2024-3-28

BMC 服务器故障预测与诊断平台

详细设计

高天润 刘开元 王永琪

哈尔滨工业大学
容错与移动计算研究中心

目录

1 数据采集与呈现

数据采集与呈现总体架构如下图 1- 1 所示:

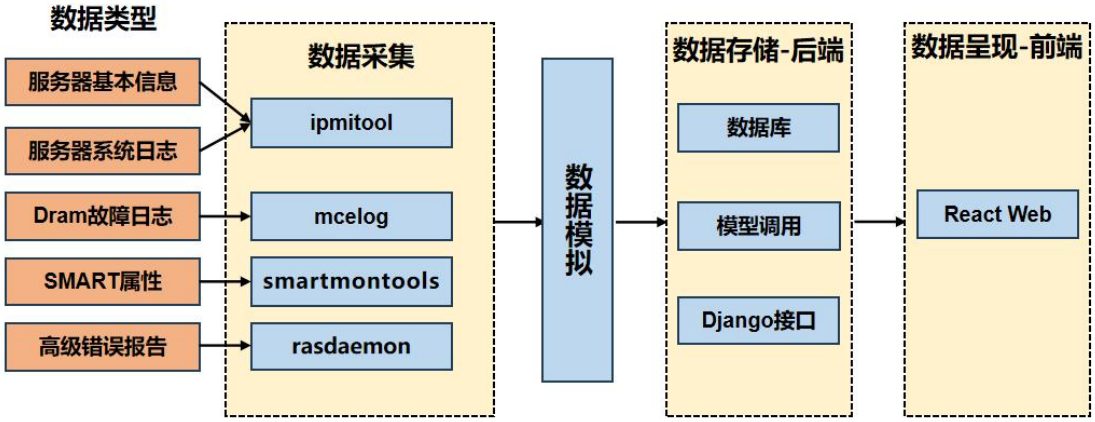

图 1- 1 数据采集与呈现模块架构

1.1 数据采集

平台使用如下 RAS 检测工具:

- Ipmitool: 一种可用在 Linux 系统下的命令行方式的 ipmi 平台管理工具，它可以实现获取传感器的信息、显示系统日志内容
- mcelog: Linux 基于 Intel 的机器检查架构（MCA）记录 DRAM 故障的标准工具
- smartmontools: Linux 用于监控和分析硬盘 SMART（自我监测、分析和报告技术）数据的工具
- rasdaemon: Linux 系统故障和错误检测工具，可生成 AER 高级错误报告

为满足故障分析和故障预测的需求，需要收集以下数据：

服务器基本信息——收集方式：impitool

```
Device ID                 : 32
Device Revision           : 1
Firmware Revision         : 1.00
IPMI Version              : 2.0
Manufacturer ID           : 10876
Manufacturer Name         : Super Micro Computer Inc.
Product ID                : 7054 (0x1b8e)
```

图 1- 2 服务器基本信息收集

服务器系统日志（SEL 日志）——收集方式：impitool

```
root@tester-VIT-E2250:~# ipmitool -I lanplus -H 100.2.76.144 -U admin -P admin sel list
   1 | 06/08/2020 | 11:43:48 | System Boot Initiated #0x6d | Initiated by hard reset | Asserted
   2 | 06/08/2020 | 11:44:16 | System ACPI Power State #0x75 | S4/S5: soft-off | Asserted
   3 | 06/08/2020 | 11:44:20 | Fan #0x8a | Redundancy Lost | Asserted
   4 | 06/08/2020 | 11:44:26 | Power Supply #0x70 | Presence detected | Asserted
   5 | 06/08/2020 | 11:44:29 | Power Supply #0x71 | Presence detected | Asserted
   6 | 06/08/2020 | 11:44:35 | Power Supply #0x28 | Config Error | Asserted
   7 | 06/08/2020 | 11:44:40 | Management Subsystem Health #0xec | Sensor access degraded or unavailable | Asserted
   8 | 06/08/2020 | 11:44:50 | Power Supply #0x74 | Redundancy Lost | Asserted
   9 | 06/08/2020 | 11:45:39 | Chassis #0xee | Transition to Non-critical from OK | Asserted
   a | 06/08/2020 | 13:45:09 | System Boot Initiated #0x6d | Initiated by hard reset | Asserted
```

图 1- 3 服务器系统日志收集

服务器 DRAM 故障日志——收集方式：mcelog

```
[Hardware Error]: section_type: memory error
[Hardware Error]: error_status:0x0000000000000400
[Hardware Error]: physical_address:0x0000000f98cf5fc0
[Hardware Error]: node: 1 card: 1 module: 0rank: 1 bank: 1 row: 42861 column: 192
[Hardware Error]: error_type: 13, scrubcorrected error
[Hardware Error]: DIMM location: not present.DMI handle: 0x0000
```

图 1- 4 服务器 DRAM 故障日志收集

服务器硬盘 SMART 属性值——收集方式：smartmontools

ID#	ATTRIBUTE_NAME	FLAG	VALUE	WORST	THRESH	TYPE	UPDATED	WHEN_FAILED	RAW_VALUE
5	Reallocated_Sector_Ct	0x0003	100	100	000	Pre-fail	Always	-	0
9	Power_On_Hours	0x0002	100	100	000	Old_age	Always	-	3023
12	Power_Cycle_Count	0x0003	100	100	000	Pre-fail	Always	-	2296
175	Program_Fail_Count_Chip	0x0003	100	100	000	Pre-fail	Always	-	0
176	Erase_Fail_Count_Chip	0x0003	100	100	000	Pre-fail	Always	-	0
177	Wear_Leveling_Count	0x0003	100	100	000	Pre-fail	Always	-	213
178	Used_Rsvd_Blk_Cnt_Chip	0x0003	100	100	000	Pre-fail	Always	-	0
179	Used_Rsvd_Blk_Cnt_Tot	0x0003	100	100	000	Pre-fail	Always	-	0
180	Unused_Rsvd_Blk_Cnt_Tot	0x0003	100	100	005	Pre-fail	Always	-	93
181	Program_Fail_Cnt_Total	0x0003	100	100	000	Pre-fail	Always	-	0

