УДК 004.33

Насыров И.Н., профессор, доктор экономических наук, доцент, Набережночелнинский институт ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Насыров И.И., доцент, кандидат технических наук, Набережночелнинский институт ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Насыров Р.И., старший преподаватель, Набережночелнинский институт ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

ПРИКЛАДНЫЕ ПРОБЛЕМЫ ОБЕСПЕЧЕНИЯ ЭФФЕКТИВНОСТИ ХРАНЕНИЯ ИНФОРМАЦИИ В DATA-ЦЕНТРАХ

Аннотация: Экспоненциальное увеличение объема генерируемых в мире данных обуславливает необходимость расширения существующих и появления новых распределенных и централизованных (data-центров) систем хранения информации. При этом из-за постоянной замены вышедших из строя накопителей информации на новые, с улучшенными характеристиками, их состав с течением времени в этих системах неизбежно становится гетерогенным. Кроме этого разные производители вкладывают неодинаковый смысл в параметры надежности накопителей, хотя и имеющие одинаковые наименования. В связи с этим является актуальным решение проблемы неоднозначности и неполноты значений параметров надежности гетерогенных наборов накопителей информации в data-центрах. Научная новизна исследований заключается в разработке эффективных способов для выявления и устранения всех реальных ошибок в значениях параметров надежности, отличающихся учетом их неоднозначности и изменчивости во времени, что позволяет повысить качество данных и сформулировать модели данных для гетерогенных наборов накопителей информации в data-центрах.

Ключевые слова: жесткий диск, накопитель, информация, надежность, параметр, эффективность, data-центр.

Введение

Расширение существующих и появление новых распределенных и централизованных (data-центров) систем хранения информации обусловлено экспоненциальным увеличением объема генерируемых данных. При этом из-за постоянной замены вышедших из строя накопителей информации на новые, с улучшенными характеристиками, их состав с течением времени неизбежно

становится гетерогенным. Кроме этого разные производители вкладывают неодинаковый смысл в SMART-параметры (self-monitoring, analysis and reporting technology – технология самоконтроля, анализа и отчетности) надежности накопителей, хотя и имеющие одинаковые названия. Более того, даже у одних и тех же производителей разные марки накопителей имеют как различный набор самих параметров, так и различное смысловое содержание этих параметров. В связи с этим возникает проблема неоднозначности и неполноты значений параметров надежности гетерогенных наборов накопителей информации в data-центрах, на решение которой направлено данное исследование.

В соответствии с общемировой тенденцией повсеместного роста числа систем хранения информации, в разных регионах нашей страны также разворачивается сеть data-центров. Они будут и уже используются как для собственных нужд различных отраслей народного хозяйства России, так и для внешних коммерческих заказчиков. В частности, создается data-центр в городе Иннополис (Республика Татарстан) как референтный прототип среднеразмерного data-центра регионального масштаба для внутренних и зарубежных проектов. Планируется обкатанную здесь технологию тиражировать на проекты в ряде стран, откуда уже поступили заявки. Исходя из этого актуальность исследования проблемы неоднозначности и неполноты значений параметров надежности гетерогенных наборов накопителей информации в data-центрах весьма высока. Научная значимость решения этой проблемы состоит в появлении научного фундамента для разработки метода оценки и прогнозирования надежности накопителей информации как наиболее ответственной части указанной критической информационной инфраструктуры.

Предлагаемые подходы и методы исследования

Конкретными задачами в рамках обозначенной проблемы являются:

1) выявление разницы в структуре параметров надежности накопителей информации, собранных в различные временные периоды;

2) приведение параметров надежности накопителей информации к

одинаковой структуре и интеграция данных в единую базу;

3) анализ параметров надежности накопителей информации на неполноту и ошибочность значений;

4) повышение качества данных за счет устранения выявленных ошибок в значениях параметров;

5) анализ неоднозначности параметров надежности вследствие гетерогенности накопителей информации;

6) формулирование моделей данных для гетерогенных наборов накопителей информации.

Значительный масштаб задач определяется средними (региональными) и крупными (страновыми) размерами рассматриваемых data-центров. Комплексность каждой из задач заключается в необходимости разработки и использования сразу нескольких способов анализа параметров и вариантов моделей данных.

Для решения указанных задач предлагается использовать следующие методы: кластеризации, группировки, структуризации.

