

数据中心硬盘故障预测技术 白皮书

版权声明

ODCC（开放数据中心委员会）发布的各项成果，受《著作权法》保护，编制单位共同享有著作权。

转载、摘编或利用其它方式使用 ODCC 成果中的文字或者观点的，应注明来源：“开放数据中心委员会 ODCC”。

对于未经著作权人书面同意而实施的剽窃、复制、修改、销售、改编、汇编和翻译出版等侵权行为，ODCC 及有关单位将追究其法律责任，感谢各单位的配合与支持。

www.ODCC.org.cn

编写组

项目经理：

原超 中移（苏州）软件技术有限公司

工作组长：

岳上 腾讯科技（深圳）有限公司

贡献专家：

潘飞 中移（苏州）软件技术有限公司

许可欣 中国信息通信研究院

于超琪 超聚变数字技术有限公司

王钦东 超聚变数字技术有限公司

张宇奇 西安三星电子研究所

李娟 浪潮电子信息产业股份有限公司

梁文俊 浪潮电子信息产业股份有限公司

www.ODCC.org.cn

前言

随着互联网、云计算的发展，数据的存储需求与日倍增，大规模海量数据存储中心是必不可少的基础性设施。在当前的数据中心存储系统中，硬盘作为核心的设备，其寿命期是有限的，往往在使用 2 到 3 年后故障率明显升高，是影响服务器运行可靠性的重要因素。以硬盘故障预测技术为主的主动容错技术不仅可以减少数据丢失的风险，还能降低数据恢复的成本，如额外存储开销，网络带宽等，已成为当前数据中心智能运维方面主流的研究方向。

本白皮书从硬盘故障预测的完整技术栈出发，对能力指标、故障数据源的选择、预测算法和模型、案例等几个方面对硬盘故障预测技术能力和架构进行描述，提供对硬盘故障预测技术的全视角剖析。

由于时间仓促，水平所限，错误和不足之处在所难免，欢迎各位读者批评指正。如有意见或建议请联系编写组。

www.ODCC.org.cn

目 录

版权声明.....	I
编写组.....	II
前 言.....	III
一、背景与发展.....	1
(一) 硬盘故障已成为大型数据中心硬件故障主要因素	1
(二) 智能化运维发展推动故障预测应用建设	4
二、硬盘故障预测关键技术和指标.....	6
(一) 业界传统解决磁盘故障方式	6
(二) 评价指标	7
(三) 数据源及预处理	10
1、S.M.A.R.T. 参数.....	10
2、其他数据.....	13
3、数据的预处理.....	20
(四) 预测模型	21
1、传统的阈值方法和统计学算法.....	21
2、基于深度学习算法的预测模型.....	23
三、主流存储故障运维软件案例.....	26
(一) 中移云秤-磁盘健康检测平台	27
(二) 超聚变 FusionDirector -故障诊断功能.....	29
(三) 三星存储健康预测	31
(四) 浪潮融合智能存储管理平台	34
四、总结和展望.....	37

参考文献.....	38
附录 缩略语和名词释义.....	39



www.ODCC.org.cn



数据中心硬盘故障预测技术白皮书

一、背景与发展

(一) 硬盘故障已成为大型数据中心硬件故障主要因素

随着数字中国、企业数字化转型的加速发展，越来越多的多元化数据和服务将部署在数据中心的存储系统中。用户对数据中心的存储可靠性要求也逐步增大，从最初的 99.99% 到 99.999%，甚至到 99.9999%，只有存储系统的可靠性得到保证，数据中心的可靠性和可用性才能得到保证。

在当前的数据中心存储系统中，机械硬盘（Hard Disk Drive, HDD）和固态硬盘（Solid State Disk, SSD）仍然是主流的外存储设备，其生命周期通常为 3 到 5 年，在 2 到 3 年后故障率明显升高，导致换盘量陡增。

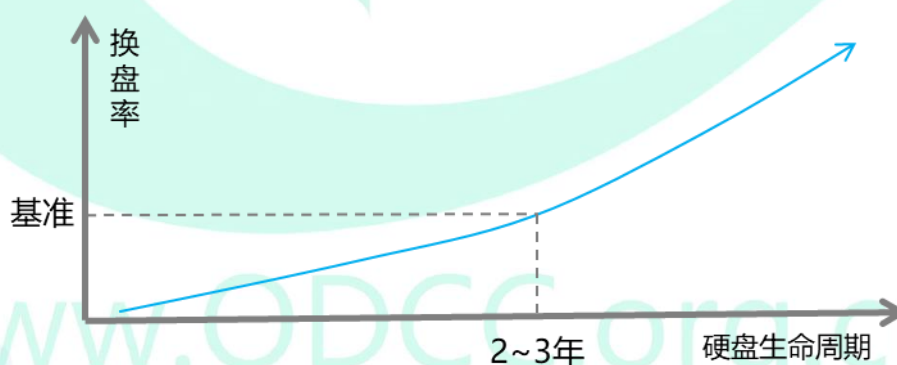


图 1

硬盘由多种机械和电子器件组成，是非常复杂且精密的设备，硬盘的故障呈现出不同的表现形式和严重程度。机械硬盘 HDD 由磁头/盘面接口、磁头堆栈组件、主轴电机/轴承和电子模块四部分单

元组成。这些单元的机制不同，其故障的触发因素也是各不相同。例如，磁头/盘面接口和主轴电机/轴承是机械部件，经常因为磨损而失效；控制板是电子电路，其经常遭受的是短路故障；固态硬盘 SSD 内部使用多通道技术来进行并行读写，控制器在 SSD 寿命方面起重要作用，它决定哪些位置的页在哪些时候进行拷贝或者擦除，并且提供逻辑地址和物理地址之间的映射。

在数据中心的存储容量不断增长，硬盘故障已成为常态而非例外的情况。国内外数据中心由于硬盘故障导致的服务中断屡见不鲜，造成的经济损失不计其数。

依据 Backblaze 在 2023 年第一季度公开报告可以看出，Backblaze 共管理着 236893 块机械硬盘，排除无法启动的、S.M.A.R.T. 信息不全的、数据异常的，累计已有 17155 块硬盘挂掉，它们的平均使用时间只有 2 年 6 个月。机械盘的平均故障率为 1.54%。

www.ODCC.org.cn

Backblaze Hard Drives Annualized Failure Rates for Q1 2023

Reporting period: 1/1/2023 through 3/31/2023 for drive models active as of 3/31/2023

MFG	Model	Drive Size	Drive Count	Avg. Age (months)	Drive Days	Drive Failures	AFR
HGST	HMS5C4040ALE640	4TB	3,722	80.0	334,847	1	0.11%
HGST	HMS5C4040BLE640	4TB	12,728	77.0	1,145,343	12	0.38%
HGST	HUH728080ALE600	8TB	1,116	59.4	99,522	5	1.83%
HGST	HUH728080ALE604	8TB	93	68.3	8,434	1	4.33%
HGST	HUH721212ALE600	12TB	2,606	41.8	232,402	1	0.16%
HGST	HUH721212ALE604	12TB	13,151	24.2	1,178,671	23	0.71%
HGST	HUH721212ALN604	12TB	10,690	47.7	959,203	77	2.93%
Seagate	ST4000DM000	4TB	18,070	88.9	1,633,962	170	3.80%
Seagate	ST6000DX000	6TB	884	95.4	79,651	2	0.92%
Seagate	ST8000DM002	8TB	9,467	77.6	855,217	49	2.09%
Seagate	ST8000NM000A	8TB	96	14.3	7,456	0	0.00%
Seagate	ST8000NM0055	8TB	14,331	65.8	1,296,710	132	3.72%
Seagate	ST10000NM0086	10TB	1,159	63.5	105,443	14	4.85%
Seagate	ST12000NM0007	12TB	1,239	40.6	112,607	23	7.46%
Seagate	ST12000NM0008	12TB	19,749	35.9	1,774,769	120	2.47%
Seagate	ST12000NM001G	12TB	12,805	27.4	1,137,007	26	0.83%
Seagate	ST14000NM001G	14TB	10,760	25.7	962,379	35	1.33%
Seagate	ST14000NM0138	14TB	1,497	27.8	134,818	23	6.23%
Seagate	ST16000NM001G	16TB	21,910	15.6	1,897,163	31	0.60%
Seagate	ST16000NM002J	16TB	310	9.5	27,590	0	0.00%
Toshiba	MD04ABA400V	4TB	94	94.3	8,460	0	0.00%
Toshiba	HDWF180	8TB	60	17.5	2,186	0	0.00%
Toshiba	MG07ACA14TA	14TB	38,156	28.9	3,421,785	109	1.16%
Toshiba	MG07ACA14TEY	14TB	577	22.7	50,665	3	2.16%
Toshiba	MG08ACA16TA	16TB	5,194	10.7	440,772	4	0.33%
Toshiba	MG08ACA16TE	16TB	5,932	17.7	529,316	9	0.62%
Toshiba	MG08ACA16TEY	16TB	5,288	15.8	470,414	1	0.08%
WDC	WUH721414ALE6L4	14TB	8,412	27.7	754,296	10	0.48%
WDC	WUH721816ALE6L0	16TB	2,699	17.8	240,256	2	0.30%
WDC	WUH721816ALE6L4	16TB	14,098	6.3	1,128,989	6	0.19%
Totals			236,893		21,030,333	889	1.54%



