УДК 004.33:338

Насыров Искандар Наилович,
профессор, д.э.н., к.ф.-м.н. (ecoseti@yandex.ru)
Набережночелнинский институт (филиал)
ФГАОУ ВО «Казанский (Приволжский) федеральный университет»
ORCID 0000-0003-3293-6965
Насыров Ильдар Искандарович,
руководитель группы, к.т.н. (ildarec@mail.ru)
ООО «Глобал Дата Консалтинг энд Сервисез»
ORCID 0000-0002-0186-2871
Насыров Рустам Искандарович,
руководитель портфеля проектов (rinasyrov@gmail.com)
ООО «Газпромнефть – Цифровые решения»
ORCID 0000-0002-4923-4532

БОЛЬШИЕ ДАННЫЕ ПО НАДЕЖНОСТИ НАКОПИТЕЛЕЙ ИНФОРМАЦИИ В DATA-ЦЕНТРАХ

Аннотация: На основании анализа научных публикаций, использующих находящиеся в открытом доступе значения параметров накопителей информации data-центров компании Backblaze за длительный период, сделан вывод о наличии проблемы «больших данных», приводящей к существенному перекосу тематики исследований.

Ключевые слова: большие данные, накопитель информации, data-центр, надежность.

Nasyrov Iskandar Nailovich, Nasyrov Ildar Iskandarovich, Nasyrov Rustam Iskandarovich

BIG DATA ON STORAGE DEVICES RELIABILITY IN DATA CENTERS

Abstract: Based on the analysis of scientific publications using the publicly available values of Backblaze data centers storage devices for a long period the conclusion is made, that there is a problem of «big data», leading to a significant disbalance in the research topics.

Keywords: big data, data storage, data center, reliability.

Введение

Актуальность исследования связана с дисбалансом между продолжающимся ростом генерируемых в цифровой экономике данных и возможностью их сохранения. В частности, в системах централизованного хранения данных (data-центрах) из-за недостаточной скорости копирования информации с дублирующих накопителей на новый в случае, если основной жесткий диск (HDD – hard disk drive) или твердотельный накопитель (SSD – solid state drive) заменен по причине выхода его из строя, значительные ресурсы отвлекаются на восстановление. Замедляется доступ остальных пользователей ко всем накопителям, с которых копируется информация. Дело в том, что рост скорости передачи данных на порядок отстает от роста емкости накопителей, в связи с чем на первый план выходит задача своевременной оценки и прогнозирования надежности накопителей информации для обеспечения возможности предварительной подготовки к проведению мероприятий по восстановлению данных при их замене. Целью исследования является выработка предложений по разработке подобной системы оперативной оценки и прогнозирования надежности накопителей информации.

Методы

Информационной базой исследования послужили ежедневно записываемые SMART-данные (self-monitoring, analysis and reporting technology – технология самоконтроля, анализа и отчетности) накопителей, находящиеся в свободном доступе на сайте одной из крупнейших в мире групп коммерческих data-центров компании Backblaze (https://www.backblaze.com/b2/hard-drive-test-data.html). Они удовлетворяют всем требованиям для прогнозирования сбоев, в связи с чем исследователи всего мира активно их используют в своей работе [1]. В качестве метода исследования выбран анализ причин выхода из строя накопителей путем сравнения значений параметров надежности продолжающих функционировать и отказавших накопителей информации.

Результаты

Изучение состояния исследований в данной области по трем базам научных публикаций (РИНЦ, Scopus, Web of Science) показало, что вместо ожидаемого увеличения количества публикаций соответственно росту доступных для анализа данных наблюдается их стабилизация с максимумом в 2019 году для иностранных авторов (рис. 1) и максимумом в 2018 году для российских авторов (рис. 2).

