数据中心云系统智能预警系统

腾讯云、希捷联合出品

腾讯云：牛犇、严勇、李靖、傅欢、宗传涛
希捷：Joshua Zhang、刘健、马勇、Matt Shumway

Tencent Cloud | SEAGATE

目录

摘要：

高效可靠稳定的服务对于云系统至关重要，一个典型的云系统使用大量物理硬盘驱动器做为储存介质，而磁盘失效是导致服务中断的最重要原因之一，磁碟有一些微小的故障信息（例如扇区错误和等待时间错误）作为灰色故障的一种形式，很难被host捕捉到，但是往往是硬盘将要发生故障的前兆。

在本文中，我们尝试主动预测磁盘错误，以免造成更严重的后果损坏云系统。我们可以根据预测结果来优化业务模型，为了建立准确的在线预测模型，我们利用磁盘级传感器(FARM)数据信号，基于可以学习特征的机器学习模型，并根据风险因子对在线的硬盘进行风险排序，根据我们对在线的500K样本进行持续3个月的检测，证实预测的方法是有效的，且表现优于传统基于SMART的评估方法。

背景：

过去的几年中由于用户数呈指数级增长，数据中心的宕机成本大大增加了，IT设备故障是造成此类停机的重要原因，而硬盘失效是最常见的故障组件，同时由于云系统对硬盘失效判断的不准确性，带来了不必要的数据迁移成本和业务影响，磁盘故障有些可以是可预测的，也有一部分是不可预测的。一方面，发生不可预测的故障，从电子元件损坏到突然的崩溃处理不当，无法通过监控来预见，另一方面，可预见的故障主要是由于缓慢磨损过程通常会持续几个月或几年。使故障预测成为可能，本文介绍了用希捷开发的FARM日志来建立机器学习模型来预测硬盘失效的案例。

1. 云系统简介

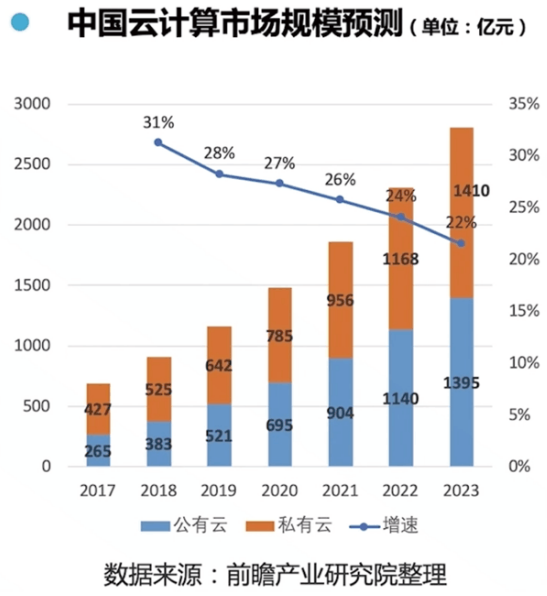
1.1. 发展趋势

中国云计算市场规模突破千亿，持续两位数增长。

云计算发展有4个阶段：

A. 1960年~2005年 – 虚拟化阶段：为了实现资源共享成本优化而形成的虚拟化技术，包括全/半/硬件辅助虚拟化，实现软硬件解耦和资源池化，从而达到局部效率和可靠性提升的目的。

B. 2006年~2018年 – 服务器化阶段：将独立的技术能力包装为整合的服务输出，包含IaaS,PaaS,SaaS等，并将原始的计算和存储服务扩展至安全、音视频、大数据等多领域服务类型。



数据来源：前瞻产业研究院整理

C. 2018年~至今 – 产业化阶段：产业互联网阶段，通过云计算技术助力产业进行变革，打造行业生态与解决方案，助力传统行业上云迁移/云转型。

D. 未来 – 标准化阶段：建立标准化流程，使公有云，私有云，多云互通无缝迁移；提高云服务便利性，像水和电一样使用云服务。

1.2. 业务需求与挑战

腾讯云以卓越科技实力，丰富的实战经验，全面发力产业互联网，助力各行各业数字化转型。当前腾讯云已在全世界多个国家、地区部署数据中心，打造百万级服务器网络，为全球数百万企业和开发者提供领先的云计算。