图 1- 5 服务器硬盘 SMART 属性收集

PCIE 高级错误报告 AER——收集方式：rasdaemon

```
)0.0: PCIe Bus Error: severity=Corrected, type=Data Link Layer, id=0000(R
)0.0:   device [8086:2030] error status/mask=000000c0/00002000
)0.0:    [ 6] Bad TLP
)0.0:    [ 7] Bad DLLP
)0.0: AER: Multiple Corrected error received: id=ae00
)0.0: PCIe Bus Error: severity=Corrected, type=Data Link Layer, id=0000(Rec
```

图 1- 6 服务器 PCIE 高级错误报告收集

1.2 数据模拟

对于真实的服务器，需要通过以上开源软件收集服务器的实时数据，但由于真实的服务器发生错误频率较低，对于故障分析和预测的样本量不足，平台在分析真实数据的格式后，使用生成的模拟数据用于进一步的分析和预测工作，除基本信息部分使用 ipmitool 读取浪潮服务器的真实数据外，各个部分模拟数据生成的实现方式如下：

1. 服务器系统日志：以第三届阿里云磐久智维算法大赛日志数据[1]为原始数据，从原始的 48 万条数据中，随机抽样 10000 条日志数据
2. 服务器 DRAM 故障日志：以第二届阿里云智能运维算法大赛 Dram[2]故障日志为原始数据，从原始的 100 万数据中，随机抽样 10000 条数据，重新调整时间戳，保持时间戳范围与服务器系统日志一致
3. 服务器硬盘 SMART 属性值 ：以 Backblaze 年度硬盘报告[3]为原始数据，从中选出 5 个硬盘数据作为服务器的模拟硬盘数据，选择时间戳范围与服务器系统日志一致内的硬盘数据

[1] https://tianchi.aliyun.com/dataset/121954
[2] https://tianchi.aliyun.com/competition/entrance/531874/information
[3] https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data

4. PCIE 高级错误报告 AER：从 rasdaemon 获取的阿里云服务器少量 AER 报告中分析格式，使用相应范围的随机数生成相同格式的模拟数据，所保持时间戳范围与服务器系统日志一致

1.3 数据存储

所有模拟数据均存储在后端数据库中，通过预先设立的 API 接口获取，后端是基于 Python 开源框架中的 Django 框架作为开发基础，使用 MySQL 作为数据库，建立的模拟服务器。如图 1-7 所示，在 Django 中，操作表的语句与直接在 MySQL 中执行的 SQL 语句有着密切的关系。Django 的 ORM（Object-Relational Mapping）提供了一种 Pythonic 的方式来操作数据库，它将 Python 代码翻译成对应的 SQL 语句，并执行这些 SQL 语句以实现数据库操作。

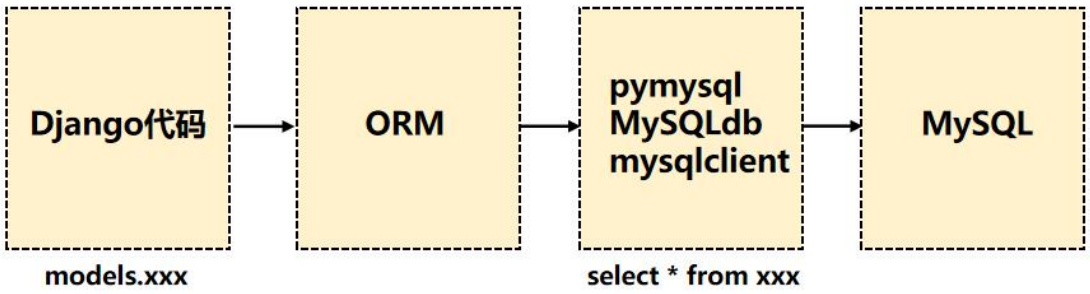

图 1- 7 Django 调用 MySQL 数据库流程

模拟服务器向前端提供数据接口并发送数据，该模拟服务器部署在带有 Ubuntu22.04 操作系统阿里云服务器上。

图 1- 8 Django 后端提供模拟服务器数据接口

1.4 数据呈现

在获取相应需要的所有的服务器模拟数据之后，Django 后端将会调用模型，将输入数据传递给模型，使得模型给出相应的故障分析和预测结果，并由 Django 将结果转化为 JSON 格式发送给前端，前端以 React 框架构造的 Web 形式展示相应的数据和结果。React 使用 BrowserRouter 路由模式，并使用 Nginx 匹配 React-Router。

图 1- 9 React Web 前端负责展示数据和检测结果

1.5 数据格式说明

数据格式-服务器系统日志 SEL

服务器系统日志 SEL 提供主要设备状态变化的历史记录，用于故障诊断。BMC 能够记录基于 IPMI 传感器的事件历史记录，IPMI 规范定义的 IPMI 标准的事件均会被记录。本次比赛中使用从第三届阿里云磐久智维算法大赛数据中生成的格式相同的模拟数据。

表 1- 1 服务器系统日志 SEL 格式

属性	描述
time	SEL 日志产生时间
msg	SEL 日志信息内容

数据格式-平台故障分析日志

在原始 SEL 数据上，使用日志解析和分类模型得到相应的结果，实现故障分类，故障定位等功能。

具体的故障分析日志格式见表 1- 2。

表 1- 2 平台故障分析日志格式

属性	描述
故障等级	日志错误等级，包括 INFO：通知 WARNING:告警 CRITICAL:关键 ALERT：严重警示
故障类型	告警事件关联的部件类型，包括 CPU PSU Memory Disk PCIE FAN INTRUSION OS STATUS ACPI STATUS Boot LAN Other 其它
事件码	告警事件的唯一故障编码 从 AUGG0001 编码到 AUGG0206，分别对应 从原始日志中解析出的每一条模板
产生时间	原始日志的产生时间

属性	描述
主机名	服务器对应的主机名称
日志信息	原始日志的告警信息

数据格式-平台内存（DRAM）故障定位日志

安装到服务器的 DRAM 单元称为双列直插式内存模块（DIMM）。本次比赛使用阿里云 PAKDD2021 数据中生成的模拟数据，一台服务器最多可以安装 24 个 DIMM，其中每个 DIMM 都属于一个通道，服务器中的 24 个 DIMM 编码为 0 到 23（以名称 memory 表示）。