Рисунок 1. Число публикаций иностранных авторов с данными Backblaze (шт.) по годам

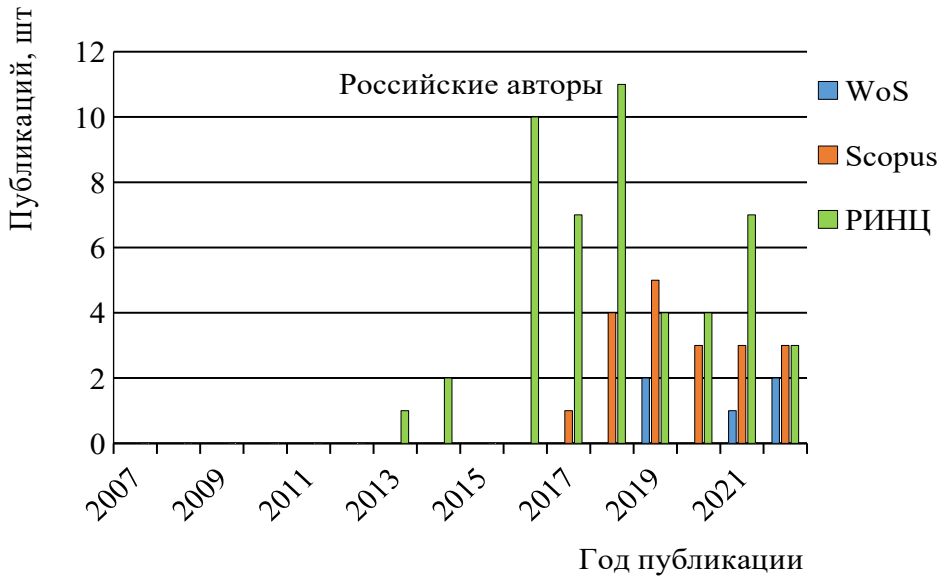

Рисунок 2. Число публикаций российских авторов с данными Backblaze (шт.) по годам

В последнем случае часть публикаций является повтором, например, упомянута в [2], а приведена в [3], или [4, 5] продублированы в [6, 7]. Это произошло вследствие ошибки службы поддержки публикационной активности аффилированной организации авторов. Отдельные единичные случаи частичных повторов есть и у других иностранных и российских авторов.

Группировка публикаций (всего 339 статей на 01.05.2022) по направлениям использования данных показала, что у иностранных (рис. 3) и российских (рис. 4) авторов в основном происходило упоминание компании Backblaze (42,9% и 32,9% соответственно), предлагались системы кодирования дублирующих частей информации типа RAID (redundant array of independent disks – избыточный массив независимых дисков) для повы-

шения надежности восстановления (12,4% и 12,3%), рассматривались методы машинного обучения для прогнозирования отказов (14,7% и 5,5%) и отдельно среди них – нейронные сети (17,7% и 11,0%). В российских публикациях по анализу данных значительную часть из общей доли в 27,4% обеспечили работы авторов настоящей статьи. У иностранных авторов доля публикаций по этой теме мала, всего 1,1%.

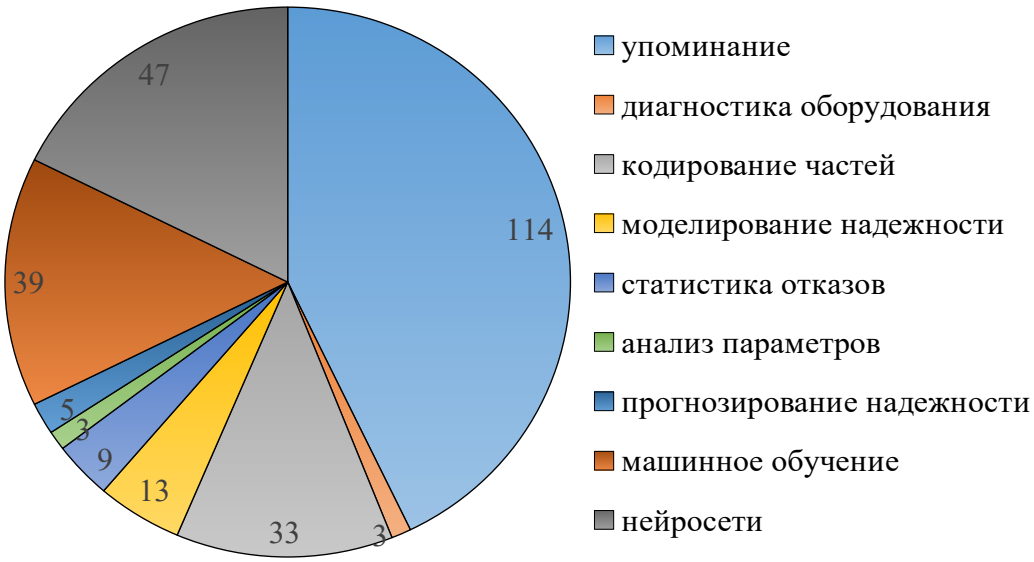

Рисунок 3. Число публикаций иностранных авторов с данными Backblaze (шт.) по темам