图 2 2023 年第一季度硬盘故障报告

根据国内某公有云的省节点资源池上 23 年第一季度硬件故障的统计，损坏率较高的服务器部件为硬盘损坏和内存损坏，两者占服务器整体硬件损坏的 70% 以上。当然，硬盘、内存本身数量在所有部件中占比较高，也是拉高损坏占比的原因之一。

损坏类型	3 月	占比
硬盘损坏	560	61.47%
内存损坏	109	11.96%
CPU 损坏	50	5.49%
主板损坏	32	3.51%
风扇损坏	24	2.63%
网卡损坏	22	2.41%
电源损坏	18	1.98%
光模块损坏	11	1.21%
RAID 卡损坏	9	0.99%
其他损坏	5	0.55%

图 3 国内某公有云 2023 年第一季度硬件故障占比报告

当前数据中心中的大规模存储系统普遍采用副本或纠删码等机制来提供高可靠性，在硬盘发生故障后的进行数据恢复，属于被动容错技术。这种方式采用牺牲较大存储空间和增加资源开销等保障存储系统的可用性。而以硬盘故障预测技术为主的主动容错技术越来越受到研究人员的关注。硬盘故障预测具有容错开销小、处理过程简单且对用户透明等显著优势，其核心思想在于实时监控运行中的硬盘，采集并记录硬盘的运行状态数据，随后通过特定算法检测硬盘状态数据并预测硬盘是否即将发生故障。

成功的硬盘故障预测不仅可以减少数据丢失的风险，还能降低数据恢复的成本，如额外存储开销，网络带宽等。随着深度学习的发展，故障预测的研究更是一个热点。

（二）智能化运维发展推动故障预测应用建设

最早期的硬盘监控技术起源于 1992 年，IBM 在 AS/400 计算机的 IBM 0662 SCSI 2 代硬盘驱动器中使用了后来被命名为

Predictive Failure Analysis（故障预警分析技术）的监控技术，它是通过在固件中测量几个重要的硬盘安全参数和评估他们的情况，然后由监控软件得出两种结果：“硬盘安全”或“不久后会发生故障”。不久，当时的微机制造商康柏和硬盘制造商希捷、昆腾以及康纳共同提出了名为 IntelliSafe 的类似技术。通过该技术，硬盘可以测量自身的健康指标并将参数值传送给操作系统和用户的监控软件中，每个硬盘生产商有权决定哪些指标需要被监控以及设定它们的安全阈值。1995 年，康柏公司将该技术方案提交到 Small Form Factor（SFF）委员会进行标准化，该方案得到 IBM、希捷、昆腾、康纳和西部数据的支持，1996 年 6 月进行了 1.3 版的修正，正式更名为 S.M.A.R.T.（Self-Monitoring Analysis And Reporting Technology），全称就是“自我检测分析与报告技术”，成为一种自动监控硬盘驱动器完好状况和报告潜在问题的技术标准。

随着机器学习和人工智能的发展，智能化运维建设也在不断升级。国务院关于印发《“十四五”数字经济发展规划的通知》中指出，数据的爆发增长为智能化发展带来了新的机遇，提升基础设施网络化、智能化、服务化、协同化水平，加快优化智能化产品和服务运营，也将进一步突出智能化运维对于业务创新的重要意义。近几年现在绝大部分关于硬盘故障预测的研究都或多或少使用了人工智能的技术，包括：传统机器学习方法，如 Random Forest（随机森林）等作为基础模型、神经网络，如 LSTM 等、特殊的神经网络模型，如 GAN 等。

人工智能技术需要大量的数据来训练，故障出现的条件和模式无法由单一因素决定，只有对故障的历史数据通过专家经验分析和标注，不断用模式识别和数据进行迭代训练，才能让训练出来的算法推理更加准确。因此，使用人工智能模型必然会有相应的数据集。目前来说，最常用的、且与硬盘故障强相关的特征属性仍旧 S.M.A.R.T.。当然也不断有研究者和硬盘厂商如希捷等提出新的与硬盘故障相关的其他特征属性，如硬盘故障的时空局域性等等，在下一章中也会加以介绍以供参考。

二、 硬盘故障预测关键技术和指标

（一） 业界传统解决磁盘故障方式

一般来说，数据中心的磁盘错误属于灰色故障，即使已经对生产系统造成了严重影响，依然隐蔽而不易被发现，目前业界共有三种方法尝试解决磁盘故障检测难题：

1、RAID。传统的提高数据可靠性的方法主要是冗余磁盘阵列 RAID 技术，通过采用数据冗余，容忍单个或多个磁盘的故障，并通过数据编码方式恢复错误数据。属于被动容错。

2、S.M.A.R.T.。S.M.A.R.T. 数据描述了磁盘的多种属性，是磁盘管理的标准接口，S.M.A.R.T. 是典型的主动容错技术。该技术监控磁盘运行过程中的多项参数，包括磁盘的寻道错、奇偶校验错等信息。基于 S.M.A.R.T. 的主动容错采用阈值方法，简单易行，但是预警准确率较低。

3、机器学习。智能运维（AIOps）是存储管理的发展趋势，预测性分析是 AIOps 的重要方向。基于大量数据，通过对数据的分析，从中发现蕴含的规律，实现“学习”的目的，从而提高系统本身的能力。在磁盘 S.M.A.R.T. 等数据的基础上，结合系统 IO 错误信息，采用机器学习方法，实现对磁盘故障的预测，可达到及时对故障磁盘进行处理，并增强存储系统可靠性的目的。AI 使能磁盘故障检测，是存储预测性分析中的关键技术和创新突破点。



图 4

本白皮书主要介绍基于机器学习的智能预测技术框架，包括以硬盘故障数据源、数据的特征处理工程、预测模型方法论等，在此之前，首先需要明确硬盘故障智能预测系统的评价指标。

（二）评价指标

在智能预测系统中，由于不存在一个适用于所有情况的评价指标，所以评估预测精度（或误差）就成为了较为不易的事情。一般来说只有通过试验，才能知道哪个性能评估指标适用于当前情况。对于数据中心硬盘故障的预测能力，基本上是基于数据驱动的方法进行的，所以评价指标基本上使用的是机器学习中模型评价的指标：

1、检出率 FDR (Failure Detection Rates)。FDR 又被称为故障检测率、召回率，表示所有的故障盘中被检测出来的比例；

2、误报率 FAR (False Alarm Rate)。FAR 又被称为虚警率、误检率，表示所有的正常盘中被预测为故障盘的比例；

3、漏报率 MAR (Missing Alarm Rate)。MAR 又被称为漏检率、漏警率，表示故障盘被预测为正常盘的比例。漏报率 MAR 和检出率 FDR 是具备关联性的，这两项指标通常只需考虑其中一项。

当然，结合硬盘故障预测的实际问题，对于具体的预测程序也有一些其他要求，如耗时要求、预警提前量（时间）指标等：

- 1、提前时间：故障预警和故障发生之间的时间间隔
- 2、执行时间：机器学习算法或者其他算法进行一次预测的时间。

硬盘的预测其他指标如 MR (Migration Rate) 和 MMR (Mismigration Rate)，分别代表了被保护的数据的比例和不必要的被保护的数据比例。这两个指标也可用来很好的衡量存储系统的可用性，尤其是在硬盘数据预测模型需要进行样本数据平衡、小样本迁移，耗费时间和数据中心系统资源的情况下，可以借助这两个指标比较预测磁盘剩余寿命模型的质量。

在当前已商用的硬盘故障预测系统中，也存在一些指标的名称不同但是实际意义相同的情况。如检出率 FDR，可能在某些系统中

用召回率 Recall 取代，以误报率 FAR，也可能用 False-PositiveRate 取代，但实际意义和计算公式都是一致的。

下表为目前业界普遍遵循的指标解释可供参考。由于某些指标值间具备一定的关联性，因此评价预测的能力并不需要囊括如下所有指标，只需选择其中 1-3 项作为参考项即可。

表 1

名称	计算公式	描述
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	Precision 表示预测为故障的盘中有多少是真的故障盘。
Recall, True-Positive Rate, Failure Detection Rate	$\text{Recall} = \frac{TP}{TP + FN}$	所有的正例中被判断为正例的比例。即，所有的故障盘中被检测出来的比例。
False-Positive Rate, False Alarm Rate	$\text{FPR} = \frac{FP}{FP + TN}$	所有的好盘中被判断为坏盘的比例。
False-Negative Rate	$\text{FNR} = \frac{FN}{TP + FN}$	所有的坏盘被判断为好盘的概率
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	所有预测正确的比例
Migration Rate(MR)	被迁移的数据比例	度量在正确的故障预测的情况下的有多少危险数据被保护起来了。
Mismigration Rate(MMR)	不必要的数据迁移的比例	度量在错误的故障预测的情况下有多少不必要的数据被迁移
Alarm Distance, Lead Time	报警点和真实的故障发生点之间的距离	这个指标越高，就越可以保证用户有足够的时间去备份数据。
BA(balanced accuracy)	$\text{BA} = \frac{P(s^- = F(\vec{s}) s^-) + P(s^+ = F(\vec{s}) s^+)}{2}$	成功检测出正例和成功检测出反例的概率的平均值