在每个 DIMM 内，其物理位置进一步按 rank 和 bank 组织，其中每个 DIMM 具有两个 rank，每个 rank 具有 16 个 bank。每次内存访问将仅访问 32 个 bank 之一。对于每个 bank，我们可以将其视为一个二维数组，其中数组中的每个元素将存储一个位数据，bank 有 2**17 行和 2**10 列，具体见表 1- 3。

表 1- 3 平台内存（DRAM）故障定位日志

属性	描述
memory	取值范围[0,23], DIMM 的代号
rankid	取值范围[0,1], DIMM 的其中一面
bankid	取值范围[0,15]
row	取值范围[0,2**17-1]
col	取值范围[0,2**10-1]
datetime	日志上报时间

数据格式-服务器硬盘 SMART 属性

Self-Monitoring Analysis and Reporting Technology（S.M.A.R.T.），即“自我监测、分析及报告技术”，是一种自动的硬盘状态检测与预警系统和规范。通过在硬盘硬件内的检测指令对硬盘的硬件如磁头、盘片、马达、电路的运行情况进行监控、记录并与厂商所设定的预设安全值进行比较，若监控情况将要或已超出预设安全值的安全范围，就

可以通过主机的监控硬件或软件自动向用户作出警告并进行轻微的自动修复，以提前保障硬盘数据的安全。SMART 属性数目超过 200 条，各个属性的意义不尽相同，也每种类型的硬盘具备的 SMART 属性也不尽相同，并且其判断方式也不尽相同。

本次比赛我们参照著名的备份厂商 backblaze 的文章[4]中的建议主要监控以下五个属性的 RAW_VALUE 的值：

SMART 5 – Reallocated Sector Count.

SMART 187 – Reported Uncorrectable Errors.

SMART 188 – Command Timeout.

SMART 197 – Current Pending Sector Count.

SMART 198 – Uncorrectable Sector Count.

如果发现该五项值中有大于 0 的情况，就说明该磁盘应该对它进行持续的关注，根据其经验这 5 项 S.M.A.R.T.指标的增长表明即将发生磁盘驱动器故障。

表 1- 4 服务器硬盘 SMART 属性

属性	描述
SMART 5 – Reallocated_Sector_Count	重定位磁区计数，记录由于损坏而被映射到无损的后备区的扇区计数
SMART 187 – Reported_Uncorrectable_Errors	报告不可纠正错误，硬件 ECC无法恢复的错误计数
SMART 188 – Command_Timeout	通信超时,由于无法连接至硬盘而终止操作的统计数，一般为 0，如果远超过 0，则可能电源问题，数据线接口氧化或更严重的问题
SMART 197 – Current_Pending_Sector_Count	等候重定的扇区计数，记录了不稳定的扇区的数量
SMART 198 – Offline_Uncorrectable	无法校正的扇区计数，记录确定出错的扇区数量

[4] https://www.backblaze.com/blog/hard-drive-smart-stats/

此外我们还通过 Xgboost 分类器选择了在故障分类中起最关键作用的 5 项 S.M.A.R.T.指标。对于前端展示的 ST4000DM000 和 ST12000NM0008 两种型号的硬盘，Xgboost 权重最高的 5 项 S.M.A.R.T. 指标按权重大小自上而下排列，如表 1-5 和表 1-6 所示。

表 1- 5 服务器硬盘 SMART 属性（ST4000DM000）

属性	描述
SMART 242 – Total LBAs Read	LBA 读取总数计数
SMART 7 – Seek Error Rate	（该属性是特定制造商才有的）磁头寻找磁道由于机械问题而出错几率
SMART 241 – Total LBAs Written	LBA 写入总数计数
SMART 193 – Load Cycle Count	计量磁头在加电时移至停泊区和移至盘片循环的值
SMART 9 – Power-On Hours	硬盘自出厂以来加电启动的统计时间，单位为小时（或根据制造商设定为分钟或秒）

表 1- 6 服务器硬盘 SMART 属性（ST12000NM0008）

属性	描述
SMART 193 – Load Cycle Count	计量磁头在加电时移至停泊区和移至盘片循环的值
SMART 242 – Total LBAs Read	LBA 读取总数计数
SMART 241 – Total LBAs Written	LBA 写入总数计数
SMART 7 – Seek Error Rate	（该属性是特定制造商才有的）磁头寻找磁道由于机械问题而出错几率
SMART 240 – Head Flying Hours	磁头处于定位中的时间