Информационной базой исследования являются опубликованные в открытом доступе ежедневные записи SMART-данных более чем 100 тысяч накопителей информации data-центров одной из крупнейших в мире компаний Backblaze (https://www.backblaze.com/b2/hard-drive-test-data.html) за период с 01.04.2013 по 31.12.2021, обеспечивающих хранение свыше одного зеттабайта (10 в 21 степени байт) собираемых со всей планеты сведений. Использование исследователями по всему миру одной и той же базы данных позволяет адекватно сравнивать полученные ими выводы, что является несомненным достоинством с точки зрения доказательства научной новизны, практической применимости и глобальности результатов исследования в условиях цифровой экономики.

Основным инструментом для загрузки, обработки и анализа больших данных, моделирования, оформления и выдачи результатов исследования в числовой и графической форме является лицензионная версия матричной

лаборатории MATLAB.

Результаты исследования

Идея записывать SMART-параметры для анализа их значений пришла к разработчикам и эксплуатантам data-центров не сразу. Поэтому в первые годы функционирования таких систем хранения данных регулярность записи и набор записываемых параметров были низкие. Затем по мере использования сохраненных значений в целях оценки и прогнозирования надежности накопителей информации на жестких дисках (HDD - hard disk drive) запись стали вести ежедневно. Количество записываемых параметров также с течением времени увеличивалось. Поэтому в первую очередь необходимо принять во внимание указанную разницу в структуре параметров надежности накопителей информации, собранных в различные временные периоды. В таблице приведены конкретные даты добавления и номера добавленных параметров. Соответствие наименований параметров и их номеров есть в Википедии https://en.wikipedia.org/wiki/S.M.A.R.T.#ATA_S.M.A.R.T._attributes).

Таблица. Динамика изменения структуры параметров

Дата добавления	Всего параметров	Номера добавленных параметры
10.04.2013	40	1-5, 7-13, 15, 183, 184, 187-201, 223, 225, 240-242, 250-252, 254, 255
01.01.2015	45	22, 220, 222, 224, 226
01.01.2018	50	177, 179, 181, 182, 235
01.04.2018	52	23, 24
01.10.2018	62	16, 17, 168, 170, 173, 174, 218, 231-233
01.10.2019	63	18
01.10.2020	72	175, 180, 202, 206, 210, 234, 245, 247, 248
01.07.2021	82	160, 161, 163-167, 169, 176, 178
01.10.2021	87	171, 172, 230, 244, 246

Выяснилось, что количество данных оказалось очень большим, всего 287 145 360 строк, в связи с чем их пришлось группировать в отдельные файлы приемлемого размера от 3,5 до 4,5 миллионов строк. Для приведения параметров к одинаковой структуре создавались файлы с такими же числами строк, но с

максимальным числом параметров и с пустыми значениями. Затем в них копировались в нужные позиции столбцы из файлов с исходными данными.

Для первоначальных целей исследования из этих файлов выделялись строки только с последними по времени значениями параметров отдельно по каждой марке накопителя. Тем не менее, даже при таком сжатии (до 30 раз) число строк оставалось очень большим. Поэтому полученные из разных файлов данные снова объединялись и из них повторно выделялись строки с последними по времени значениями параметров. В конечном итоге удавалось осуществить интеграцию данных в единую базу с приемлемым для исследования размером.

Однако даже такая простая предварительная обработка данных занимает настолько большое время (по оценкам – свыше четырех месяцев), что заявленные в работе с третьей по шестой задачи целиком будут выполнены позднее. Здесь же приведены результаты исследования только по части исходных данных за временной отрезок с 01.04.2013 по 31.12.2016.

В последнее время в связи с широким распространением твердотельных накопителей информации (SSD - solid state drive) их параметры с операторских и обслуживающих компьютеров data-центров тоже стали сохраняться. Но хотя наименования в системе SMART у них такие же, как и у жестких дисков, тем не менее их надо выделить в отдельную группу, т.к. физика процессов в них совершенно иная. Поэтому требуется разделить данные по HDD и SSD и привести структуру записываемых параметров надежности накопителей информации в соответствие с каждой группой, а также накапливать данные в отдельные базы, единые по каждой группе.

Анализ параметров надежности накопителей информации на неполноту и ошибочность значений показал, что иногда значения параметров отсутствуют или они все нулевые. Кроме этого случаются ошибки записи данных. Самая частая из них связана с человеческим фактором, когда вышедший из строя накопитель помечен как отказавший, но не заменен вовремя. В связи с чем данные по нему могут не записываться в базу несколько дней.