其中，TP、FP、FN、TN 解释如下：

TP (True Positive)：预测结果为故障，实际上就是故障

FP (False Positive)：预测结果为故障，实际上是正常

FN (False negative)：预测结果为正常，实际上是故障

TN (True negative)：预测结果为正常，实际上就是正常

表 2

		预测结果	
		P/故障	N/健康
真实结果	P/故障	TP	FN
	N/健康	FP	TN

针对上述情况，这里我们举一个简单的例子来加以说明：

表 3

真实情况			预测准确		预测错误	
故障	FALSE	14	TP	12	FN	2
正常	TRUE	574	TN	572	FP	2

检出率 $FDR=Recall=85.71\%$

误报率 $FAR=False-PositiveRate (FPR)=0.35\%$

漏报率 $MAR=False-NegativeRate (FNR)=14.29\%$

精确率 $Precision=85.71\%$

准确率 $Accuracy=99.32\%$

(三) 数据源及预处理

1、S.M.A.R.T. 参数

S.M.A.R.T. 数据是目前最常用的、与硬盘故障相关性最高的特征属性数据，其全称为“Self-Monitoring Analysis and Reporting Technology”，即“自我监测、分析及报告技术”。正

如前文中所述，S.M.A.R.T. 是由康柏公司发起，由 IBM、希捷、昆腾、康纳和西部数据共同支持的自动监控硬盘驱动器完好状况和报告潜在问题的技术标准。作为行业规范，S.M.A.R.T. 规定了硬盘制造厂商应遵循的标准。

S.M.A.R.T. 功能不断从硬盘上的各个传感器收集信息，并把信息保存在硬盘的系统保留区（service area）内，这个区域一般位于硬盘 0 物理面的最前面几十个物理磁道，由厂商写入相关的内部管理程序。这里除了 S.M.A.R.T. 信息表外还包括低级格式化程序、加密解密程序、自监控程序、自动修复程序等。用户使用的监测软件通过名为“S.M.A.R.T. Return Status”的命令（命令代码为：B0h）对 S.M.A.R.T. 信息进行读取，且不允许最终用户对信息进行修改。

S.M.A.R.T. 数据可以是硬盘在使用过程中所生成的数据，如通电时间累计，即硬盘出厂后通电的总时间，各硬盘性能不同其预设通电时间阈值也不同，当通电总时间越接近预设通电时间阈值甚至高于预设通电时间阈值时，则未来一段时间硬盘大概率存在故障；还可以是硬盘运行时的温度数据、底层数据读取错误率、磁盘通电次数等。

S.M.A.R.T. 数据对每一项目都有相应的不同属性值，如下图所示，一般来说 S.M.A.R.T. 数据包括 ID、属性名（ATTRIBUTE_NAME）、

当前值 (Value)、Worst (最差值)、临界值 (Threshold)、原始值 (RAW-VALUE)、FLAG、TYPE、UPDATED、WHEN_FAILED 等。

SMART Attributes Data Structure revision number: 10
Vendor Specific SMART Attributes with Thresholds:

ID#	ATTRIBUTE_NAME	FLAG	VALUE	WORST	THRESH	TYPE	UPDATED	WHEN_FAILED	RAW-VALUE
1	Raw_Read_Error_Rate	0x000f	077	064	044	Pre-fail	Always	-	146936992
3	Spin_Up_Time	0x0003	090	089	000	Pre-fail	Always	-	0
4	Start_Stop_Count	0x0032	100	100	020	Old_age	Always	-	13
5	Reallocated_Sector_Ct	0x0033	014	014	010	Pre-fail	Always	-	13248 (0, 5)
7	Seek_Error_Rate	0x000f	084	060	045	Pre-fail	Always	-	274033350
9	Power_On_Hours	0x0032	096	096	000	Old_age	Always	-	3625
10	Spin_Retry_Count	0x0013	100	100	097	Pre-fail	Always	-	0
12	Power_Cycle_Count	0x0032	100	100	020	Old_age	Always	-	13
187	Reported_Uncorrect	0x0032	001	001	000	Old_age	Always	-	555
188	Command_Timeout	0x0032	100	100	000	Old_age	Always	-	0
190	Airflow_Temperature_Cel	0x0022	054	050	040	Old_age	Always	-	46 (Min/Max 39/47)
192	Power-Off_Retract_Count	0x0032	100	100	000	Old_age	Always	-	47
193	Load_Cycle_Count	0x0032	100	100	000	Old_age	Always	-	178
194	Temperature_Celsius	0x0022	046	050	000	Old_age	Always	-	46 (0 10 0 0 0)
195	Hardware_ECC_Recovered	0x001a	004	001	000	Old_age	Always	-	146936992
197	Current_Pending_Sector	0x0012	100	100	000	Old_age	Always	-	200
198	Offline_Uncorrectable	0x0010	100	100	000	Old_age	Offline	-	200
199	UDMA_CRC_Error_Count	0x003e	200	200	000	Old_age	Always	-	0
200	Multi_Zone_Error_Rate	0x0023	100	100	001	Pre-fail	Always	-	0
240	Head_Flying_Hours	0x0000	100	253	000	Old_age	Offline	-	27741193768064
241	Total_LBAs_Written	0x0000	100	253	000	Old_age	Offline	-	31278699856
242	Total_LBAs_Read	0x0000	100	253	000	Old_age	Offline	-	46599094092

图 5

这些属性值的介绍可以在各公开报告和技术研讨论坛中查询到，这里做针对重要属性做一个简单的整理和介绍：

1) 属性 ID，通常是一个 1 到 255 之间的十进制或十六进制的数字。目前，各硬盘制造商的绝大部分 S.M.A.R.T. ID 代码所代表的参数含义是一致的，但厂商也可以根据需要使用不同的 ID 代码，或者根据检测项目的多少增减 ID 代码。

2) ATTRIBUTE_NAME 硬盘制造商定义的属性名，是 ID 代码的文字解释。

3) 当前值 (Normalized value)：ID 项在硬盘运行时根据实测原始数据 (Raw value) 通过公式计算的结果，计算公式由硬盘厂商自定。新硬盘刚开始使用时显示的当前值可以认为是预设的最大正常值 (有些 ID 项如温度等除外)，随着硬盘的

使用该值会不断刷新并改变。当前值接近临界值可能意味着硬盘寿命的减少，发生故障的可能性增大。

4) 最差值 (Worst) 是硬盘运行时各 ID 项曾出现过的最差值，是硬盘运行中某项数据变劣的峰值统计。通常最差值与当前值是相等的，如果最差值出现较大的波动 (小于当前值)，表明硬盘曾出现错误或曾经历过恶劣的工作环境。

5) 临界值 (Threshold)，是硬盘厂商指定的表示某一项可靠性的门限值，通过特定公式计算而得。如果当前值接近了临界值，表示该项值变得不可靠，可能会导致硬盘故障。

6) 原始值 (RAW-VALUE) 是由制造商定义的原始值，从 VALUE 派生。

2、其他数据

S.M.A.R.T. 体系自投入使用已有 25 年时间，可涵盖 10 多项的参数。可以说 S.M.A.R.T. 是当前覆盖范围最广、使用时间最长、成熟度最高的硬盘健康度标准。但随着云计算的快速发展，超大规模数据中心对于硬盘故障预测能力的要求越来越高，而 S.M.A.R.T. 体系较为陈旧、只涵盖有限的信息、不够灵活、占用系统资源较大，难以适应不同应用场景对于硬盘健康管理的个性化需求。为应对如上挑战，三星以及希捷等厂商各自开发了新的日志数据系统，并逐步得到推广。

(1) Telemetry 数据

三星公司率先在固态硬盘上引入了更加丰富的 Telemetry 属性来实现对存储装置更加细粒度的监控，以便更好的分析和识别固态硬盘的故障。

Telemetry 是 NVMe spec 1.3 版本中引入的可选特性，这个特性可以使 SSD 厂商能够收集内部数据日志，以改善产品的功能和可靠性。目前 Telemetry 定义了数据收集的机制，但具体的数据日志仍然是由各个厂商自己定义的，不同的厂商支持的数据日志的数量也是不一样的。以三星 PM9A3 为例，支持的 Telemetry 属性可以达到 90 多个。相比于 S.M.A.R.T.，Telemetry 的内容更全面，更细粒度：

- 1) 多达 90 多个属性，对于 SSD 的各个硬件组件，温度，以及设备 IO 情况都有记录。
- 2) 对于 NAND 闪存，DRAM，SRAM 等都有详细记录，对于 IO、性能、WAF 等也记录的更加详细。

表 4 S.M.A.R.T. 与 Telemetry 对比

日志类型	数量	粒度	目的
S.M.A.R.T.	22	粗粒度 设备级	硬盘状态监控
Telemetry	93	细粒度 组件级	SSD 状态监控 问题/故障分析 故障预警/预测

Telemetry 属性通过多个维度上对 SSD 内部状态进行监控来全面了解 SSD 内部组件的健康状况，主要分为读写相关、块相关、温度相关、错误相关、RAM 相关、PCI 相关、电源相关等 8 个部分。通