2 故障分析

目标：给定服务器系统日志判断服务器发生的故障等级，故障类型并根据不同的故障类型，结合服务器 DRAM 故障日志，服务器硬

盘 SMART 属性值，PCIE 高级错误报告 AER 进行定位。

故障分析架构如下图 1- 1 所示：

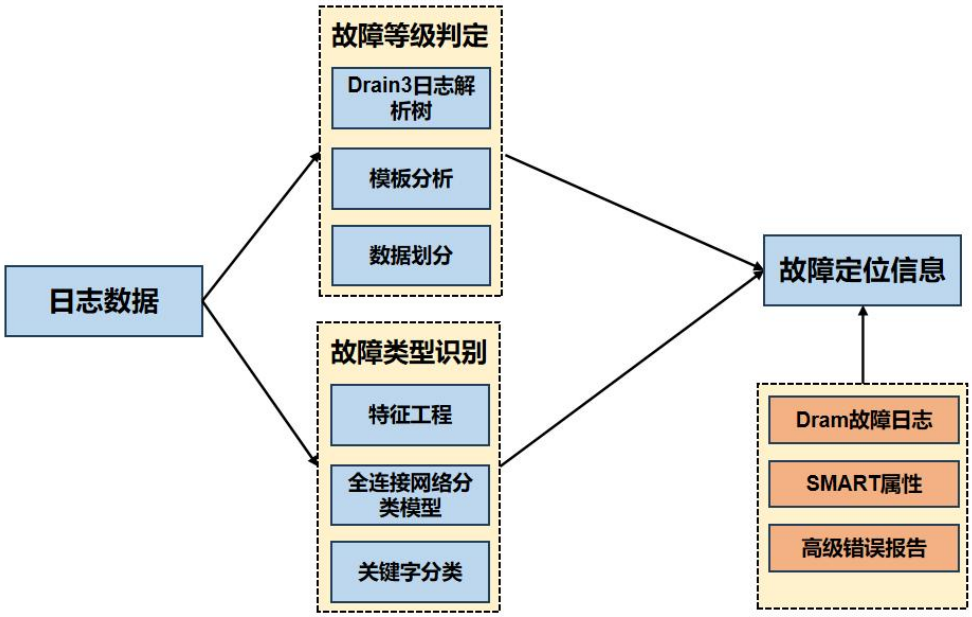

图 2- 1 故障分析架构

2.1 故障等级判定

Drain 是《Drain: An Online Log Parsing Approach with Fixed Depth Tree》[5]中提出的日志解析方法，Drain3 作为 Drain 的 Python3 改造版本，常用于各类日志解析。原始数据为包含所有服务器系统日志的样本集，Drain3 库从原始日志中提取出正则表达式模板，共计 206 条。

[5] https://ieeexplore.ieee.org/document/8029742/

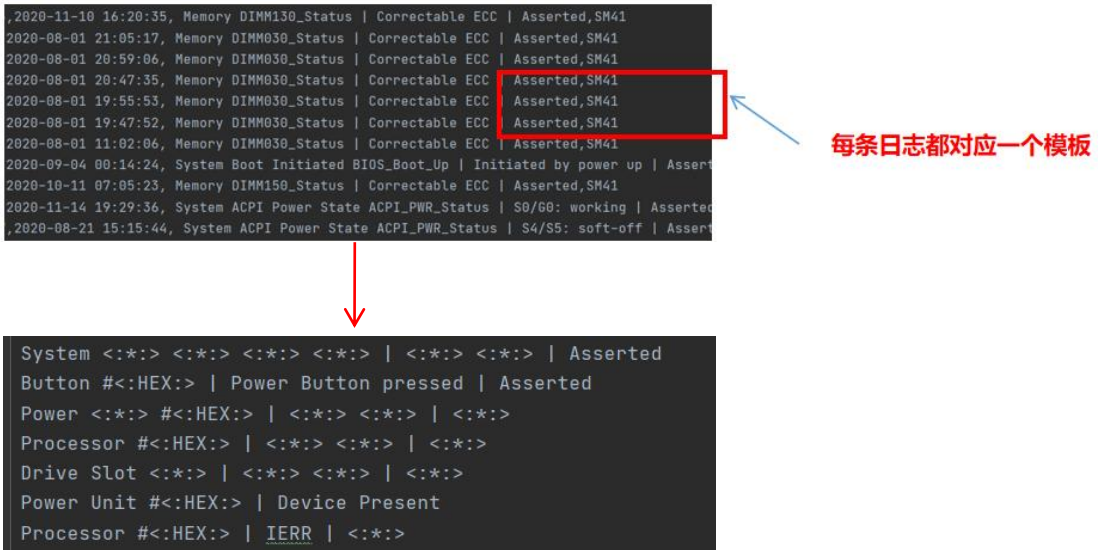

图 2- 2 Drain3 从原始日志中提取模板

在模板提取完成之后，人工逐条分析每种模板在服务器故障中的具体含义，并把每个模板归类于一种故障等级，包括 INFO：通知，WARNING:告警，CRITICAL:关键，ALERT：严重警示。其中 INFO 等级代表服务器正常运行中发生的事件，例如"*System Boot Initiated .* .* .* .* .* .* | Asserted*"代表系统启动事件，WARNING 等级及以上则被认为服务器发生了错误。

```
r"System Boot Initiated .* .* .* .* .* | Asserted": "INFO",
r"Processor .* .* | Uncorrectable machine check exception | .*": "CRITICAL",
r"Memory .* | .* .* | .*": "WARNING",
r"System .* .* .* | .* .* | Asserted": "INFO",
r"Cable/Interconnect #((?<=[^A-Za-z0-9])|^)(0x[a-f0-9A-F]+)((?=[^A-Za-z0-9])|$) | C
r"Add-in Card #((?<=[^A-Za-z0-9])|^)(0x[a-f0-9A-F]+)((?=[^A-Za-z0-9])|$) | Device P
```

图 2- 3 每种模板归类于一种故障等级

2.2 故障类型识别

故障类型识别的目标是从日志信息中识别出故障发生的具体类型，参照浪潮 BMC 服务器分析指南，故障最终被分为 12 种类型：CPU 故障， INTRUSION 故障，PSU 故障，OS STATUS 故障，Memory

