SRE学堂：OSS监控告警案例分析

原创

阿里云开发者 于 2021-01-15 13:49:29 发布 1555 收藏 1

文章标签： 资源调度 监控 网络协议 前端开发 网络安全 tsar 对象存储 Perl 容器

本文链接：https://blog.csdn.net/alitech2017/article/details/112663371

简介：【SRE学堂】OSS从入门到精通第四章：OSS监控告警案例分析处理

阿里云
阿里云智能GTS-平台技术部-SRE混合云技术服务赋能团队

通过OSS第一章的学习，大家知道了OSS是什么、OSS的各项优势、OSS的架构组成以及OSS相关的一些基本概念；通过OSS的第二章的学习，大家知道了如何创建、查看、删除、修改OSS的存储空间，如何实现object上传、下载、删除的操作；通过OSS第三章的学习，大家知道了OSS如何执行天基的终态检查、赤骥白屏巡检以及黑屏巡检；第四章将为大家介绍OSS监控TAC以及OSS集群出现大量5xx错误时该如何分析。

1. OSS监控TAC

使用域名登录至TAC报警中心，查看OSS的监控能力、告警详情，还需要关注盯屏及钉钉群，当TAC报警中心接收到报警之后会向钉钉群推送告警通知，可根据报警信息查看和分析问题。