过引入 Telemetry，我们可以预测和分析更多的故障信息，进一步提升故障盘的预测准确度，也能大大提高运营商数据中心云存储系统的智能运维效率。

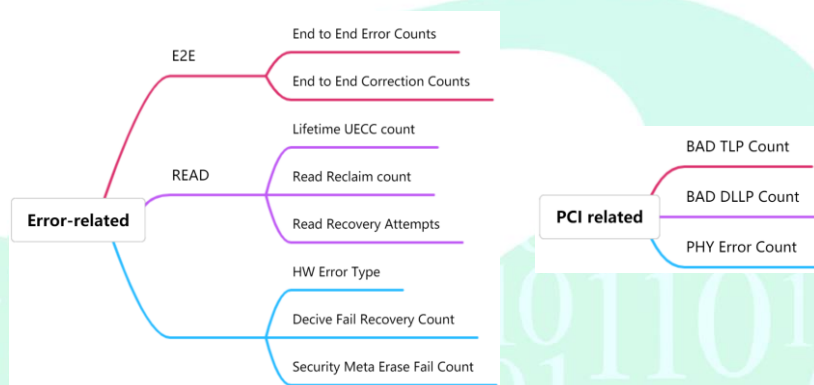


图 6 Telemetry 属性示例

在真实的 Telemetry 数据采集，由于数据采集周期不规律，SSD 的 Telemetry 日志在采样间隔和条数上存在很大的差异，因此在进行数据分析和模型训练之前需要对数据进行预处理，主要包括去除唯一值属性、处理缺失值和重采样：

- 1) 去除唯一值属性：具有唯一值的属性并不能刻画数据的分布规律，所以删除这些属性。
- 2) 缺失值处理：将属性分为连续型特征和离散型特征两类。连续型特征指的是该特征的数据类型是连续的，比如温度、读写相关等属性，这类特征采用线性插值来填充缺失值。离散型特征指的是该特征的数据类型是离散的，比如状态、类型等属性，这类特征采用就近插补，替换为缺失之前的最后一次观测值。最后删除无法填充缺失值的日志。

数据重采样：在时间序列中，数据的一致性是很重要的，重采样可以确保数据以一致的频率分布，根据真实数据的采样周期选择以天为频率对数据进行重采样，方便后续特征提取和模型训练。

（2）FARM 数据

Field Accessibility Reliability Metrics（简称 FARM）是由机械硬盘巨头希捷开发的硬盘故障数据体系，是对 S.M.A.R.T. 体系的重大改良。FARM 本身获取方便，参数覆盖广，可以说是超级加强版的 S.M.A.R.T.。FARM 的 120 多项参数包含了硬盘更加底层的传感器和磁头级别的相关参数，以及飞行高度、工作负载、环境等使用状况的参数，并且已经在多个希捷产品上实施，给硬盘做详细的“全身体检”。同时，FARM 是对 S.M.A.R.T. 体系的重大改良。

FARM 日志按照分页（page）的方式组织，每个分页包含 32 个 512 Bytes 大小的块，按数据类别分为 6 个分页，整个日志大小为 96 kB。

SATA FARM Pro 日志地址为 0xA6，其结构如表所示：

表 5 SATA FARM Log Structure

Page	Description
0	FARM Header-参见表 6
1	General Drive Information
2	Workload Statistics
3	Error Statistics
4	Environmental Statistics
5	Reliability Statistics

表 6 SATA FARM Header Structure

Byte Offset	Data Type	Description
0..7	Qword	Log Signature = 0x00004641524D4552
8..15	Qword	Log Major Revision
16..23	Qword	Log Minor Revision
24..31	Qword	Number of Pages Supported
32..39	Qword	Log Size in Bytes
40..47	Qword	Page Size in Bytes
48..55	Qword	Maximum Drive Heads Supported
56..63	Qword	Number of Historical Copies
64..16383	Qword	Reserved

表 7 SATA FARM Pages 1-5 Structure

Byte Offset	Data Type	Description
0..7	Qword	Log Page Number
8..15	Qword	Log Copy Number
16..23	Qword	Field 1
24..31	Qword	Field 2
...	Qword	...
N..16383	Qword	Reserved

表 8 SAS FARM Sub Pages Structure

Bit Byte	7	6	5	4	3	2	1	0
0	DS	SPF	Page Code (0x3D)					
1	Subpage code (0x03 or 0x04)							
2	Page Length (n-3)							
3								
	FARM Log Page log parameters							
4	FARM Log Page log parameter [First]							
...								
...	FARM Log Page log parameter [Last]							
N								

表 9 SAS FARM Log Parameter Definitions

Parameter Code	Description
0x0000	FARM Header Parameter
0x0001	General Drive Information Parameter
0x0002	Workload Statistics Parameter
0x0003	Error Statistics Parameter
0x0004	Environmental Statistics Parameter
0x0005	Reliability Statistics Parameter
0x0006	General Drive Information Parameter Continued
0x0007	Environmental Statistics Parameter Continued
0x0008-0x000F	Reserved for future statistics
0x0010-0x008F	Reliability by Head Statistics

以下简要介绍各个分页所包含的内容：

General Drive Information: 记录如 SN、WWN、容量，支持的 feature 等基本信息，除此之外，还包含磁头数量、马达等零部件的运行时间、最近一次记录到的硬盘状态等。

Workload Statistics: 记录读写负载数据，除读写 LBA 数量之外，还包含按类别统计的读写命令数量（总数，随机命令数，非读写命令数），以及最近几个小时内读写命令在磁碟不同区域的数量等。

Error Statistics: 记录硬盘错误处理相关数据，除 S.M.A.R.T. 已经包含的错误统计之外，还包含固件内部异常事件、读写重试、机构部件重试等。同时对于不可恢复的错误，按照读写分别统计。

Environmental Statistics: 记录外界环境相关参数，除温度之外，还包含湿度，5V/12V 输入电压，马达电压等。

Reliability Statistics: 记录硬盘可靠性相关的参数, 以及部分诊断功能的记录。包含周期性及空闲时的后台评估、IDD (In Drive Diagnostic, 如果使用过此功能)、偏心率、以及磁头级别的底层参数 (误码率、信道补偿、寻道错误率、磁阻、飞高等) 等。

FARM 日志的获取和解析可以通过 READ LOG (DMA) EXT (SATA Command) 或者 LOG SENSE/LOG SELECT (SAS Command) 来获取 FARM 日志。同时希捷也提供了预编译的日志获取及解析工具, 支持常见的操作系统, 减少客户的二次开发负担。

下图展示了如通过 SeaDragon_LogsUtils 和 SeaDragon_LogParser_FARM 来获取和解析 FARM 日志。

```
root@localhost:~# ./SeaDragon_LogsUtil -d /dev/sg0 --farm
=====
SeaDragon_LogsUtil - Seagate drive utilities - NVMe Enabled
Copyright (c) 2014-2020 Seagate Technology LLC and/or its Affiliates, All Rights Reserved
SeaDragon_LogsUtil Version: 4.13.0-1_21_30 X86_64
Build Date: Feb 19 2020
Today: Mon Jul 6 10:33:45 2020
SEAGATE PRIVATE UTILITY - FOR YOUR EYES ONLY - DO NOT DISTRIBUTE
=====

/dev/sg0 - ST8000NM0055-1RM112 - ZA1056YN - ATA
..
Successfully pulled FARM log
```

图 7 获取 FARM 日志

```

root@localhost:~# ll
total 2844
drwx--x--x+ 1 root    root      236 Jul  6 10:36 .
drwx--x--x+ 1 root    root      132 Jul  6 10:31 ..
-rwx--x--x+ 1 root    root 1470792 Apr 29 23:20 SeaDragon_LogParser_FARM
-rwx--x--x+ 1 root    root 1172560 Feb 21 21:36 SeaDragon_LogsUtil
-rwx--x--x+ 1 root    root  98304 Jul  6 10:33 ZA1056YN_FARM_2020-07-06__10_33_45.bin #上一步获取
到的 FARM 二进制日志
root@localhost:~# ./SeaDragon_LogParser_FARM \
--inputLog ./ZA1056YN_FARM_2020-07-06__10_33_45.bin \ #需要解析的二进制 FARM 日志路径
--logType farmLog \ #日志类型
--printType flatcsv \ #解析输出格式, 这里选择 flatcsv, 也可以使用 json 格式
--outputLog ./ZA1056YN_FARM_2020-07-06__10_33_45.csv #解析后的输出文件路径
=====
SeaDragon_LogParser - Seagate drive utilities
Copyright (c) 2018-2020 Seagate Technology LLC and/or its Affiliates, All Rights Reserved
SeaDragon_LogParser Version: 1.2.3-2.2.4 X86_64
Build Date: Apr 16 2020
Today: Mon Jul  6 10:36:51 2020
SEAGATE PRIVATE UTILITY - FOR YOUR EYES ONLY - DO NOT DISTRIBUTE
=====
Parsing completed with no issues

```

图 8 解析 FARM 日志

按照 flflatcsv 解析后的 FARM 日志内容（前 10 列），可以看到 flflatcsv 格式输出的是一行表头一行数据的结构化数据，方便后续的 ETL，分析等工作。

