ISSN: 2688-9323

Data Representation for Primary Features Identification of Hard Drives

Liyu Zhu

College of Software Engineering, Guangdong university of science and technology, Dongguan, Guangdong 532000, China

zhuliyu6@foxmail.com

Abstract

Hard drive failure is the most important reason for cloud storage data loss. Hard drive failure is a gray failure, and the system's failure detector may not notice potential signs of failure. This paper characterizes the hard drive data for identification to help cloud service providers identify primary features that cause hard drive failures. We first selected the authoritative and extensive dataset and then characterized the SMART features in the dataset. The observation and abstraction of SMART features are the basis of feature engineering.

Keywords

Hard Drives Failure; Smart Feature; Representation.

1. Introduction

To meet the increasing demand for Internet services, cloud service providers provide users with different levels of performance and configuration services, of which cloud storage services are a very essential part. Although many cloud service providers' goal is high service availability (such as 99.99%), failures rarely occur in computing clusters and data centers, but any server failure may lead to user dissatisfaction and irreversible data loss.

Hard drive failure is widespread in large storage systems. Literature [1] statistics that the main cause of server failure is hard drive failure, accounting for about 81.84%, and other components such as memory, RAID, etc., accounting for only about 8%, as shown in Figure 1.

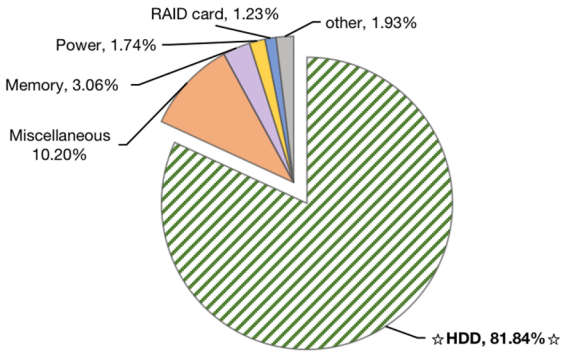

Fig 1. Proportion of various causes leading to server failure

Unexpected downtime in cloud systems is often caused by subtle gray failures [2]. Therefore, hard drive failure is a gray failure. To enhance user confidence in cloud storage resources, cloud service providers need to identify primary features that cause hard drive failures to predict hard drives' failure in advance. The purpose of this paper is to characterize the hard drive data and provide the necessary information for identifying primary features to maintain a highly reliable storage system.

ISSN: 2688-9323

The remainder of the paper is organized as follows. First, we discuss the selection of datasets in Section 2. Then, we characterize the data of the selected dataset in Section 3. followed by the conclusions in Section 4.

2. Dataset Selection

Previous hard drive failure prediction works usually builds models based on small datasets or datasets collected under a unified control environment. For example, the number of hard drives used by Hamerly et al. [3] is 1936, and the number of hard drives used by Hughes et al. [4] is 3744. The quantities are all in the thousands level, and the size of the dataset is one of the bottlenecks in the early improvement of the model performance.

Over time, more and more large datasets become available, including datasets collected from real industrial environments. Pinheiro et al. [5] studied the failure trend of a large number of hard drives deployed in Google systems. The dataset used by Ma et al. [6] included 6 types of 1 million hard drives. The work of Li Jing et al. [7] and Zhu et al. [8] is based on the 23395 Baidu hard drive dataset, in which the number of normal hard drives is 22,962, and the number of failed hard drives is 433. The sampling frequency of hard drive SMART (Self-Monitoring, Analysis, and Reporting Technology) statistics is once an hour. However, this dataset only contains one hard drive model ST31000524NS of Seagate hard drive manufacturer.

Data service provider Backblaze has released SMART statistics of hard drives in its data center since 2013. This dataset is the largest public hard drive dataset in history [9]. Since 2013, the data has been continuously updated every quarter, including various models of hard drives from 7 hard drive manufacturers such as Seagate, HGST, Toshiba, Western Digital, Hitachi, and Samsung. The extensiveness and authority of hard drives in the data set make it possible to design more complex models, which are very popular in the academic world, such as Kadekodi et al [10]. Therefore, this paper chooses Backblaze's dataset as the research object.

3. Data Representation

The Backblaze dataset has taken "snapshots" of running hard drives every day since 2013, and the sampling frequency of hard drive SMART statistics is once a day. According to the literature [11], the data center recorded 122,507 hard drive logs in 2019 alone, with an annual failure rate of about 1.89%.