故障，ACPI STATUS 故障，Disk 故障，Boot 故障，PCIE 故障，LAN 故障，FAN 故障。

考虑到作为模拟数据源的阿里云磐久智维算法大赛日志数据给出了故障类型标签，分类任务采用有监督预分类加关键字分类相结合的方式。在进行预分类之前，首先将原始的文本数据提取为向量特征，提取方法为：以一个小时为单位，每个时间点对应的向量为单位内所有模板发生次数的 one-hot 编码，共计 206 个模板形成 206 维向量。

```
me_gap,template_id_1,template_id_2,template_id_3,template_id_4,template_id_5,template
2020-06-16 03:00:00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
2020-06-16 03:00:00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
2020-06-16 03:00:00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
2020-06-16 03:00:00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

图 2- 4 原始文本数据特征提取

得到的向量需要输入分类器进行分类，为了尽可能提高分类效果，我们选择使用逻辑回归，支持向量机，随机森林，XGboost，全连接网络等多个模型分别进行对比实验，选择具有最高分类 F1-Score 的分类器，根据实验结果最终我们选择全连接网络作为分类器。

	precision	recall	f1-score	support
0.0	0.622047	0.229651	0.335456	344.000000
1.0	0.679887	0.650407	0.664820	738.000000
2.0	0.815593	0.940765	0.873718	1857.000000
3.0	0.762376	0.700000	0.729858	440.000000
accuracy	0.773602	0.773602	0.773602	0.773602

图 2- 5 具有最高 F1-Score 的全连接网络分类结果，F1-Score 为 0.7736

由于原始标签只分为了 cpu1 类,cpu2 类，内存,以及其它共 4 类，我们使用关键字补充的方式在预分类的基础上进一步分类，得到最终的 12 分类结果。

根据以往经验，我们采用以下关键字：

- Processor 对应 CPU 故障
- INTRUSION 对应 INTRUSION 故障
- Power Supply 对应 PSU 故障
- OS STATUS 对应 OS STATUS 故障
- Memory 对应 Memory 故障
- ACPI 对应 ACPI STATUS 故障
- Drive Slot 对应 Disk 故障
- Boot 对应 Boot 故障
- Critical Interrupt 对应 PCIE 故障
- LAN 对应 LAN 故障
- FAN 对应 FAN 故障

如果包含相应关键字则为对应类型，如不包含以上关键字则判定为其它（Other）类型；对于使用分类模型得到的预分类结果，使用以上关键字重新调整分类结果。最终得到的故障等级判定和故障类型识别结果在前端 Web 的日志告警界面呈现，如图 2- 6 所示。

图 2- 6 故障等级判定和故障类型识别 Web 呈现

对于由故障类型识别的 12 种分类结果，选择其中 3 种类型：Memory 内存，Disk 硬盘，PCIE 进行进一步的故障定位。对每一条故障等级大于 INFO 的日志，根据 mcelog，smartmontools，rasdaemon 得到的内存，硬盘，PCIE 信息，输出与故障 log 产生时间最接近的定