3、数据的预处理

在获取硬盘故障相关的数据后，可以针对这些数据进行预处理，以提高模型的效果，例如，对 S.M.A.R.T. 数据进行数据清理的预处理，具体可以是处理脏数据，包括填写缺失值、清除噪声数据（降噪）、识别或删除离群点等；对 S.M.A.R.T. 数据进行数据集成的预处理，具体可以是处理不同厂商硬盘特征的不一致和冗余；对 S.M.A.R.T. 数据进行数据变换的预处理，具体可以是对 S.M.A.R.T. 数据进行归一化处理，将数据值限定在预设范围内，以便于后续输入模型进行处理。

数据特征工程的目的是最大限度地从原始数据中提取特征以供后续算法和模型使用。直接采用从硬盘中得到的原始数据作为特征

信息往往是不全面的，硬盘是连续工作的，孤立的看某一时刻的数据不能反映硬盘信息全貌，因此要考虑时序特征，进而可以引入差分特征反映硬盘变化情况，还可以考虑提取硬盘数据均值，中位值等统计特征反映硬盘基本信息。在对数据进行预处理后，还可以采用预设的特征工程，提出特征数据。在特征工程中，可以提取硬盘潜在数据特征，其特征数据包括但不限于原始特征数据，统计特征数据（如均值、中位数等），拟合特征、分类特征等。

（四） 预测模型

近年来越来越多的研究人员展开了对硬盘故障预测的相关研究，按照各种技术问世的先后顺序，大致可以分为传统基于统计的预测方法和基于深度学习预测方法。

1、传统的阈值方法和统计学算法

传统的统计方法基本是依赖于 S.M.A.R.T. 本身具备的阈值属性，加以优化。单纯依赖阈值的方法非常简单，即在硬盘实际运行过程中，如果检测到任一属性的数值超过了硬盘制造商预设的阈值，S.M.A.R.T. 机制发出一个警告提示该硬盘可能会发生故障。但这种检测法的机制过于简陋，而且一般硬盘制造商的阈值设置也比较保守，导致这种故障预测方法的效果极差，仅能达到 3%~10% 的故障检测率（Failure Detection Rate, FDR）和 0.1% 的误报率（False Alarm Rate, FAR）。

在阈值检测法中，以下三个检测项是采用此种方法的主要依赖的 S.M.A.R.T. 数据：

005 Reallocated-Sector-Ct （重映射扇区数）

197 Current-Pending-Sector （等待更新或恢复的扇区计数）

198 Offline Uncorrectable Sector Count （脱机无法纠正扇区计数）

为了提高硬盘故障预测的性能，也有些研究进行了基于 S.M.A.R.T. 属性的硬盘故障统计学预测方法。例如使用马氏距离（Mahalanobis Distance, MD）将属性转换为一维索引，通过定义一个适当阈值或半径度量来检测硬盘异常情况。这种方法最高的 FDR 能接近到 70%。

但这种方法并不能预测所有的故障，以 HDD 硬盘为例：

表 10

故障模式	根因	是否可预测
Media Error	磁头故障	可以预测（对于瞬间造成的磁头/磁碟损坏不可预测）
	磁碟故障	
性能慢	磁头性能弱化	根据硬盘性能的变化能够部分预测（硬盘性能与客户业务相关）
	碟片数据信号退化	
环境问题	特定业务模式下的硬盘固件性能问题	几乎不可预测
固件缺陷	温湿度、振动、误操作等	
其他	开发问题、业务配合问题	
	PCBA、污染问题、兼容性问题、硬盘连接问题等	

但不可否认的是，S.M.A.R.T. 技术提供的与硬盘运行时状态相关的各种硬盘 S.M.A.R.T. 属性信息十分重要，具有很高的隐藏价值，为硬盘故障预测方向的进一步研究打下了坚实的数据基础。

2、基于深度学习算法的预测模型

随着人工智能技术研究热潮的兴起，很多研究人员发现了该技术应用在硬盘故障预测领域，如机器学习、神经网络等，取得了非常优秀的成果，提高了故障预测性能、适应了硬盘异质化、解决了冷启动等硬盘故障预测领域的核心问题，标志着硬盘故障预测研究步入了人工智能算法驱动硬盘故障预测的黄金时代。其处理流程大致为：

输入数据 > 预处理 > 模型训练 > 模型优化集成 > 预测输出

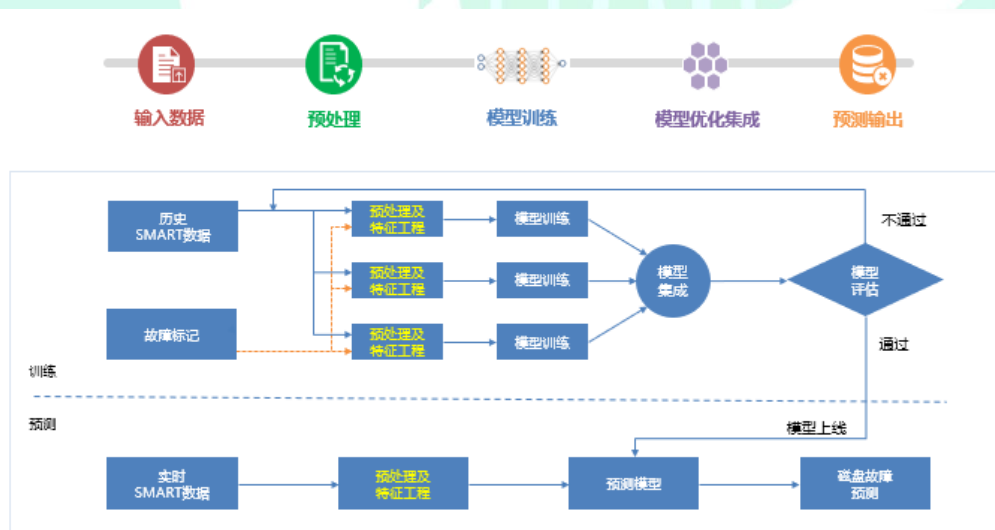


图 9

具体来看，分为线下的训练和线上的预测两部分。

（1）线下训练部分：

数据输入主要是具有故障标记的历史数据（如 S.M.A.R.T. 信息）。选择特定的模型算法，进行预处理及特征工程，模型训练，训练后的模型通常会根据选定的评估指标进行调优集成，以确保预测稳定、准确，当新模型满足准确率、召回率、误判率等一系列量化评估指标后，允许上线。

（2）线上预测部分：

线上预测的数据，不进行故障标记，使用同样的预处理方法及参数，对线上的实时数据进行预处理，然后送入训练好的集成模型中进行预测。通过集成模型，给出磁盘故障的综合多分类预测结果。

1) 传统机器学习

最早的研究最经典的 SVM（Support Vector Machine，支持向量机）模型，最佳效果达到 50.6% 的 FDR，和没有任何误报（FAR=0%）。随着集成学习（Ensemble Learning）技术的发展与成熟，各种集成学习算法开始崭露头角，并在硬盘故障预测方面表现出远超 SVM 的性能。代表性算法包括 XGBoost（Extreme Gradient Boosting）、GBDT（Gradient Boosting Decision Tree，梯度提升树）和 RF（RandomForest，随机森林）等。较好的效果达到了 97.67% 的 FDR 和 0.017% 的 FAR。

2) 神经网络

自从硬盘故障预测步入了以人工智能算法驱动的时代，预测性能再次实现了飞跃性的提升。而除了传统机器学习方法之外，深度神经网络算法也是后来居上，在硬盘故障预测领域大放异彩。

相较于传统的机器学习算法，神经网络的优势在于大数据，理论上神经网络学习的数据越多，其表现就越好。而传统的机器学习算法在达到一定水平后，再学习更多数据后并不会提高其性能。较为经典的算法是使用长短期记忆网络（Long Short Term Memory, LSTM）预测硬盘的剩余使用寿命，可以在训练样本和测试样本上分别达到 0.27 和 1.85 的平均绝对误差（Mean Absolute Error, MAE）。

3) 单类别分类算法

传统机器学习和人工神经网络模型训练过程都需要庞大的数据做支撑，所以其应用还是具有局限性。如果需要在一個全新的存储系统中应用故障预测技术，会遇到前期数据匮乏无法支撑故障预测模型训练的情况。而且通常来说，在真实的存储场景下故障盘的数量会远远少于正常盘数量。在极度缺乏故障盘样本的情况下，一般的机器学习和神经网络算法就失去了用武之地。这就是硬盘故障预测领域里所谓的“冷启动问题”，即当无法采集到足够多的硬盘数据时，常规的硬盘故障预测方法无法取得好的效果。此时硬盘故障预测从一个多类别分类问题变成了一个单类别分类问题。经典算法之一是基于 GAN（Generative Adversarial Network，生成对抗网