"Snapshot" consists of two parts: basic drive information and S.M.A.R.T. statistics. Among them, the basic drive information includes five parts, namely "Date", "Serial Number", "Model", "Capacity Bytes" and "Failure". If the value of the "Failure" field is "0", the drive is healthy; if the value of the "Failure" field is "1", the value of "Date" is the last day of operation before the drive fails. Each SMART attribute in S.M.A.R.T. statistical data is composed of its raw value and normalized value, which are collectively referred to as SMART features in this paper.

The two-dimensional structure of SMART feature of a hard drive in this dataset is shown in Table 1. Among them, SMART feature $s_m = \{s_1, s_2, ..., s_n\}, m \in [1, n]$, time $t_i = \{t_1, t_2, ..., t_j\}, i \in [1, j]$, x_{im} represents the feature value of SMART feature s_m at time t_i.

ISSN: 2688-9323

Table 1. Two-dimensional representation of SMART features

		S_1	S_2	S_3	S_4	⋯	S_m	⋯	S_n	
t_1		x_{11}	x_{12}	x_{13}	x_{14}	⋯	x_{1m}	⋯	x_{1n}	
t_2		x_{21}	x_{22}	x_{23}	x_{24}	⋯	x_{2m}	⋯	x_{2n}	
⋮		⋮	⋮	⋮	⋮		⋮		⋮	
t_i		x_{i1}	x_{i2}	x_{i3}	x_{i4}	⋯	x_{im}	⋯	x_{in}	
⋮		⋮	⋮	⋮	⋮		⋮		⋮	
t_j		x_{j1}	x_{j2}	x_{j3}	x_{j4}	⋯	x_{jm}	⋯	x_{jn}	

Up to now, the Backblaze data center has collected 124 columns of SMART feature data. However, not all SMART features can effectively represent the failure characteristics of hard drives, and too many features will reduce the weight of primary failure features and affect the accuracy of the prediction model.

4. Conclusion

In this paper, We selected the authoritative and extensive dataset and then characterized the SMART features in the dataset. Misuse of SMART feature data and ignoring the information in the process of hard drive degradation will affect the quality of the prediction model. Therefore, we can continue to explore the characteristics of the nature of hard drive failure, find out the primary features that cause the failure phenomenon and characterize it as a calculable object.

References

[1] Wang G, Zhang L, Xu W. What Can We Learn from Four Years of Data Center Hardware Failures? [C]. In Proceedings of the 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 2017: 25-36.

[2] Huang P, Guo C, Zhou L, et al. Gray Failure: The Achilles' Heel of Cloud-Scale Systems[C]. In Proceedings of the 16th Workshop on Hot Topics in Operating Systems. ACM, 2017: 150-155.

[3] Hamerly G, Elkan C. Bayesian approaches to failure prediction for disk drives[C]. In ICML. 2001, Vol. 1: 202-209.

[4] Hughes G F, Murray J F, Kreutz-Delgado K, et al. Improved disk-drive failure warnings[J]. IEEE Transactions on Reliability, 2002, 51(3): 350-357.

[5] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. Failure trends in a large disk drive population[C]. In Proceedings of the 5th USENIX Conference on File and Storage Technologies. USENIX Association, 2007: 17–23.

[6] Ao Ma, Rachel Traylor, et al. RAIDShield: Characterizing, monitoring, and proactively protecting against disk failures[C]. ACM Transactions on Storage (TOS) .2015,11(4): 1-28.

[7] Li J, Stones R J, Wang G, et al. Being Accurate Is Not Enough: New Metrics for Disk Failure Prediction[C]. In Proceeding of the 35th IEEE Symposium on Reliable Distributed Systems (SDRS). IEEE, 2016: 71-80.

[8] Zhu B, Wang G, Liu X, et al. Proactive drive failure prediction for large scale storage systems[C]. In Proceeding of the 29th IEEE Symposium on Mass Storage Systems and Technologies (MSST). IEEE, 2013: 1-5.

[9] Backblaze. Downloading the Raw Hard Drive Test Data [EB/OL]. https:// www.backblaze. com/b2/ hard- drive-test-data.html#downloading-the-raw-hard-drive-test-data.

ISSN: 2688-9323

[10] Kadekodi, Saurabh, K. V. Rashmi, and Gregory R. Ganger. Cluster storage systems gotta have HeART: Improving storage efficiency by exploiting disk-reliability heterogeneity[C]. In Proceeding of the 17th Conference on File and Storage Technologies (FAST).USENIX, 2019: 345-358.

[11] Backblaze. Hard Drive Stats, 2019 Snapshot [EB/OL]. https://www.backblaze. com/b2/ hard-drive -test-data.html.