随着云业务的扩张与服务器数量的增加，腾讯云硬盘已突破千万量级，硬盘的自动化运维与监控预警系统建设在海量服务器运维体系中变的尤为重要。尤其作为客户数据的承载，硬盘健康监控预警的精准性要求更高。一份调查报告显示，57.8%的企业客户将数据安全考虑作为企业上top1的顾虑点，所以硬盘的精确监控与预警，对数据安全与业务连续具有深远而重大的意义。

1.3. 腾讯云业务监控架构与逻辑

腾讯云服务器定期采集服务器硬盘的SMART日志和FARM日志，并通过大数据离线评分模块、实时健康评分模块以及FARM健康预测模块对原始日志进行建模分析，统一输出硬盘健康度评分，并结合用户业务模式给出处理建议。

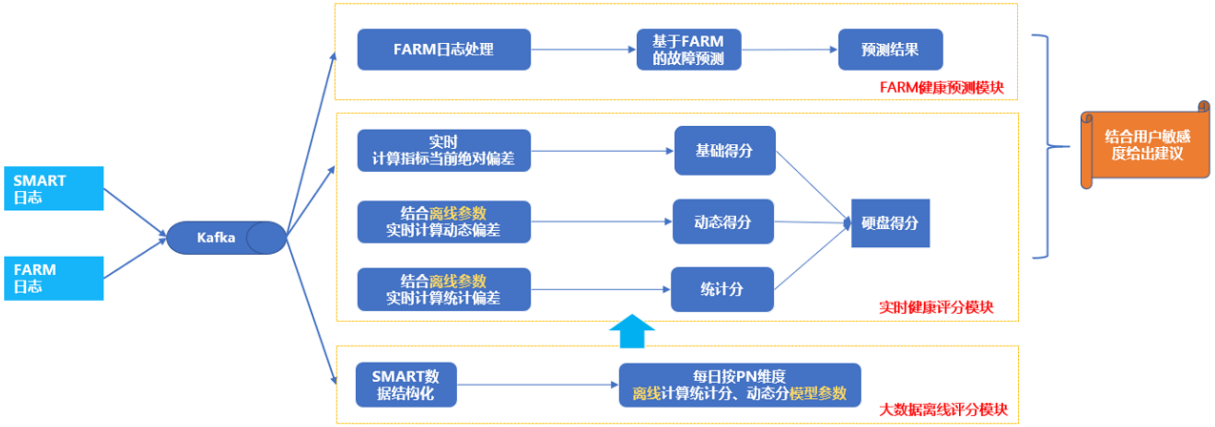


Fig 1. 腾讯云业务监控架构

通过对50万+硬盘样本量进行分析，SMART健康评分与FARM故障风险分呈现正比关系，即基于SMART的健康评分越低，FARM故障风险分数越高。在腾讯云服务器上，将SMART健康评分和FARM故障风险分数进行建模加权后，形成了一套多维度硬盘综合健康分数评估体系。

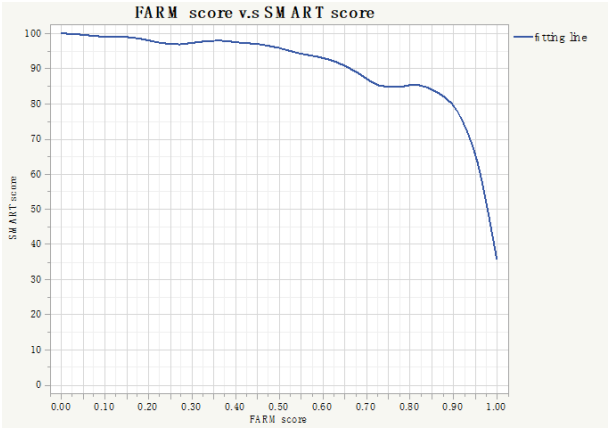


Fig 2. Scores Correlation

2. FARM Pro的简介

2.1. FARM Pro介绍

随着数据中心规模的扩张，在线硬盘的数量也在快速增长。在目前的运维实践中，硬盘驱动器的数据采集主要基于SMART，而SMART体系较为陈旧、只涵盖有限的信息、不够灵活，难以适应不同应用场景对于硬盘健康管理的个性化需求。为了应对这样的挑战，希捷开发了FARM这样一个全新的日志。