络)的全生命周期硬盘故障预测方法,这种技术仅需要在健康硬盘的样本上进行训练,因此不存在冷启动问题。这种方法可以达到 95% 以上的 FDR。

4) 迁移学习

在大规模的存储系统中,为了增加存储容量或替换故障硬盘会经常添加新的硬盘,这会导致随着时间的推移存储系统逐渐由来自不同供应商或同一供应商但不同型号的少量新型硬盘组成。这些不同型号的硬盘之间的 S.M.A.R.T. 或其他属性也有较大差异,因而只有当训练数据和测试数据都来自相同型号的硬盘时,传统的机器学习算法才能取得良好的故障预测性能。直接使用常规的故障预测方法进行跨硬盘型号的故障预测是行不通的。为了优化在小样本数据集上的硬盘故障预测性能,研究人员使用了迁移学习技术。该技术的核心思想就是将大样本中的知识通过迁移学习算法迁移到小样本中去,从而达到扩充小样本数据集的目的,使得常规的故障预测算法有足够的数据支撑训练出性能良好的故障预测模型。目前现有的研究表明,优化迁移学习算法,预测成效可以达到 96% 的平均 FDR,和低至 0.5% 的 FAR

三、主流存储故障运维软件案例

早期比较成熟的存储类 AI 运维软件主要有: EMC 的 AI 存储管理产品 CloudIQ、飞康 Intelligent Predictive Analytics、NetAPP 的 Elio 存储管理聊天机器人和智能运维系统 ActiveIQ 等,

能够一定程度的检测磁盘的运行数据和状态，预测性能问题，提出配置建议。下面重点介绍几个已在大型数据中心中得到广泛应用的案例：

（一）中移云秤-磁盘健康检测平台

磁盘健康检测平台，是提供针对服务器设备中 HDD、SSD 的多种主流品牌型号的磁盘开展健康 AI 模型建模与检测服务。平台支持基于 API 的磁盘运行数据采集，依托 AI 算法建立磁盘健康度预测模型开展磁盘健康评估，并将磁盘健康度预测结果通过监控大屏展示。其关键技术包括：

1、数据采集和预处理：

在服务器部署 Agent 实现自动化采集主机磁盘属性信息，借助软件 S.M.A.R.T. montools 采集磁盘的 S.M.A.R.T. 数据，对数据进行预处理生成磁盘预测模型所需的标准数据。

2、健康度模型

基于 AI0ps 的深度学习 GAN 网络用于预测 HDD 磁盘健康度，基于 LSTM 神经网络用于预测 SSD 磁盘健康度。这两种模型具体使用如下：

HDD 磁盘的预测模型由编码器-解码器-编码器子网络组成，其中所有网络均利用到卷积神经网络。包括下列过程：二维 S.M.A.R.T. 数据图像构建过程、磁盘型号状态嵌入过程、磁盘型号

融合过程、模型训练过程以及模型预测过程。通过模型输出的磁盘健康值，判断磁盘的健康状态

SSD 磁盘深度学习算法 LSTM 构建不同品牌 SSD 盘预测模型。首先选定模型输入过去 30 天的 SSD 盘 S.M.A.R.T. 各指标数据，输出未来 30 天的预测生命值数据，并与提前设定好的建议更换/报修阈值进行对比，若大于阈值，则磁盘处于健康状态，若小于阈值，则磁盘大概率会发生故障。



图 10

中移云秤磁盘健康检测平台的应用价值和应用场景如下图所示。

www.ODCC.org.cn



图 11

平台功能覆盖磁盘 S.M.A.R.T. 性能采集、数据处理与磁盘故障预测服务及预测结果的大屏可视化呈现，共预测出 158 块磁盘故障，成功预测的磁盘故障已达到 98%，使磁盘故障率下降，提高了业务运行的可靠性。

异常磁盘建议更换表									
资源池	主机名	品牌	序列号	型号	规格	接口	寿命参数	预计剩余可用时间(天)	建议维保日期
哈池	IP	INTEL	BTWA63510028480FGN	SSDSC2BB480G6	480GB	SATA	Media_Wearout_Indicator	3	2022-03-18
哈池	IP	INTEL	BTWA634605GM480FGN	SSDSC2BB480G6	480GB	SATA	Media_Wearout_Indicator	3	2022-03-18
哈池	IP	INTEL	BTWA63510028480FGN	SSDSC2BB480G6	480GB	SATA	Media_Wearout_Indicator	3	2022-03-18
哈池	IP	INTEL	BTWA634605GM480FGN	SSDSC2BB480G6	480GB	SATA	Media_Wearout_Indicator	3	2022-03-18
哈池	IP	INTEL	BTWA63510028480FGN	SSDSC2BB480G6	480GB	SATA	Media_Wearout_Indicator	3	2022-03-18

图 12

(二) 超聚变 FusionDirector-故障诊断功能

FusionDirector 故障诊断功能，提供超聚变服务器的内存和 HDD 硬盘开展健康检测服务，该功能基于带外 iBMC 收集的日志，根

据日志中内存和硬盘的日志内容，用于提前预测硬盘故障，提示更换高风险内存或硬盘，提供预测结果查看功能。

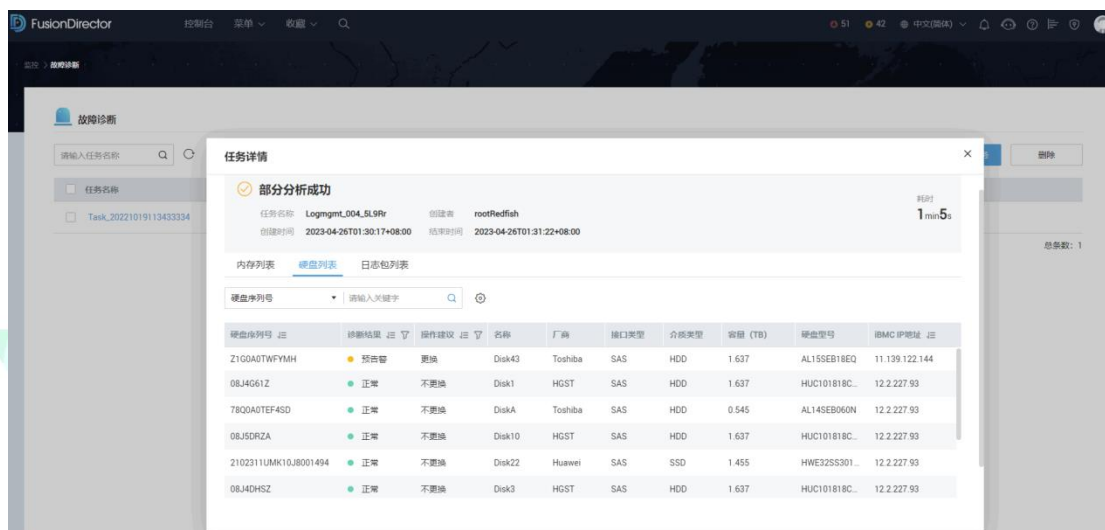


图 13

FusionDirector 采用的数据流向为“Disk > RAID > iBMC > FusionDirector”，带外通道采集不影响用户业务和带内网络带宽，且监控数据会在 iBMC 持久化一段时间。极端情况时，OS 宕机数据丢失后，不影响 FusionDirector 通过带外通道获取这部分数据。具体流程图如下：

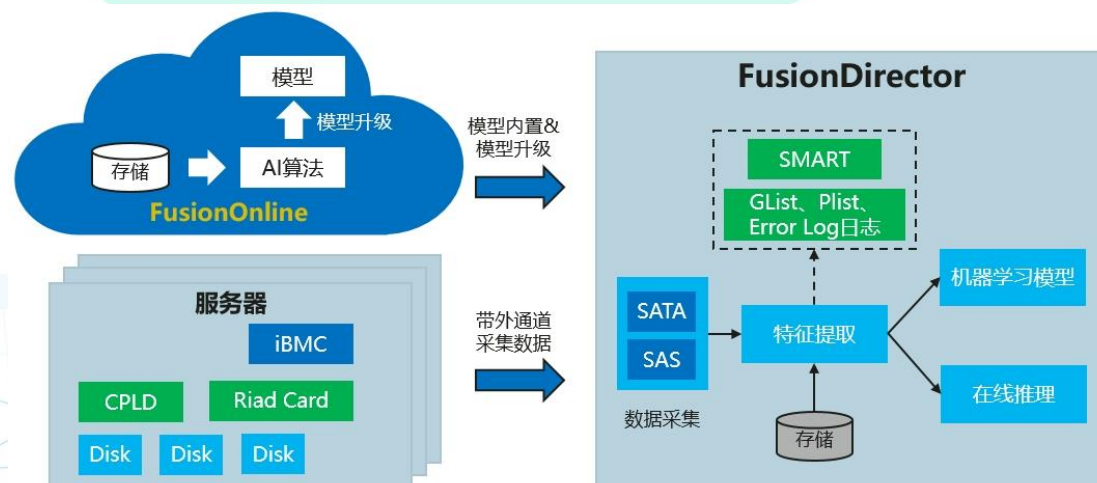


图 14

(三) 三星存储健康预测

三星存储健康预测基于 Telemetry 数据采集，采用多视角的随机森林故障预测方案。目前的评估结果如下表所示，多视角的随机森林故障预测方案与原始的随机森林相比在 FDR 和 FAR 两个指标上都有显著的改进，这表明通过结合不同的视角可以更准确和全面的预测 SSD 故障。