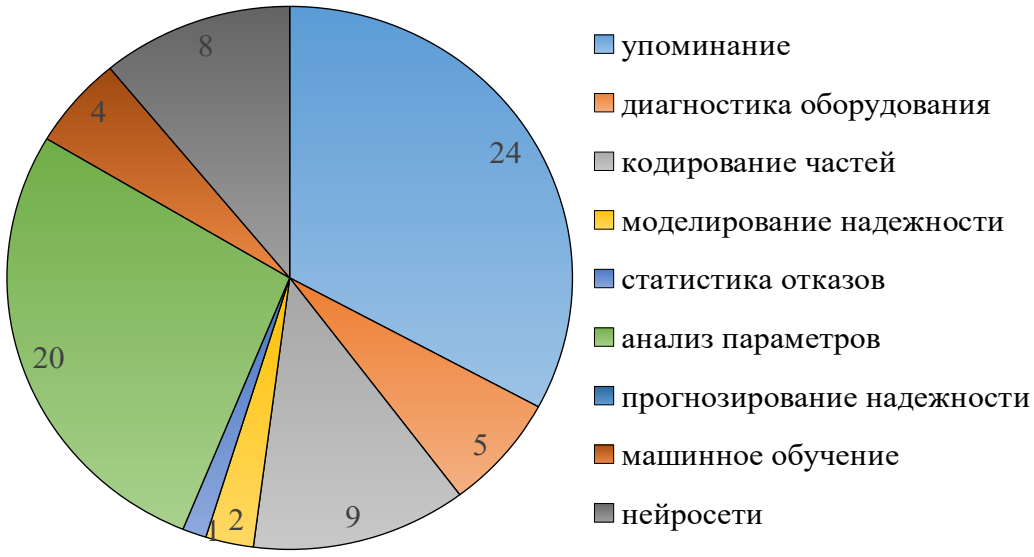

Рисунок 4. Число публикаций российских авторов с данными Backblaze (шт.) по темам

Углубленный анализ состояния исследований по данной тематике выявил некото-

рые особенности. Первая: вначале объявляется, что исследования проводятся на реальном массиве больших данных, однако по факту из него выбирается только их незначительная доля. Вторая: разрабатываются сложные системы прогнозирования на нейронных сетях с многоступенчатой предварительной подготовкой исходных данных и хорошими итоговыми показателями, но в конце сообщается, что система для накопителей других производителей или для других моделей тех же самых производителей скорее всего так хорошо прогнозировать вероятность отказов уже не будет. Наглядным примером является одна из последних публикаций на эту тему [8]. В ней была рассмотрена всего одна модель ST4000DM000 за период с 2016 по 2019 годы. А в выводах говорится, что из-за появления новых моделей накопителей эффективность разработанной системы прогнозирования может снизиться.

Из-за того, что основные направления исследований в мировой науке обусловлены грантовой поддержкой их финансирования, выявилась и третья особенность. Она заключается в том, что хотя главные закономерности можно извлечь из изучения уже накопленных в мире больших данных, но необходимость соблюдения жестких сроков предоставления отчетности заставляют исследователей сосредоточиться в основном на разработке новых и модификации имеющихся методов анализа только небольшой их совокупности. Примером является еще одна из последних публикаций [9]. В ней опять же рассматривались всего две модели: у компании Baidu это ST31000524NS, у компании Backblaze это снова ST4000DM000.

Также известна группа исследователей, работающая с наиболее полными по производителям, хотя и ограниченными периодом с 2013 по 2019 годы, данными компании Backblaze [10]. Скорее всего такое ограничение вызвано четвертой особенностью – тем, что данных становится слишком много. Именно поэтому другое направление – большие данные по надежности накопителей – пока остается недостаточно изученным. Оно требует или доступа к специальным вычислительным мощностям с высокой производительностью или разработки особой системы прогнозирования. На втором варианте мы и намерены сосредоточиться.