FARM日志的主要目的之一是从多个来源汇总数据到单一日志，除了部分现有的SMART日志、T10/T13工业标准日志，还包含了更加底层的传感器和磁头级别的相关参数。同时它的结构简单、开销极小，抓取日志的动作对业务负载不会造成影响。这些特性便于运维系统长期以较细的粒度收集所需要的硬件数据并进行监控和分析。

目前的FARM Pro版本已经包含了120多个参数。随着和客户合作的深入， FARM Pro日志在后续版本上还会进一步完善。

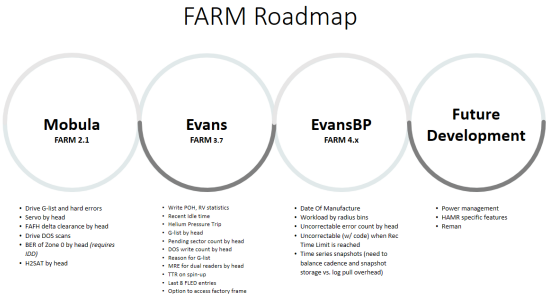


Fig 3. FARM_Roadmap

2.2. FARM Pro日志结构

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。

2.2.1. SATA FARM Pro日志地址为0xA6，其结构如表1-3

表 1 SATA FARM Log Structure

Page	Description
0	FARM Header – 参见**表 2**
1	General Drive Information
2	Workload Statistics
3	Error Statistics
4	Environmental Statistics
5	Reliability Statistics

表 2 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 3 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

2.2.1. SATA FARM Pro日志地址为0xA6，其结构如表1-3

表 4 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 5 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

由于篇幅关系，这里只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SMART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...）等。

2.3. FARM Pro日志的获取和解析

如 1.2 中的日志地址，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。

这里通过希捷 SeaDragon_LogsUtils（日志获取工具）和 SeaDragon_LogParser_FARM（FARM 日志解析工具）来演示:

2.3.1. 获取 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
```

2.3.2. 解析 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 jason 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
```

2.3.2. 按照flatcsv解析后的 FARM 日志内容 （前 10 列），可以看到 flatcsv 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。

```
    root@localhost:~# cut -d',' -f-10 ZA1056YN_FARM_2020-07-06__10_33_45.csv
    SeaDragon_LogParser,Build Version,Build Date,Run as Date,FARM Log,Log Signature,Major
Revision,Minor Revision,Pages Supported,Log Size,...
    NONE,2.2.4,Apr 16 2020,07-06-2020__10:36:51,NONE,0x4641524D4552,1,9,6,98304,...
```

3.模型训练和评估

3.1. 数据流

考虑尽量少的影响对系统IO的影响，我们每天取一帧FARM数据，通过parser自动解析并存在数据库，同时系统报错的硬盘序列号也会同步到数据库，作为训练数据的标签。