表 11 基于多视角的随机森林故障预测方案与原始随机森林结果对比

方法	FDR (%)	FAR (%)
随机森林	20	0.01
多视角随机森林	30	0.008

三星存储健康预测采用了基于多视角的随机森林故障预测方法，其总体架构如图所示，主要遵循两个思路：1) 使用长期数据的波动和趋势的相关特征来捕获长期的故障症状和模式；2) 结合短期，长期等不同视角的特征分组学习并联合决策来更加准确的预测 SSD 故障。基于多视角随机森林故障预测方案分为两个部分：离线训练和在线预测。离线训练主要包括两个步骤。首先是特征提取，基于离线数据提取原始特征、序列相关特征和服务器相关特征。原始特征关注短期 SSD 数据的异常，序列相关特征和服务器相关特征关注 SSD 长期数据的波动和趋势。然后进行模型训练，生成的多视角特征分别输入到不同的随机森林模型进行训练，以从不同的视角中获取信息。在线预测同样分为两个步骤。首先是特征提取，基于在线数据提取与离线训练相同的多视角特征。然后进行故障预测，将生

成的特征分别输入到相应视角的模型中，结合来自不同视角的决策来预测 SSD 是否故障。

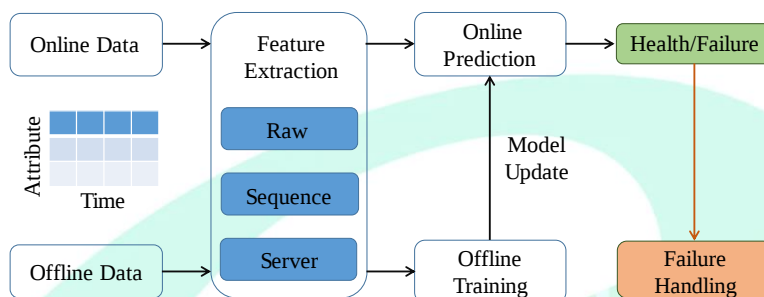


图 15 基于多视角随机森林故障预测方案总体架构

1、多视角特征提取

通过对收集到的 Telemetry 数据进行分析，发现 SSD 故障不仅体现在短期数据的异常值上，还隐藏在长期数据的波动和趋势上，并且故障 SSD 的某些属性的趋势与同一服务器上其他健康 SSD 的趋势有很大不同。因此除了短期原始数据外，引入了序列相关的特征来表征时序数据波动和变化的程度，还引入了服务器相关特征来反映与同服务器上其他 SSD 序列波动和变化的差异，最后将以上特征合并作为组合特征来综合反映 SSD 的状态。

(1) 原始特征：单条 Telemetry 日志的属性值。

(2) 序列相关特征：在固定大小的时间窗口中，分别计算序列数据的趋势和波动程度，如斜率、峰度等。

(3) 服务器相关特征：序列特征与同一服务器上其他 SSD 相同序列特征的平均值的差值。

(4) 组合特征：将上述三个特征连接在一起形成具有全局视角

的组合特征。

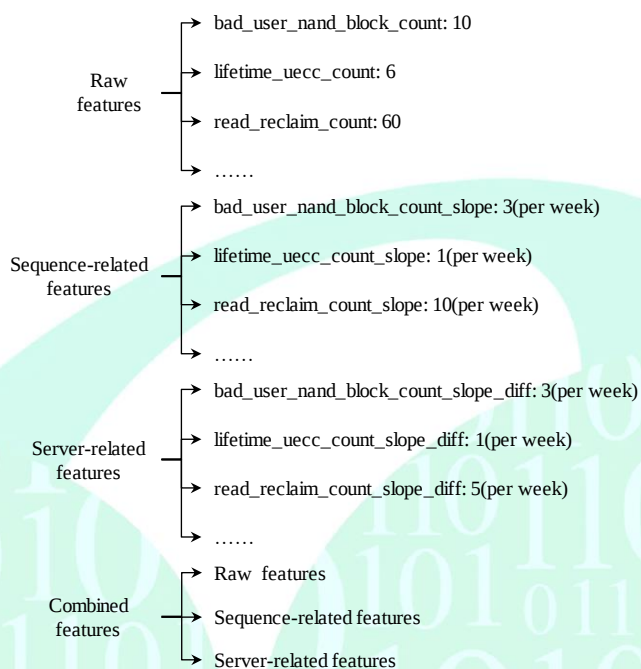


图 16 多视角特征提取

2、多视角随机森林

由于原始特征、序列相关特征、服务器相关特征和组合特征分别从不同的角度表征了 SSD 的状态，从这些不同的角度独立预测 SSD 故障然后共同做出决策会更可靠。因此，我们设计了多视角随机森林，即使用不同的决策树集合来并行学习不同类型的特征。随机森林的所有决策树被平均分为四组，分别学习原始特征、序列相关特征、服务器相关特征和组合特征。然后，四组中的所有决策树进行投票，得到最终的预测结果。得票最多的类别为预测类别，即健康或故障，投票分数为置信概率。

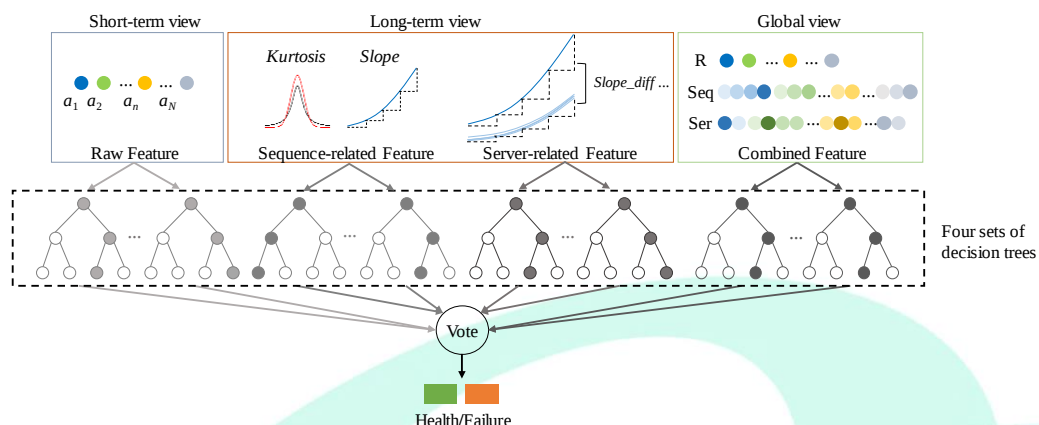


图 17 多视角随机森林架构

（四）浪潮融合智能存储管理平台

浪潮融合智能存储管理平台 InView，是浪潮推出的专门用于数据中心内的监控管理软件，在硬盘预测方面，该平台提供 HDD 故障预测，SSD 寿命预测，可以提前告知用户硬盘风险。

1、HDD 故障预测

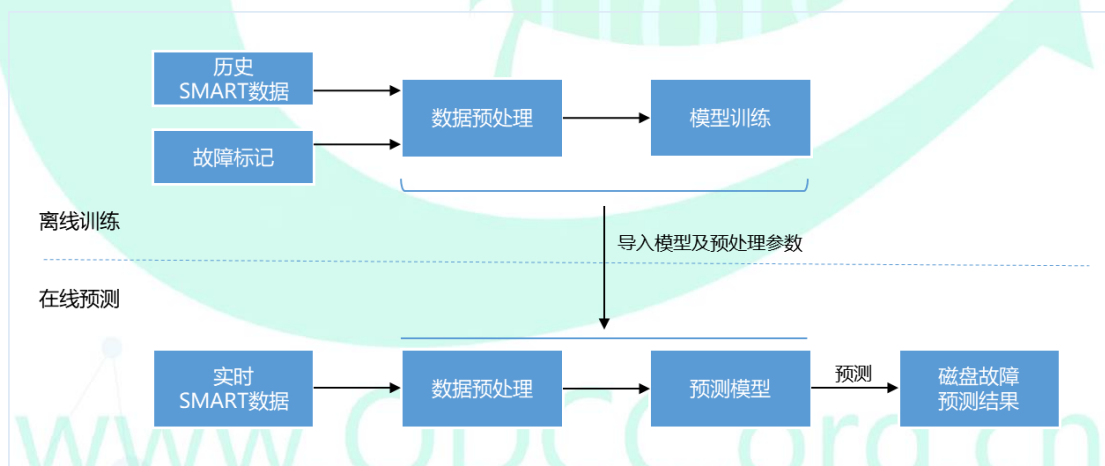


图 18

离线训练：基于历史 S.M.A.R.T. 数据及对应的故障标记，经过数据预处理后进行模型训练。训练采用 xgboost 算法。将训练好的模型及对应的预处理参数保存成文件，用于在线预测。

在线预测：引擎工作时，从数据库中拉取实时采集的 S.M.A.R.T. 数据，采用同样的预处理参数进行数据预处理，然后使用导入的训练过程生成的模型进行预测，输出预测结果。