Первые же попытки анализа данных за период с 2013 по 2021 год показали, что требуемое время даже на самые элементарные варианты их обработки на простом компьютере составляет до четырех месяцев [11]. В связи с этим предлагается сначала создать выборку из данных на последнюю известную дату по каждому накопителю и только потом, сгруппировав их по состоянию накопителей (отказавшие, снятые досрочно, функционирующие), исследовать детально связь значений параметров с состоянием по отдельным, наиболее интересным случаям.

На основе этого предложения были проведены предварительные исследования на 20% данных по HDD за период с 2013 по 2016 годы, по результатам которых был разработан матричный метод многопараметрической оперативной оценки и прогнозирования надежности накопителей информации [12]. Перспектива дальнейших исследований проистекает из необходимости раздельного анализа HDD и SSD накопителей.

Обсуждение и выводы

Согласно определению, приведенному в указе Президента РФ от 09.05.2017 № 203 «О Стратегии развития информационного общества в Российской Федерации на 2017-2030 годы», цифровая экономика – это хозяйственная деятельность, в которой ключевым фактором производства являются данные в цифровом виде, обработка больших объемов и использование результатов анализа которых по сравнению с традиционными формами хозяйствования позволяют существенно повысить эффективность различных видов производства, технологий, оборудования, хранения, продажи, доставки товаров и услуг. Сама обработка больших объемов данных – это совокупность подходов, инструментов и методов автоматической обработки структурированной и неструктурированной информации, поступающей из большого количества различных, в том числе разрозненных или слабосвязанных, источников информации, в объемах, которые невозможно обработать вручную за разумное время.

Конкурентным преимуществом на мировом рынке обладают государства, отрасли экономики которых основываются на технологиях анализа больших объемов данных. Такие технологии активно используются в России, но они основаны на зарубежных разработках. Повсеместное внедрение иностранных информационных и коммуникационных технологий, в том числе на объектах критической информационной инфраструктуры, усложняет решение задачи по обеспечению защиты интересов граждан и государства в информационной сфере.

Для предоставления безопасных и технологически независимых программного обеспечения и сервисов необходимо создать российское общесистемное и прикладное программное обеспечение, телекоммуникационное оборудование и пользовательские устройства для широкого использования гражданами, субъектами малого, среднего и крупного предпринимательства, государственными органами и органами местного самоуправления, в том числе на основе обработки больших объемов данных, применения облачных технологий и интернета вещей.

Следовательно, одним из основных направлений развития российских информационных и коммуникационных технологий является обработка больших объемов данных.

Кстати можно отметить, что в предыдущей «Стратегии развития информационного общества в Российской Федерации», утвержденной Президентом РФ 07.02.2008 № Пр-212, про большие данные ничего не говорится.

Однако обнаружившиеся трудности не позволяют полноценно применить для оперативного анализа надежности накопителей информации в крупных data-центрах имеющиеся в наличии инструменты. Отсюда вытекает необходимость создания новых методов работы с подобного рода большими данными, что и предполагается сделать в дальнейшем. Предлагаемый метод включает два аспекта. Первый заключается в выявлении параметров, значимых для оценки и прогнозирования надежности, и построения на их основе проекции конечных значений этих параметров в виде матрицы. Разделение ее на уровни по практически обоснованным критериям позволяет визуализировать оценку текущего состояния накопителей в удобной для оператора форме. Второй аспект исходит из того, что отказу накопителя во многих случаях предшествует резкий скачок значений указанных параметров. Поэтому для прогнозирования лучше брать всего два значения – текущее и предшествующее, чтобы простроить интересующий прогноз. Такой подход также дает возможность исключить необходимость каждый раз обрабатывать весь объем данных для уточнения матрицы. Достаточно только добавлять данные на последний момент измерений.

Заключение

Таким образом, главной причиной отсутствия продолжения роста публикаций на основе данных Backblaze считаем проблему «больших данных», порожденную как увеличением с течением времени числа самих накопителей информации в data-центрах, так и увеличением их емкости, а также числа записываемых параметров состояния, связанного с постоянным изменением состава накопителей вследствие непрерывного появления новых модификаций при их замене. Предлагаемый матричный метод многопараметрической оперативной оценки и прогнозирования надежности, основанный в том числе и на физических процессах деградации накопителей, а не только на статистике, позволяет преодолеть указанные трудности.