1.1 TAC中查看OSS的监控能力

登录报警中心，选择系统>监控管理，查看OSS的告警监控项。

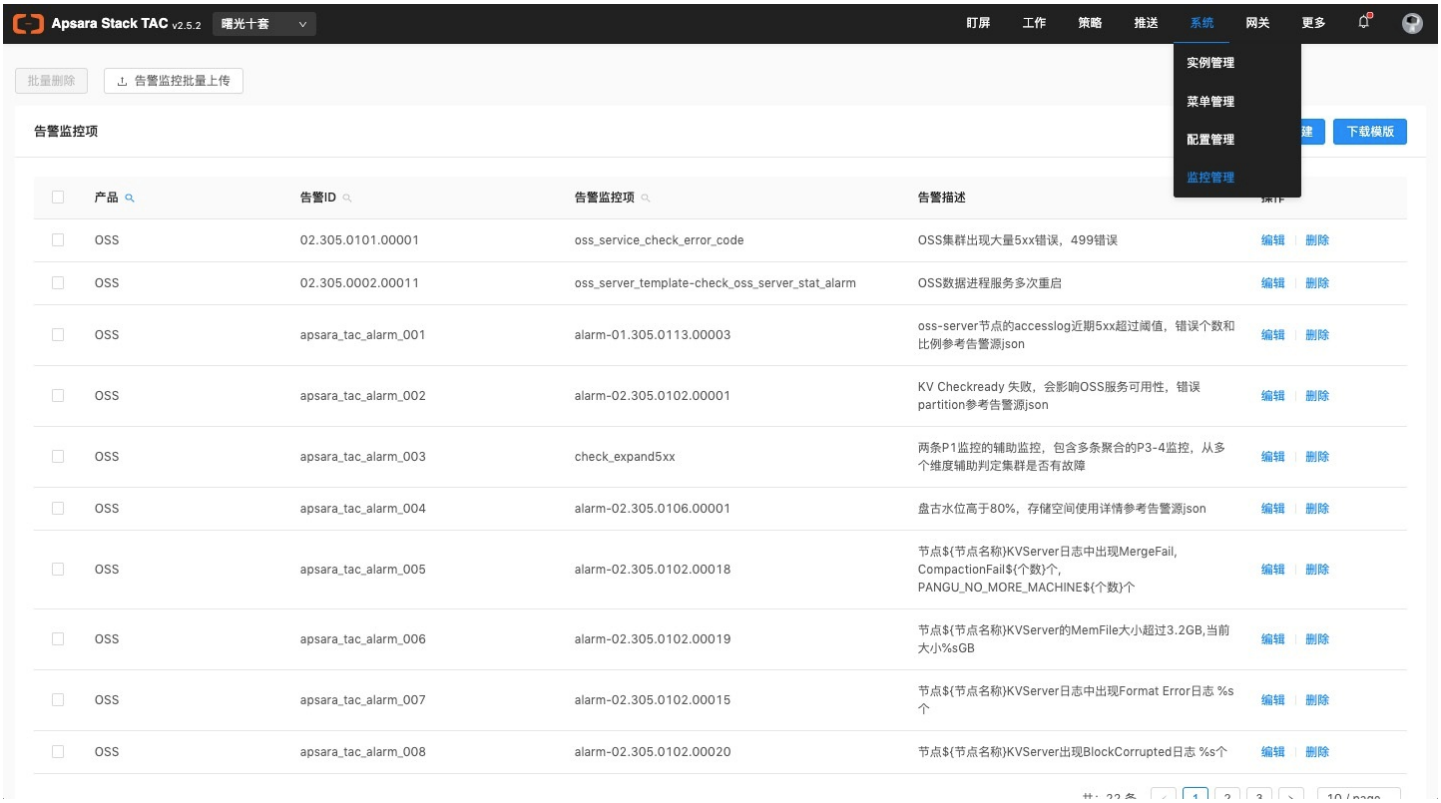

图1：监控管理

1.2 TAC中查看告警详情

登录报警中心，选择工作>平台告警，查看OSS的平台告警详情。

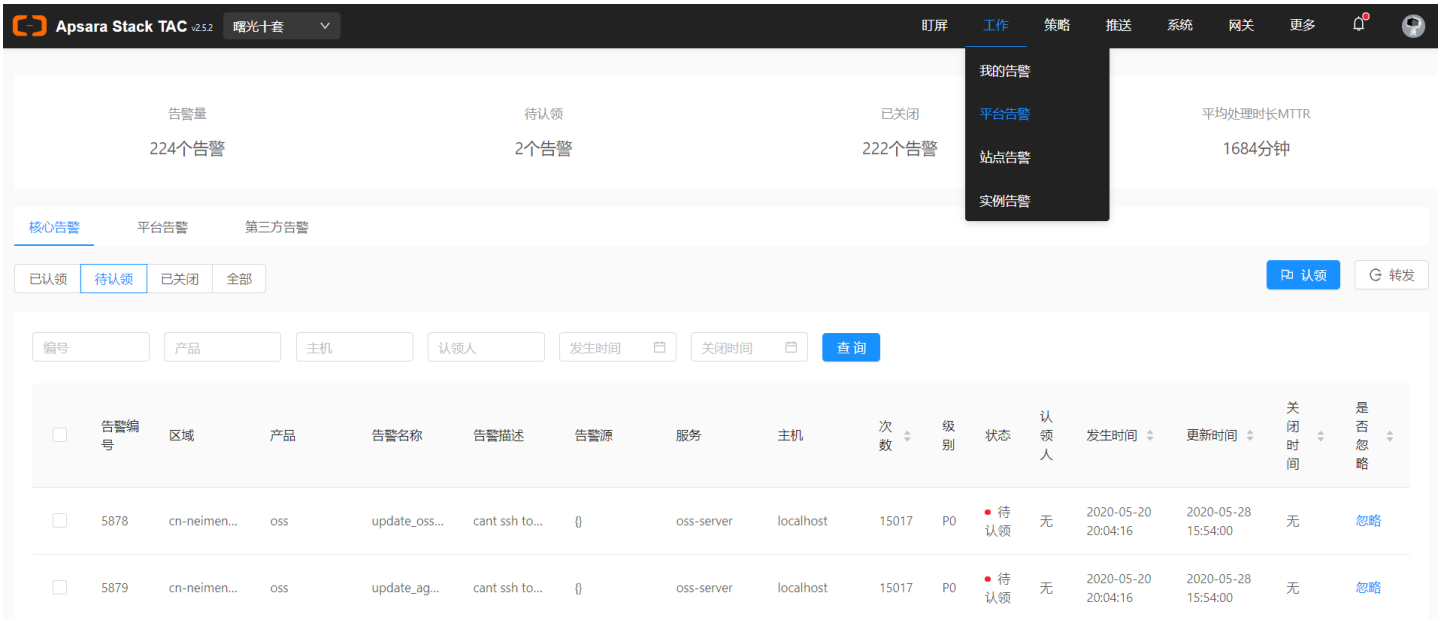

图2：平台告警详情