效果截图如下：

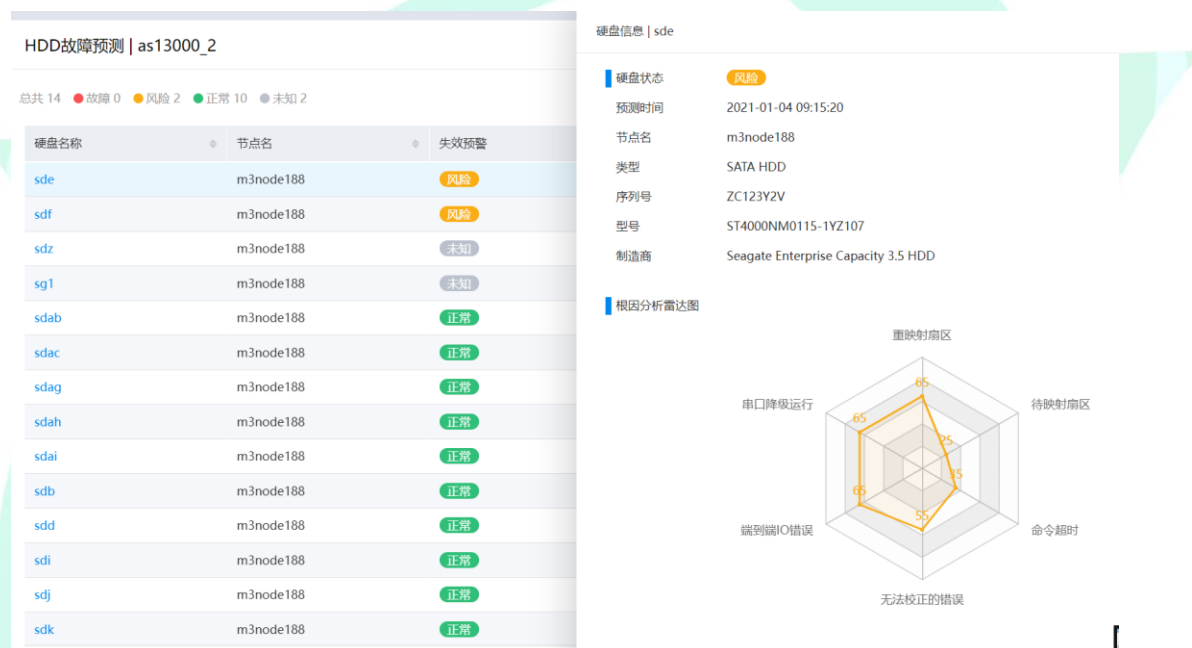


图 19

2、SSD 寿命预测技术

SSD 寿命预测基于硬盘 S.M.A.R.T. 标准采集预测所需的历史时间序列数据集，再使用浪潮自研的 AI 预测算法，预测 S.M.A.R.T. 指标未来变化，计算 SSD 剩余寿命天数。

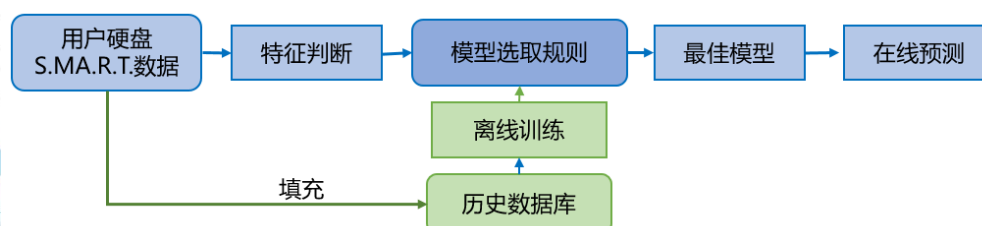


图 20 SSD 寿命预测流程

SSD 寿命预测流程如上图所示。整体的预测流程分为两个阶段，绿色为离线训练阶段，目的是为了确定模型选取规则；蓝色为在线预测阶段，用于在用户环境中预测 SSD 寿命。

离线训练阶段使用大量的 SSD 全生命周期的 S.M.A.R.T. 数据，人工将磨损度变化曲线形态标记为 3 类：平稳变化、减速变化和加速变化，再对三种类型的 SSD 数据分别进行测试。测试过程中实验了多种数据预处理方式和预测模型，比如 Prophet、ARIMA、移动平均法 (Moving Average, MA)、指数平滑法 (Exponential Smoothing, ES)、神经网络等。

效果截图如下：

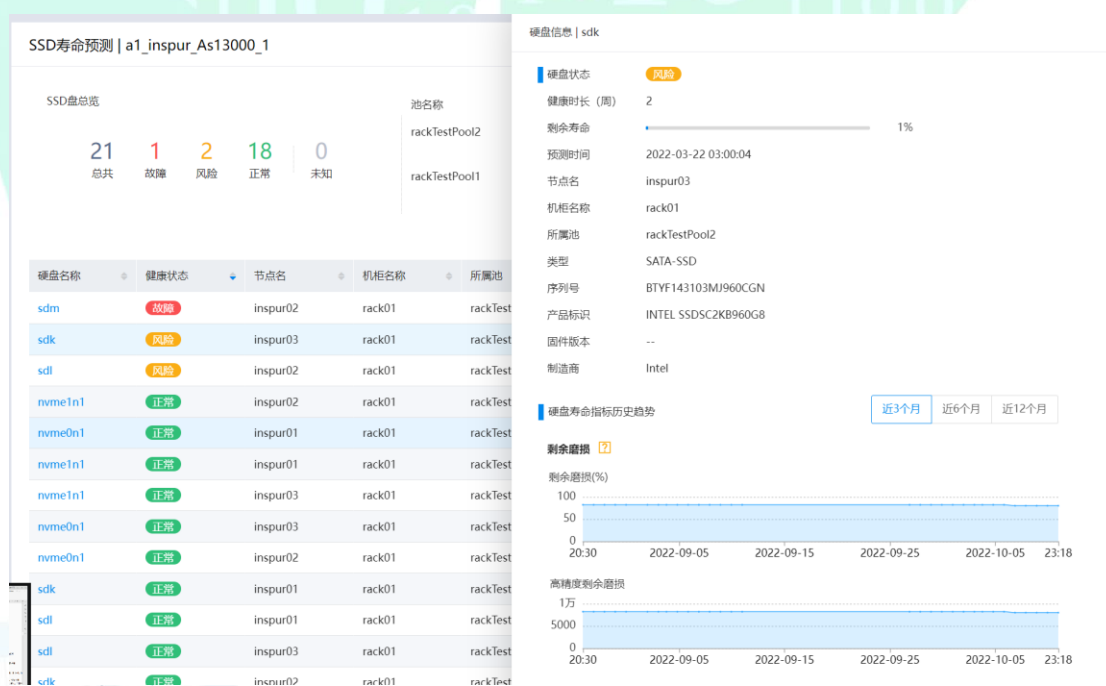


图 21

四、 总结和展望

随着大数据和云计算技术的发展，越来越多的数据和服务将会托管到数据中心中，其停机成本也越来越高，对存储系统故障的容忍性也越来越低。硬盘故障作为存储系统中几乎最为常见的一种故障，对存储系统的可靠性和可用性带来了极大的影响。所以，建立高效的硬盘故障预测模型对于提高存储系统的可靠性和可用性具有深远的意义。

硬盘故障预测系统中，最为重要的是数据源的选择。S.M.A.R.T. 数据是当前主流的数据源，但很多研究表明仅仅使用 S.M.A.R.T. 数据是不能够完全预测出所有的故障盘的。未来还应该依托磁盘技术发展，衍生出更多更精确的数据。

硬盘的使用到故障通常是一个比较长的时间，因此离线建立预测模型需要很长时间段内数据的采集和积累，这对于刚刚建立的数据中心并不友好，因为建立时间比较短的数据中心中没有足够的数据支撑来建立模型。未来还需要在模型上解决这一问题。

数据中心的存储系统本身是海量的、异构和复杂的。ODCC 作为运维工作开放生态的倡导者，我们将促进智能运维在故障预测等方面的技术发展和融合，加强智能运维的发展。

参考文献

- [1] 康艳龙. 基于深度学习的硬盘故障预测技术研究[博士学位论文]: 西安电子科技大学, 2019
- [2] 张廷雷. 基于改进随机森林的硬盘故障预测方法研究[博士学位论文]: 郑州大学, 2019
- [3] 万举. 一种基于强化学习的硬盘故障预测方法[硕士学位论文], 华中科技大学, 2022
- [4] Corinna Cortes, Vladimir Vapnik. Support-vector networks. Machine learning, 1995, 20(3): 273-297
- [5] Austin Coursey, Gopal Nath, Srikanth Prabhu and Saptarshi Sengupta. Remaining Useful Life Estimation of Hard Disk 2109
- [6] 单凯. 基于代价敏感学习的硬盘故障预测[硕士学位论文] 华中科技大学, 2020
- [7] 周志华. 机器学习. 清华大学出版社. 2016.
- [8] Li J., Ji X., Jia Y., et al., Hard drive failure prediction using classification and regression trees. in: 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2014.
- [9] Rincón C. A. C., Pâris J., Vilalta R., et al., Disk failure prediction in heterogeneous environments. in: 2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2017.
- [10] Shen J., Wan J., Lim S.-J., et al. Random-forest-based failure prediction for hard disk drives. International Journal of Distributed Sensor Networks. 2018

附录 缩略语和名词释义

表 12

中文表述	完整的英文表述	英文缩略语
自我检测分析与报告技术	Self-Monitoring Analysis And Reporting Technology	S. M. A. R. T.
固态硬盘	Solid State Disk	SSD
机械硬盘	Hard Disk Drive	HDD
替换率	Annual Replacement Rate	ARR

www.ODCC.org.cn



ODCC公众号



ODCC订阅号