Список литературы

1. Diallo M.S., Mokeddem S.A., Braud A., Frey G., Lachiche N. Identifying benchmarks for failure prediction in industry 4.0 // Informatics. 2021. 8(4), 68. https://doi.org/10.3390/informatics8040068

2. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. HDD ranking according to

failure hazard degree in large data centers // Ad Alta: Journal of Interdisciplinary Research. 2019. Vol.9, Is.2. P. 159-162. URL: https://www.webofscience.com/wos/woscc/full-record/WOS:000507312200042

3. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Study of Failure Hazard Degree in Large Data Centers // Helix. 2019. Vol.9, Is.5. P. 5345-5349. URL: http://helix.dnares.in/2019/10/31/loss-of-pressure-in-a-smooth-pipe-with-a-pulsating-turbulent-course/

4. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Reallocated sectors count parameter for analysing hard disk drive reliability // Journal of Computational and Theoretical Nanoscience. 2019. Vol.16, Is.12. P. 5298-5302. URL: https://www.ingentaconnect.com/content/asp/jctn/2019/00000016/00000012/art00063

5. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Spin retry count relation with other HDD parameters // Journal of Computational and Theoretical Nanoscience. 2019. Vol.16, Is.12. P. 5303-5306. URL: https://www.ingentaconnect.com/content/asp/jctn/2019/00000016/00000012/art00064

6. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Reallocated sectors count parameter for analysing HDD reliability // International Journal of Psychosocial Rehabilitation. 2019. Vol. 23, Is. 3. P. 755-765. URL: https://www.psychosocial.com/article/PR190364/9062/

7. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Spin retry count relation with other HDD parameters // International Journal of Psychosocial Rehabilitation. 2019. Vol. 23, Is. 3. P. 766-775. URL: https://www.psychosocial.com/article/PR190365/9064/

8. Ircio J., Lojo A., Lozano J.A., Mori U., Lozano J.A. A Multivariate Time Series Streaming Classifier for Predicting Hard Drive Failures [Application Notes] // IEEE Computational Intelligence Magazine. 2022. Vol. 17. Issue 1. P. 102-114. https://doi.org/10.1109/MCI.2021.3129962

9. De Santo A., Galli A., Gravina M., Moscato V., Sperli G. Deep Learning for HDD Health Assessment: An Application Based on LSTM // IEEE Transactions on Computers. 2022. Vol. 71. Issue 1. P. 69-80. https://doi.org/10.1109/TC.2020.3042053

10. Tomer V., Sharma V., Gupta S., Singh D.P. Hard disk drive failure prediction using SMART attribute // Materials Today: Proceedings. 2021. Vol. 46. Part 20. P. 11258-11262. https://doi.org/10.1016/j.matpr.2021.03.229

11. Насыров И.Н., Насыров И.И., Насыров Р.И. Прикладные проблемы обеспечения эффективности хранения информации в data-центрах // Социально-экономические и технические системы: исследование, проектирование, оптимизация. 2022. № 1 (90). С. 67-76. URL: https://kpfu.ru//staff_files/F1651550418/SETS._1_90_.2022_67_76.pdf

12. Насыров И.Н., Насыров И.И., Насыров Р.И. Метод многопараметрической оценки надежности жестких дисков // Приборы. 2021. № 2. С. 13-19. URL: https://kpfu.ru//staff_files/F24737354/Metod_mnogoparametricheskoj_ocenki_nadezhnosti_zhestkikh_diskov.pdf

References in Cyrillics

1. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I. (2022) Prikladnye problemy obespecheniya effektivnosti hraneniya informacii v data-centrah // Social'no-ekonomicheskie i tekhnicheskie sistemy: issledovanie, proektirovanie, optimizaciya. 2022;(1);67-76. https://kpfu.ru//staff_files/F1651550418/SETS._1_90_.2022_67_76.pdf

2. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I. Metod mnogoparametricheskoj ocenki nadezhnosti zhestkih diskov // Pribory. 2021;(2);13-19. https://kpfu.ru//staff_files/F24737354/Metod_mnogoparametricheskoj_ocenki_nadezhnosti_zhestkikh_diskov.pdf

JEL classification:

C53 Методы прогнозирования. Методы моделирования,

C55 Большие объемы данных: моделирование и анализ,

L86 Информационные и интернет-услуги. Компьютерные программы