Для повышения качества данных нужно классифицировать выявленные ошибки и определить частоту их появления. Это значительно облегчит их поиск и исправление. Возможно некоторые процедуры по устранению типичных ошибок удастся автоматизировать. Эффективность будет повышаться за счет уменьшения последствий ошибок и за счет экономии времени персонала на обслуживающие мероприятия.

Анализ неоднозначности параметров надежности вследствие гетерогенности накопителей информации показал, что различные производители жестких дисков вкладывают не всегда одинаковый смысл в параметры с одним и тем же названием. Более того, даже у одного и того же производителя используемые для разных марок дисковых накопителей наборы параметров могут различаться. Получается, что вроде бы по смыслу надо использовать предусмотренный для этого параметр, а его значения у разных производителей отличаются на несколько порядков или имеют разный тип - накапливаемый или текущий.

Понятно, что в таких условиях модели, описывающие зависимости значений параметров от времени эксплуатации, будут различными для разных производителей и, возможно, даже для разных марок накопителей одних и тех же производителей. Предлагается составить классификацию таких моделей данных.

Обсуждение и выводы

Для решения поставленных задач были использованы полученные ранее результаты исследований по способам выбора параметров для оценки и прогнозирования надежности накопителей информации по относительным [1] и абсолютным [2] значениям, по зависимости числа переназначенных секторов от времени эксплуатации жестких магнитных дисков для анализа их надежности [3, 4], по порядку индикации ошибок позиционирования по частоте ошибок поиска и другим параметрам жесткого диска [5], по связи числа попыток раскрутки дисков с другими параметрами [6], по степени

опасности отказов накопителей [7], по подбору критериев и построению модели и алгоритма метода многопараметрической оценки жестких дисков по риску отказа [8, 9], по эффективности применения программы многопараметрической оценки их надежности [10] и налогообложению подобных результатов интеллектуальной деятельности при продаже [11].

Результатами исследования являются подобранная оптимальная структура параметров надежности дисковых накопителей информации, единая база данных значений параметров надежности жестких дисков для отдельного data-центра, установленная неполнота и ошибочность в этих значениях, повышенное качество данных за счет устранения выявленных ошибок, зарегистрированная неоднозначность параметров надежности разных производителей, сформулированные модели данных для гетерогенных наборов накопителей информации.

Научная новизна исследования заключается в разработке способов для выявления и устранения всех реальных ошибок в значениях параметров надежности, отличающихся учетом их неоднозначности и изменчивости во времени, что позволяет повысить качество данных и сформулировать модели данных для гетерогенных наборов накопителей информации в крупных data-центрах.

Заключение

Таким образом, конкретными научными результатами исследования являются:

1) выявленная и зафиксированная разница в структуре параметров надежности накопителей информации за различные временные периоды;

2) приведенные к одинаковой структуре параметры надежности накопителей информации, записанные в единую базу данных;

3) проанализированные на неполноту и ошибочность значений параметры надежности накопителей информации с классификацией ошибок;

4) данные повышенного качества за счет устранения выявленных

ошибок в значениях параметров с оценкой эффективности;

5) проанализированные и сгруппированные неоднозначности параметров надежности, возникшие вследствие гетерогенности накопителей информации;

6) сформулированные и протестированные модели данных для гетерогенных наборов накопителей информации;

Значимость результатов исследования состоит в возможности их применения как непосредственно в data-центрах, так и переноса наработанных знаний, умений, навыков и способов на иные объекты в виде глобально распределенных систем хранения данных, на другие предметы исследования наподобие интеграции данных из всех многочисленных разнородных производственных структур большого масштаба, на выбор методов исследования типа тестирования моделей собираемых данных для разнообразных подразделений крупных промышленно-финансовых организаций с соответствующим формированием исследовательской команды в процессе планирования и реализации проекта.