位信息作为故障定位结果，由 Web 界面上满足条件日志后的详情按钮呈现。对于硬盘和 PCIE，Web 界面会输出所有时间点的报错信息作为额外参考。

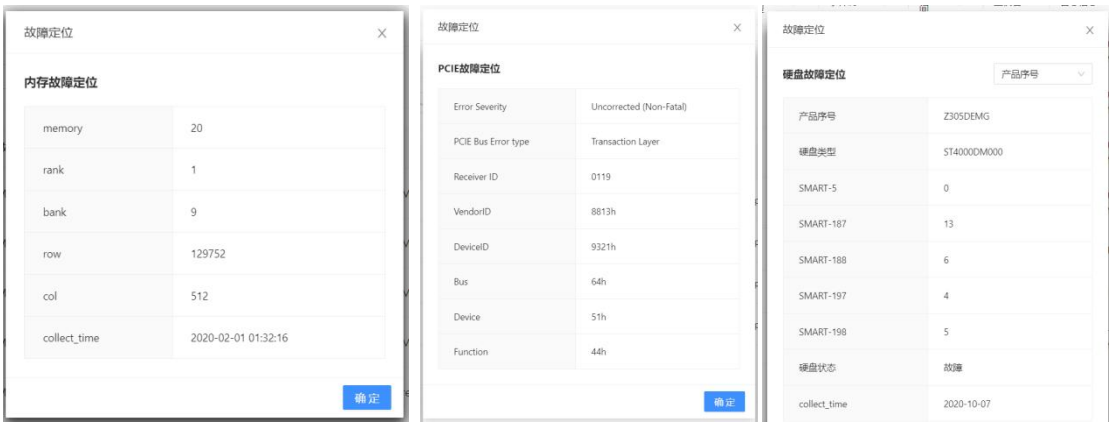

图 2- 7 Web 呈现

3 故障预测

目标：给定硬盘的 SMART 检测数据，预测该硬盘是否会在 20 天内损坏，并给出损坏的倒计时。在模拟数据中，我们使用 BackBlaze 2022 年数据中，ST4000DM000 和 ST12000NM0008 两种型号的硬盘作为案例，实际中的任何类型的硬盘理论上均可使用说明中的方法实现系统中的效果。

故障预测模块架构如下图 1- 1 所示：

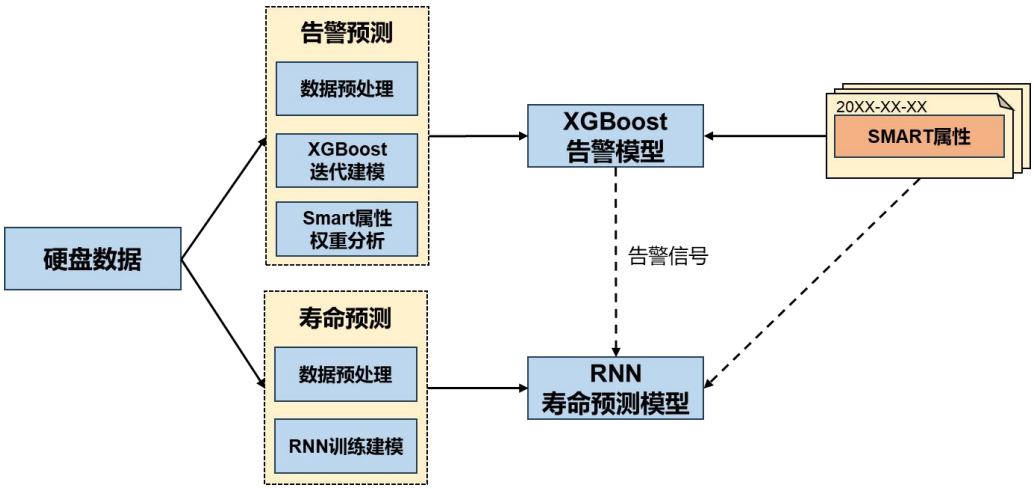

图 3- 1 故障预测模块架构

3.1 硬盘故障告警

硬盘故障告警基于 XGBoost 模型实现。我们将损坏硬盘作为正样本（Positive Sample），并随机抽取等量于正样本数量的正常硬盘作为负样本（Negative Sample）作为训练样本集。对于损坏硬盘，我们取损坏前 20 天的 SMART 数据作为真正的正样本，而其他天数的

SMART 数据依然属于负样本；对于正常硬盘，我们取全部天数的 SMART 数据作为负样本，构成样本数据集，并将样本数据集划分为 80%的训练集和 20%的测试集进行训练，并使用多种不同组合的参数进行尝试，最终选择 F1 分数和 Recall 分数最高的模型作为投产模型使用。最终选取的模型的训练参数如下表 3- 1 所示:

表 3- 1XGBoost 参数表

参数	值	参数	值
round(epoch)	10000	colsample_bytree	0.75
booster	gbtree	min_child_weight	1
objective	binary:logistic	eta	0.025
eval_metric	auc	seed	0
max_depth	5	nthread	8
lambda	1	gamma	0.05
subsample	1	learning_rate	0.01

该模型的实际效果为，将当前硬盘的前 20 天的 SMART 值（SMART 选取表见后端文件/Disk/module/disk_smart_parameter.py 中的 REALIST 元组）输入，模型会给出 0 或 1 作为输出，1 表示为告警，预估硬盘会在未来 20 天内有较高概率发生损坏。

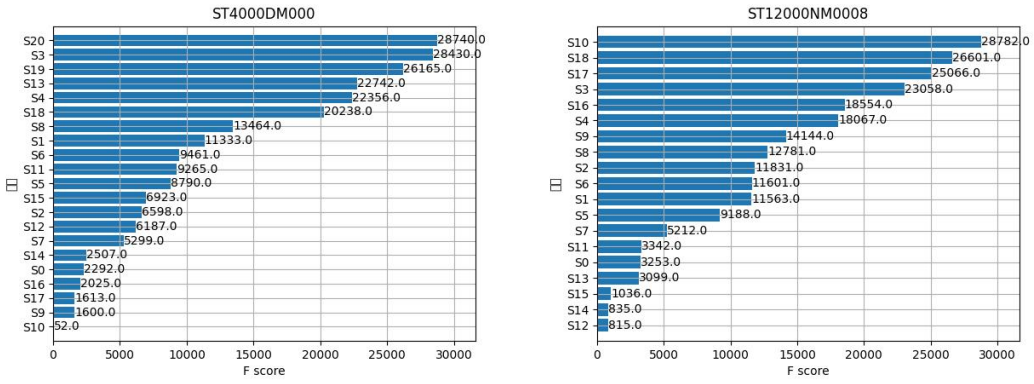

图 3- 2XGBoost 变量权重图

如图 3- 2 所示为 XGBoost 训练的结果，左侧标签栏为按照 REALIST 元组重编号的标签，SX 表示第 X 号 SMART 在 XGBoost 模型中的权

重。

我们发现，权重总和最高的 SMART 与 BackBlaze 推荐的 SMART 存在差异，XGBoost 选中的排名前五 SMART 如下表 3- 2 所示，其中加黑的 SMART 表示此监控指标超出安全范围会对性能严重影响，甚至导致数据丢失。

表 3- 2SMART **选取比对**

#	BackBlaze 推荐		ST4000DM000		ST12000NM0008	
1	**5**	**Reallocated Sector Count 重定位磁区计数**	242	Total LBAs Read LBA 读取总数	193	Load Cycle Count 磁头释放收回循环