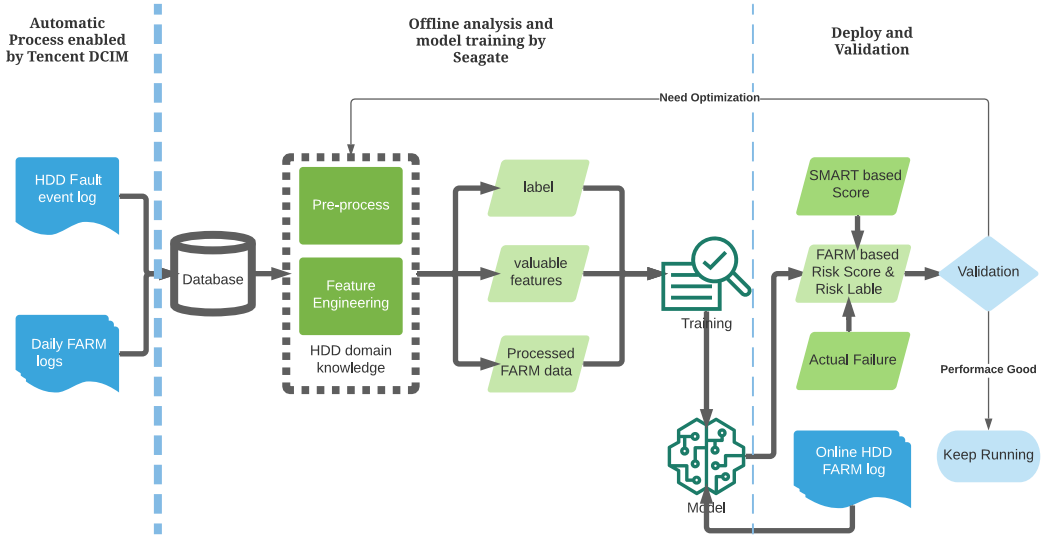


Fig 4. Model Training & Deployment Flow

3.2. 挑战

3.2.1. 样本不平衡

对于大规模的云服务系统，每天在线的上百万硬盘中最多只有不到100个磁盘失效，对于极端的不到0.1%的失效标签，难以训练有效的分类模型，很容易过拟合，不能有效的预测正样本，虽然有些数据重新平衡技术，例如过采样和或技术或SMOTE技术，可疑缓解这一挑战。虽然这些方法有助于提高召回率，但同时可以引入大量的误差，从而大大减少了精度。在我们的场景中，错误的代价很高，会造成不必要的运营成本，我们用采用重复采样（正样本，多帧的数据）来应对这一挑战。

3.2.2. 硬盘运行时间与失效模式

硬盘失效率是典型的 bath curve 曲线(如下图），不同时间的失效率和失效模式会很大的不同，所以我们对正样本过采样，尽可能多的取到一年之内的失效样本。

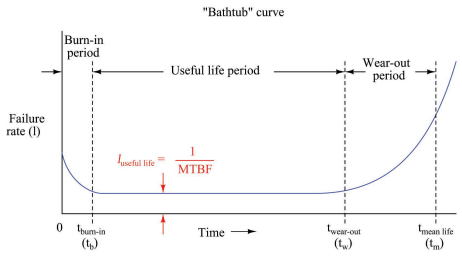


Fig 5. 典型硬盘失效率曲线

3.2.3. 标签的准确性

由于系统判断硬盘失效的机制对于不同的业务和子系统都有很大的不同，还有其他原因的系统误判，比如数据一致性、无效IO命令等等，错误的标签会给模型带来很多不必要的噪音，我们通过线下的验证剔除了假阳性的样本来保证标签的准确性。

3.2.4. 训练与评估结果的差异

通常我们以交叉验证的方式验证模型预测精度，但是，我们发现不适合评估我们的磁盘错误预测模型。在交叉验证中，数据集随机分为训练和测试组。因此，训练集可能包含未来数据的一部分，而测试集只有部分过去的数据。但是，当涉及在线预测时（使用历史数据来训练模型并进行预测未来），训练和测试数据将没有时间重叠。此外，一些数据，尤其是系统级信号，对时间敏感（它们的值不断变化随时间推移）或对环境敏感（负载，读写方式等，从训练集中学到的规律可以应用于测试数据集，从而在交叉验证中具有很高的准确性，但评估新数据时效果不佳，所以我们采用全新的数据做为测试数据集来验证模型的指标。

3.3. 特征工程

3.3.1. 数据预处理

FARM的数据分两部分，一部分是硬盘的环境变量和错误信息（以每个盘一条记录），一部分是硬盘运行性能指标变量（以每个磁头为一条记录），我们要关联这两部分数据集做为训练数据，对于后者来讲，简单的aggregation可能不能表征真正的磁头表现，我们用KNN的方法找出“性能最差”的磁头作为单个硬盘的特征，因为大多数情况下只是单个磁头的性能下降导致硬盘失效。

另外有部分参数的均值偏离其他很多，要做标准化处理

$X\_scale = (Xi - u)/stdev（Xi)$

对于与运行时间相关的参数也要标准化处理，以减小POH（Power on Hours) 造成的偏差

$X\_norm = Xi / (\log (WL) + a*POH)$

有部分参数相对变化量更相关，计算梯度，对于给定的时间窗口w, 变量x在时间t的梯度定义为：

$Diff (x; t;w) = （x(t) - x (t -w) ）/(\log（WL）+ a*POH)$

3.3.2. 参数优选

我们从FARM里已经确定了120个有效的参数。但是，我们发现并非所有参数都有效，而且有些参数和其他参数高度相关，这些无效或者冗余数据会降低模型的效果，所以特征选择对于构建机器学习模型非常关键，现有的特征选择方法分为两大类：统计指标（卡方，相互信息，等）和基于机器学习的方法，例如随机森林，但是，在我们的情况下，常规特征选择方法并不能无法取得良好的效果，因为失效模式存在时间敏感性和对环境敏感性的特点，我们用Evaluation Algorithm的方法来初选特征，然后根据专家的判断剔除冗余或者有干扰的参数 。

