.

[8] Voronin V.V., Davydov O.A. Technology of monitoring the technical condition of the equipment in local computer networks // Informatics and control systems. - 2016. - No. 2 (48). - Pp. 79-91. - Access mode: https://elibrary.ru/item.asp?id=26181365.

[9] 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.

[10] 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/.

[11] 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.

[12] 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.