2	187	Reported Uncorrectable Errors 报告不可纠正错误	7	Seek Error Rate 寻道错误率	242	Total LBAs Read LBA 读取总数
3	**188**	**Command Timeout 通信超时**	241	Total LBAs Written LBA 写入总数	241	Total LBAs Written LBA 写入总数
4	**197**	**Current Pending Sector Count 等候重定的扇区计数**	193	Load Cycle Count 磁头释放收回循环	7	Seek Error Rate 寻道错误率
5	**198**	**Offline Uncorrectable 无法校正的扇区计数**	9	Power-On Hours 硬盘加电时间	240	Transfer Error Rate 传输错误率

我们相信，BackBlaze 给出的指标是推荐指标，但不同的硬盘实际上会有不同的更为典型的特征 SMART 作为即将损坏的标志，例如希捷的这两款型号的硬盘就在寻道错误率在距离损坏大约只剩下 20 天的时候，就会有更为明显的特征，比“重定位磁区计数”、“通信超时”等这些明显代表硬盘即将损坏的 SMART 变量，更早出现损坏特征。而 XGBoost 相对精准的找到了更为适合的告警变量，用于评估硬盘现在的状态。

此外，LBA 读取总数、LBA 写入总数和磁头释放收回循环这三个

变量本身就代表了硬盘的使用寿命，它们成为了主要分类指标，实际上说明了 ST4000DM000 和 ST12000NM0008 这两个型号的硬盘，具有典型的读写寿命特征，换句话说，厂商生产的硬盘读写头的使用次数寿命是在一个特定范围之内的，基于它们的增长率以及极限，理论上便可较好的估算剩余寿命，因此寿命预测理论上是可以使用机器学习模型实现的。

针对模型在试产环境的真实效果的评估，我们使用了原始数据集中所有正样本和 10%的负样本来进行测试，我们将所有的数据按每一天输入 XGBoost，如果 XGBoost 告警，则停止并记录距离硬盘损坏的天数。

正样本的结果如下所示图 3- 3:

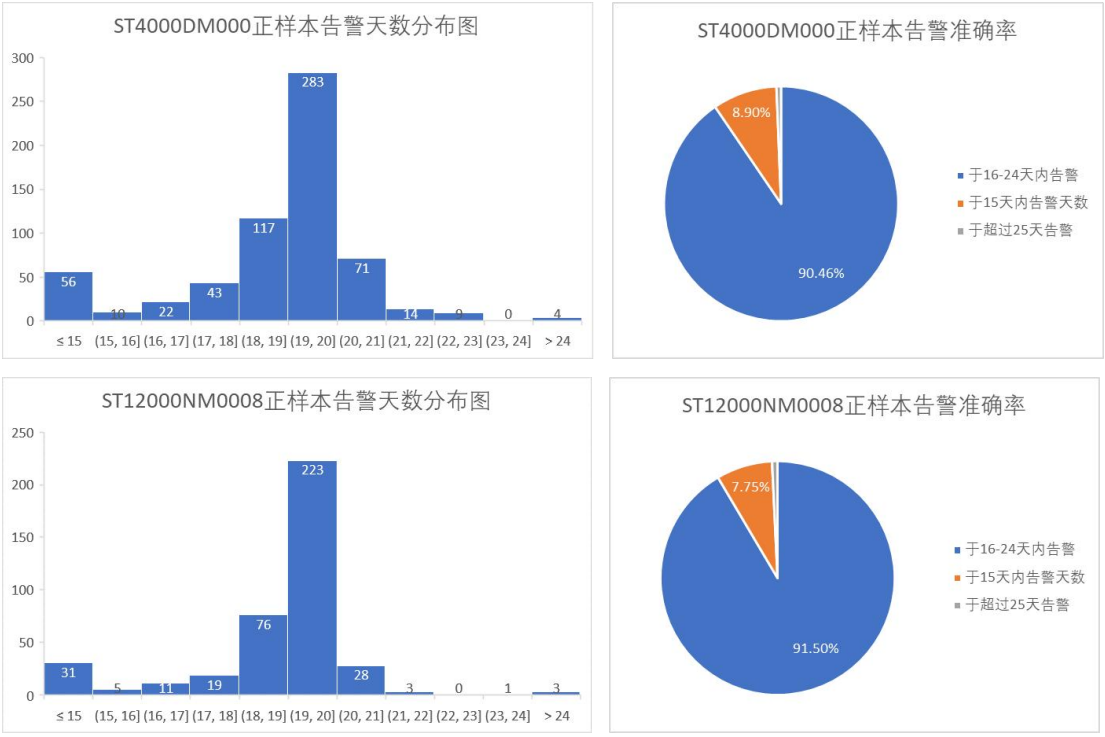

图 3- 3 正样本测试

结果表明，XGBoost 的告警天数误差在 5 天以内的准确率可以达到的 90%，即意味着绝大部分通常方式损坏的硬盘，均可以提前 20

天左右由 XGBoost 模型进行告警。

负样本的结果如下图 3- 4 所示:

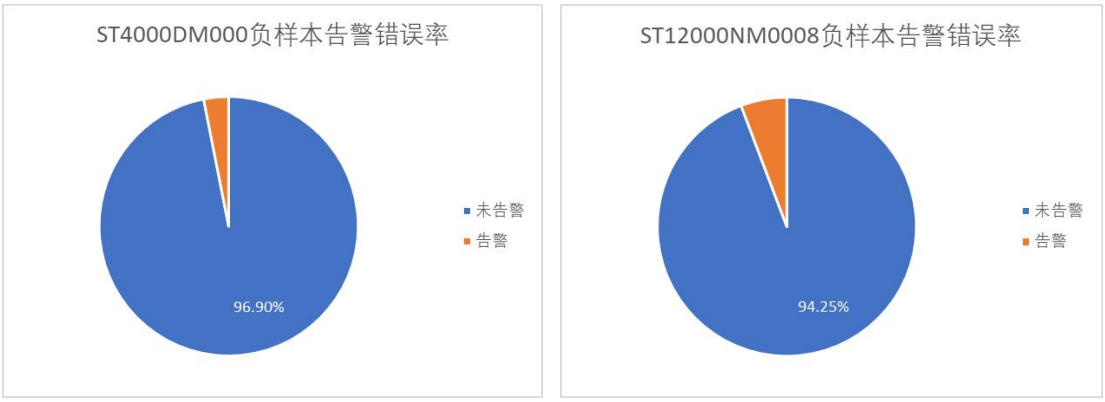

图 3- 4 负样本测试

可以看到，负样本意外告警的比率在 5%左右，即 XGBoost 发生误报的可能性比较低，使用 XGBoost 模型进行告警并不太可能造成不必要的浪费。

综上所述，XGBoost 在实际投产环境中，对硬盘告警有较高的准确率，是可以作为一个告警依据的。

3.2 硬盘寿命预测

硬盘寿命预测基于 RNN 模型实现。将上述 XGBoost 中使用的正样本，数据作为样本数据，并按照 9：1 的比例划分训练集和测试集，输入进一个含有 3 层隐藏层，并在最后有一个全连接层的 RNN 网络，输入层则取决于具体的 SMART 变量数。

对于该 RNN 模型，其公式化的函数形式为:

$$\boldsymbol{Y_t} = f(\boldsymbol{X_{SMART,t}})$$

其中$X_{SMART,t}$为每 20 天的各天 Smart 值，实际上是一个 2 维矩阵，

Y_t为预测的寿命，为一个向量。

RNN 模型的具体参数如下表 3- 3 所示:

表 3- 3RNN 模型参数

型号	ST4000DM000	ST12000NM0008
input_size	21	19
hidden_size	20	20