3.4. 模型选择

FARM的大部分有效数据都不是正态分布，很多是计数型的类似POISSON分布，并且根据描述性分析来看线性不可分，参数模型和非参数模型在验证的时候都表现很好，因为我们随机分割训练和验证数据集，总体上看训练和验证数据集还是同分布，但是在测试数据集上表现都会有不同程度的下降，显示还有许多工作来减少训练和测试之间的差异。考虑到参数模型对Outlier敏感，而且对数据的分布变化更敏感，训练开比较大，所以我们选择Xgboost来做为base model 来优化 。

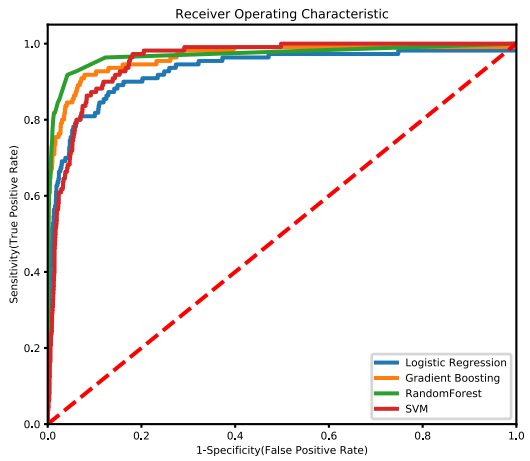


Fig 6. Validation ROC Curve

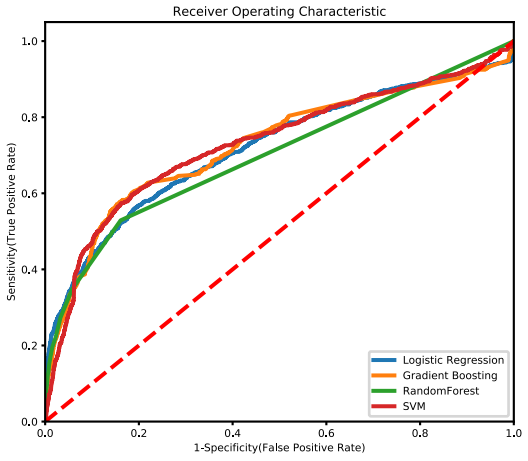


Fig 7. Test Data ROC Curve

3.5. 模型评估

3.5.1. 模型指标

- AUC (Area Under Curve), 用于整体模型的表现
- Recall ，表征 模型的实际失效检测能力
- Precision，模型的预测精度
- Miss Classification Rate，误分类率

我们设定最低的Recall (>30%), 也就是说模型能够预测至少30%实际失效，在次基础上来优化模型达到最优的精度。

		Prediction			Precision	Recall
		0	1	2		
Truth	0	143,645	112	89	0.99	0.99
	1	1,767	70	113	0.30	0.04
	2	568	48	164	**0.45**	**0.21**

Precision = Correct Prediction / Total Prediction = **0.45**

Recall = Correct Prediction / Total Actual Truth = **0.21**

Misclassification = 89 / (89 + 112 + 143645) = **0.06%**

Fig 8. Test Result Confusion Matrix

3.5.2. 风险因子

High Risk (prob_02): 硬盘预测一周之内失效的概率

Low Risk (prob_01): 硬盘预测点7 天 到 30 天之间失效的概率

我们定义风险因子为：

Risk_Score = (0*prob_01 + prob_1 + 2*prob_2) /2

目的是赋予高风险更多权重，更切合实际的应用场景，用单一的风险因子来表征风险的概率。下左图是用来用来划分风险因子区间的依据表征，下右图是预测的风险因子在实际标签中的分布。从下右图可以看出，实际场景中，风险因子在不同的标签中分布有比较明显的差别，因此对风险具有很强的区分能力。