Список использованных источников

1. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Data mining for information storage reliability assessment by relative values // International Journal of Engineering and Technology (UAE). 2018. Vol.7, Is.4.7 Special Issue 7. P. 204-208. https://www.sciencepubco.com/index.php/ijet/article/view/20545, https://www.elibrary.ru/item.asp?id=38622793

2. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Parameters selection for information storage reliability assessment and prediction by absolute values // Journal of Advanced Research in Dynamical and Control Systems. 2018. Vol.10, Is.2 Special Issue. P. 2248-2254. https://www.jardcs.org/backissues/abstract.php?archiveid=5363, https://www.elibrary.ru/item.asp?id=38621781

3. 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. https://www.ingentaconnect.com/content/asp/jctn/2019/00000016/00000012/art00

063;jsessionid=hm2u6b1m8chy.x-ic-live-03, https://www.elibrary.ru/item.asp?id=43244242

4. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Dependence of reallocated sectors count on HDD power-on time // International Journal of Engineering and Technology (UAE). 2018. Vol.7, Is.4.7 Special Issue 7. P. 200-203. https://www.sciencepubco.com/index.php/ijet/article/view/20544, https://www.elibrary.ru/item.asp?id=38621729

5. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Positioning errors indication by Seek error rate and other HDD parameters // Journal of Advanced Research in Dynamical and Control Systems. 2019. Vol.11, Is.8 Special Issue. P. 1797-1805. https://www.jardcs.org/abstract.php?id=2522, https://www.elibrary.ru/item.asp?id=41649555

6. 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. https://www.ingentaconnect.com/content/asp/jctn/2019/00000016/00000012/art00064, https://www.elibrary.ru/item.asp?id=43243589

7. 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. http://helix.dnares.in/2019/10/31/loss-of-pressure-in-a-smooth-pipe-with-a-pulsating-turbulent-course/

8. Насыров И.Н., Насыров И.И., Насыров Р.И. Метод многопараметрической оценки надежности жестких дисков // Приборы. 2021. № 2. С. 13-19. https://kpfu.ru//staff_files/F24737354/Metod_mnogoparametricheskoj_ocenki_nadezhnosti_zhestkikh_diskov.pdf, https://www.elibrary.ru/item.asp?id=44906823

9. Nasyrov I.N., Nasyrov I.I., Nasyrov R.I., Khairullin B.A. Method for HDD Reliability Multiparametric Assessment // Revista San Gregorio. 2021. Is. 44, Special edition. P. 167-178. https://revista.sangregorio.edu.ec/index.php/REVISTASANGREGORIO/article/view/1607

10. Насыров И.Н., Насыров И.И., Насыров Р.И. Эффективность применения программы многопараметрической оценки надежности накопителей информации в крупных data-центрах // Социально-экономические и технические системы: исследование, проектирование, оптимизация. 2021. № 1 (87). С. 56-65.

https://kpfu.ru//staff_files/F1435289090/_SETS._1_87_.2021_Eff_prim_prog.pdf, https://www.elibrary.ru/item.asp?id=45662207

11. Насыров Р.И., Насыров И.Н. Налогообложение результатов интеллектуальной деятельности при переходе к цифровой экономике // Социально-экономические и технические системы: исследование, проектирование, оптимизация. 2019. № 1 (80). С. 85-92. https://kpfu.ru//staff_files/F478614232/SETS_1_80_2019_85.pdf, https://www.elibrary.ru/item.asp?id=39196182

Nasyrov I.N., professor, doctor of economic Sciences, assistant professor, Naberezhnye Chelny Institute of Kazan (Volga region) Federal University

Nasyrov I.I., assistant professor, candidate of technical Sciences, Naberezhnye Chelny Institute of Kazan (Volga region) Federal University

Nasyrov R.I., senior teacher, Naberezhnye Chelny Institute of Kazan (Volga region) Federal University

APPLIED PROBLEMS OF ENSURING INFORMATION STORAGE DEVICES EFFICIENCY IN DATA CENTERS

Abstract: The exponential increase in data volume generated in the world necessitates the expansion of existing and the emergence of new distributed and centralized (data centers) information storage systems. At the same time, due to the constant replacement of failed data storage devices with new ones with improved characteristics, their composition inevitably becomes heterogeneous over time in these systems. In addition, different manufacturers put different meanings into drives reliability parameters, although they have the same names. In this regard, it is urgent to solve the problem of ambiguity and incompleteness of reliability parameters values of information storage devices heterogeneous sets in data centers. The scientific novelty of the research lies in the development of effective ways to identify and eliminate all real errors in reliability parameters values, differing in their ambiguity and variability over time, which makes it possible to improve data quality and formulate data models for information storages heterogeneous sets in data centers.

Key words: hard disk drive; storage; information; reliability; parameter; efficiency; data-center