learning_rate	0.01	0.01
layer	3	3
dropout	0.2	0.2
bidirectional	TRUE	TRUE
Optimizer	SGD	SGD
Loss Function	MSE	MSE

训练情况如图 3- 3 所示，左图为 ST4000DM000 型号的训练情况，右图则为 ST12000NM0008 型号的训练情况，我们选取测试集 Loss 最小的作为投产模型使用。

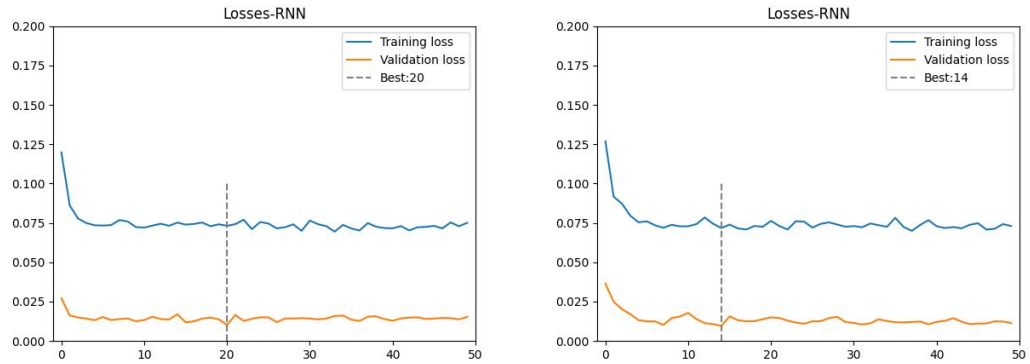

图 3- 3 RNN 训练 Loss 变化图

在实际投产时，我们将告警硬盘产生的最近 20 天的 SMART 数据输入进训练好的模型，模型将给出未来 20 天的寿命预期。

但在实际投产时，我们认为该模型的效果并不是很理想，换用理论上更能保留长序列特征的 LSTM 模型训练效果甚至不如 RNN，为此，我们正在设计新的解决方案，因此有可能会在未来进一步修正该模块，以期达到更好的效果。

故障告警和寿命预测在用户界面为一个综合模块——硬盘故障诊断，如下图 3- 5 所示：

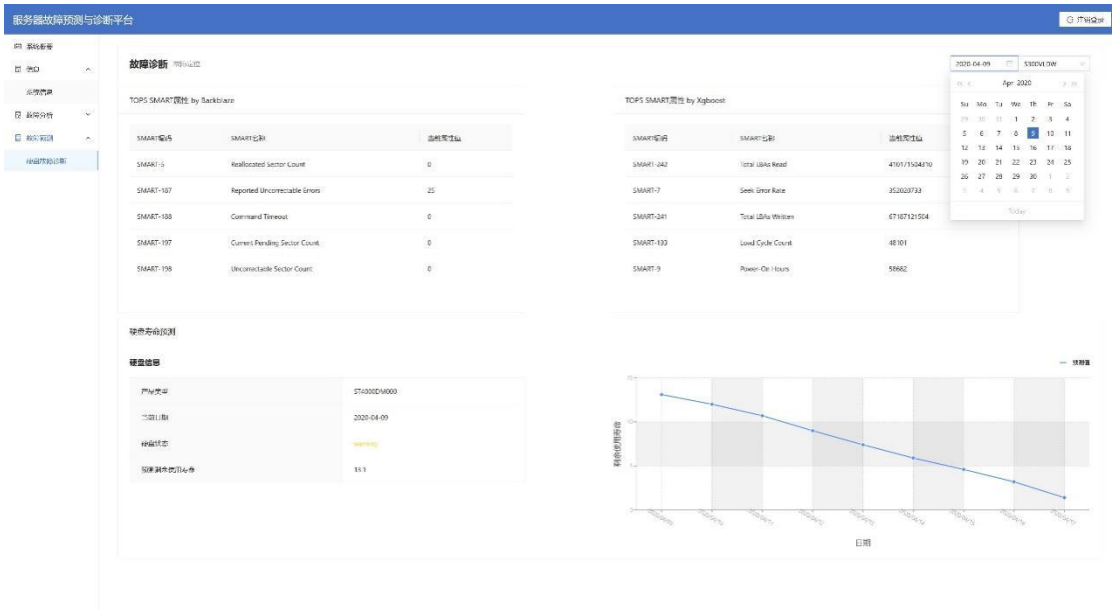

图 3- 5 硬盘故障诊断模块效果图

具体来说，我们会展示 2 个部分，分别是 SMART 数值，和寿命预测。

SMART 数值分成两个部分：Backblaze 推荐的属性，和 XGBoost 识别的属性：

TOP5 SMART属性 by Backblaze

SMART编码	SMART名称	当前属性值
SMART-5	Reallocated Sector Count	0
SMART-187	Reported Uncorrectable Errors	47
SMART-188	Command Timeout	0
SMART-197	Current Pending Sector Count	32
SMART-198	Uncorrectable Sector Count	32

图 3- 6Backblaze 推荐的 SMART 属性

TOP5 SMART属性 by Xgboost

SMART编码	SMART名称	当前属性值
SMART-242	Total LBAs Read	425187246004
SMART-7	Seek Error Rate	415548673
SMART-241	Total LBAs Written	67747247104
SMART-193	Load Cycle Count	49915
SMART-9	Power-On Hours	58680

图 3- 7XGBoost 识别的 SMART 属性

大部分情况下，Backblaze 推荐的属性并没有明显的变化，但 XGBoost 推荐的属性变化比较明显，我们为了方便用户比对效果，将两组 SMART 值均呈现。

而硬盘寿命预测部分，左侧为 XGBoost 的告警结果，如下图 3- 8 所示，硬盘状态已为 warning，并给出了预测的剩余使用寿命。如果硬盘状态为 normal，则剩余使用寿命是 N/A。

硬盘寿命预测

硬盘信息

产品类型	ST4000DM000
当前日期	2020-04-09
硬盘状态	warning
预测剩余使用寿命	13.3

图 3- 8XGBoost 告警结果

而右侧为 RNN 对硬盘使用寿命的预测，如下图 3- 9 所示，大多数情况下，硬盘的使用寿命在 XGBoost 告警后，都呈现一种线性下滑的趋势，并在大约 20 天后达到 0。

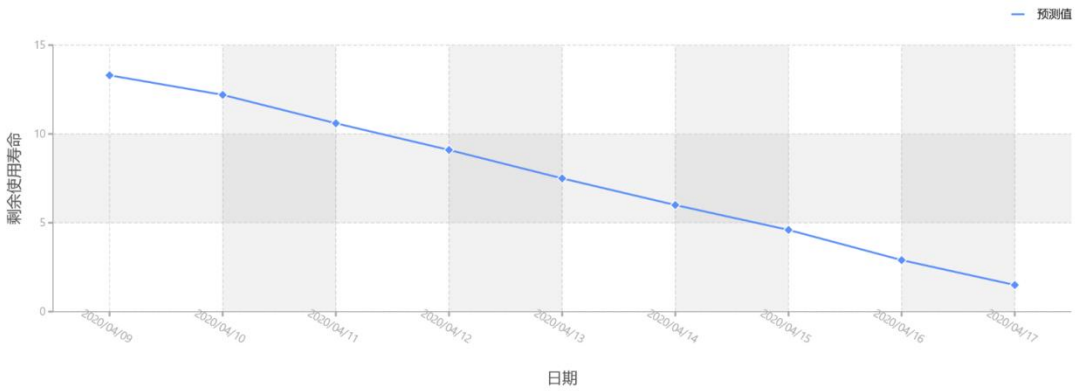

图 3- 9RNN 寿命预测结果

在未来，如果 LSTM 和 RNN 有更好的体验的话，我们会向用户提供选项，用户可以自行选择使用 RNN 还是 LSTM 进行寿命预测。