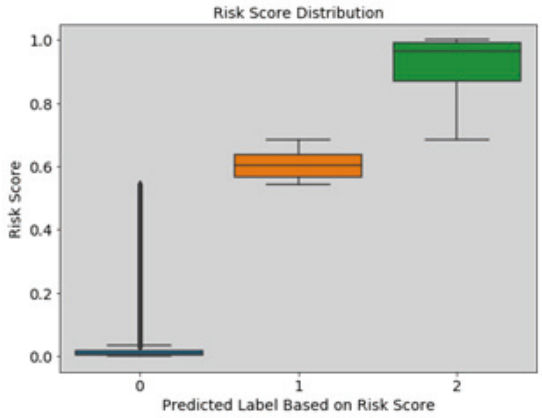


Fig 9. 风险因子基于预测标签的区间划分依据

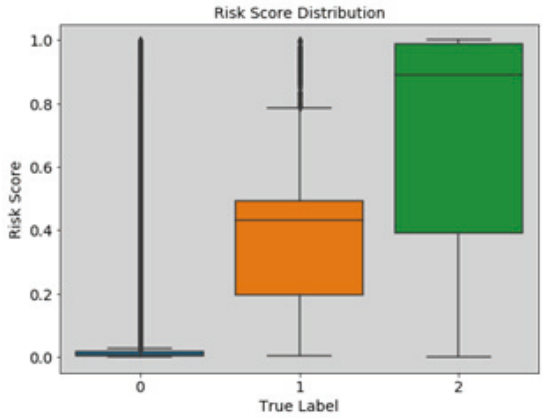


Fig 10. 风险因子在实际标签中的分布

4.生产环境部署

FARM数据可以远程抓取，并在中央服务器进行解析，我们在腾讯数据中心管理系统-TITAN之之上运行打包好的预测模型，可以无缝连接腾讯现有的管理系统。

5.业务应用

5.1. 数据迁移

一旦有硬盘失效，会人工更换一个新的盘，并且把数据迁移到其他地方去，无计划的换盘给业务带来困扰，我们用模型给出的风险因子提前做出换盘计划，可以集中在业务空闲时间段替换掉高风险的硬盘并且做数据迁移。

5.2. 关键任务分配

对于关键业务，不确定的宕机造成的损失非常大，我们会根据模型给出的风险概率，优先分配关键业务到健康的磁盘阵列，同时一旦有风险出现会做提前迁移或备份。

5.3. 提高失效判断准确率

当前情况下，被HOST认为有问题的硬盘的原因有很多种，除了硬盘本身的原因外，网络，驱动，其他的硬件包括应用程序本身的问题有可能被HOST认为是硬盘失效，具我们统计非磁盘问题的比例在30%到40%左右，因此带来额外的换盘和数据迁移成本，同时对业务的影响也非常显著，通过我们根据FARM的风险评级，以及其他的一些系统判断，可以有效降低误判率50%以上。

5.4. Cloud运营状态检测

现代化的数据中心都有DCIM (Data Center Real-Time monitoring system)，用来检测数据中心的整体运行状态，硬盘风险因子作为其中重要一环，极大的提高数据中心运营效率。

6.总结

本文介绍了一套自动、简易的磁盘故障预测方法，用于判断磁盘在接下来一段时间内是否需要替换，通过选择FARM属性、生成时间序列、解决数据不平衡性等步骤，将磁盘故障预测转化为对时间序列数据的分类问题。使用XGBoost算法对磁盘状态进行分类，从而找出可能发生故障的磁盘。在保证预测准确性的同时减少提取数据对业务的影响开销。

关于腾讯云

腾讯云，是中国领先的互联网综合服务提供商腾讯集团旗下的云计算品牌，面向全世界各个国家和地区的企业、组织、机构和个人开发者，提供全球领先的云计算、人工智能、大数据等技术产品与服务。作为产业互联网的基础设施，腾讯云以卓越的技术能力打造丰富的行业解决方案，构建开放共赢的云端生态，助力各行各业实现数字化升级。

关于希捷

希捷致力于打造数据圈，不断创新，提供世界一流、精密设计的数据管理解决方案，专注可持续发展的合作伙伴关系，帮助最大限度地发挥人类的潜力。详情请浏览希捷官方网站www.seagate.com，www.seagate.com/cn/zh，希捷官方微博http://weibo.com/seagatecn，希捷官方微信“Seagate希捷官微（Seagate1979）”，或希捷官方YouTube频道https://www.youtube.com/user/SeagateTechnology。

Tencent Cloud | SEAGATE