1.3 TAC中的盯屏

登录报警中心，选择盯屏，查看盯屏中相关告警信息。

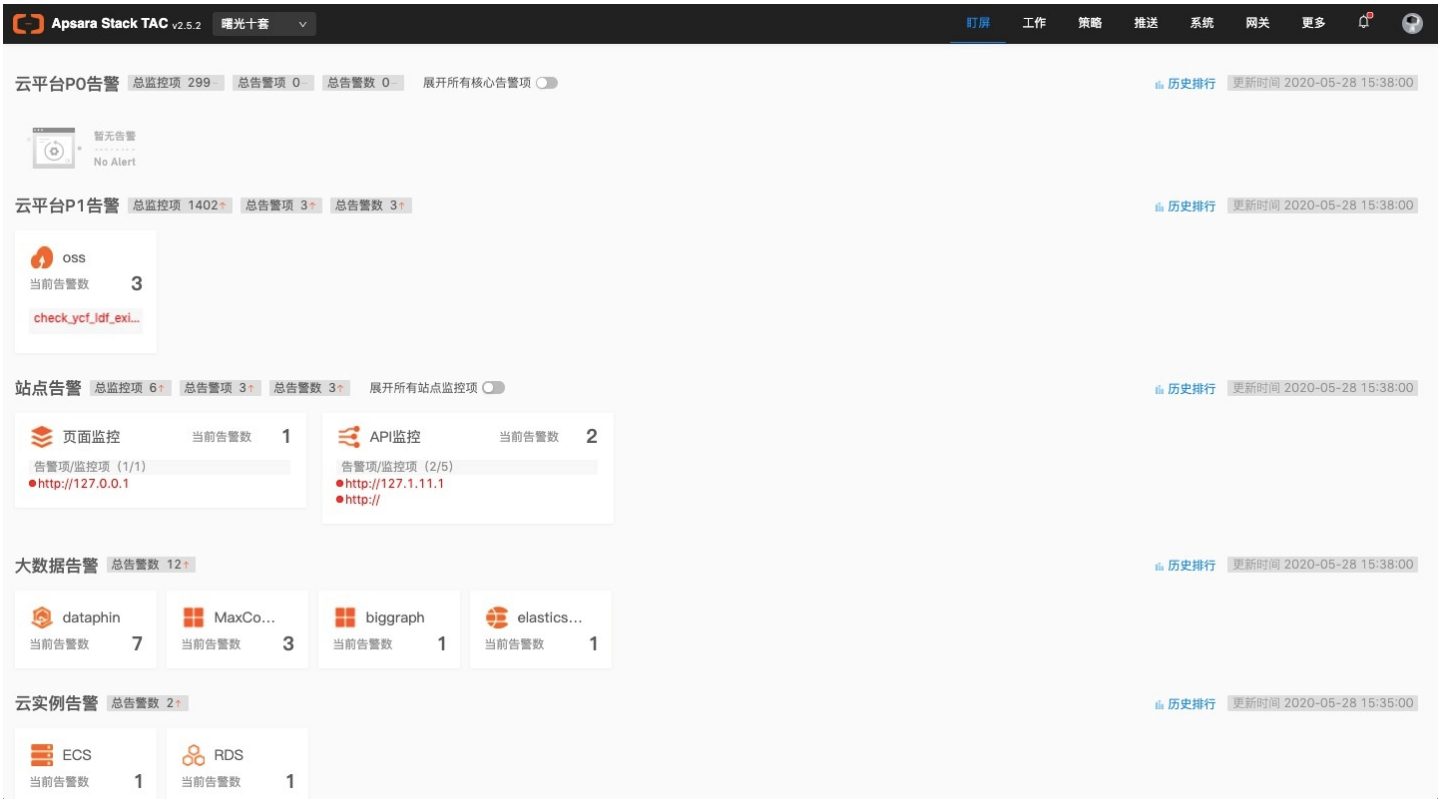

图3：盯屏

1.4 钉钉群中收到的TAC告警

登录钉钉，选择TAC告警通知群，查看TAC告警通知。

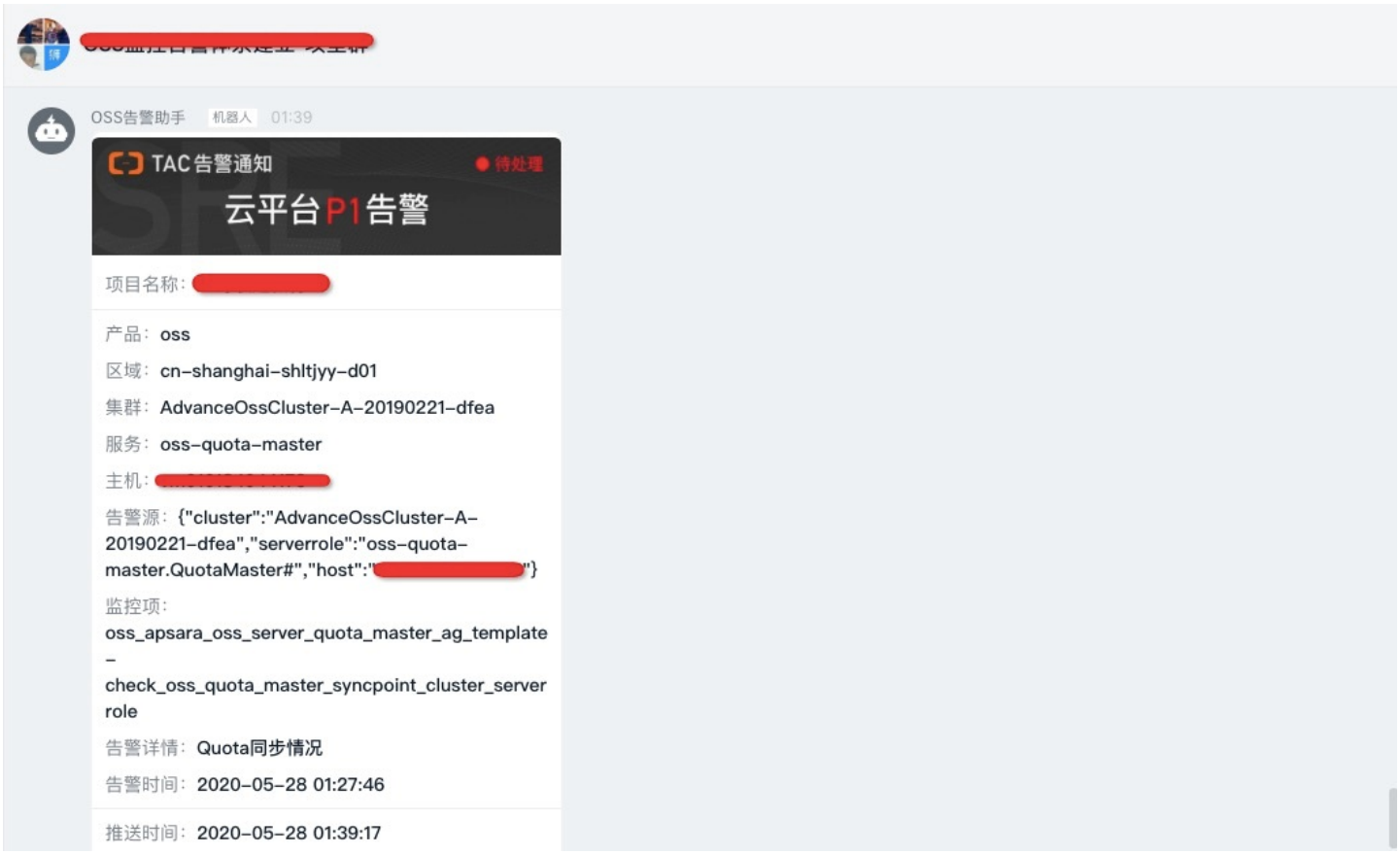

图4：TAC告警通知群

2. OSS集群出现大量5xx错误分析

因OSS集群内部错误导致资源无法访问时，会出现5xx错误。当铜雀、TAC、ASO或者封神产品收到告警信息时，可以通过人肉或者赤骥工具来定位问题，具体的定位方式可以参考5xx预案的脑图，后续会将其中的重点内容进行详细讲解。

2.1 5xx预案脑图

OSS的5xx预案可以从三个方面来详细说明，分别是：告警来源、人肉定位思路、赤骥定位思路，具体内容见下图。

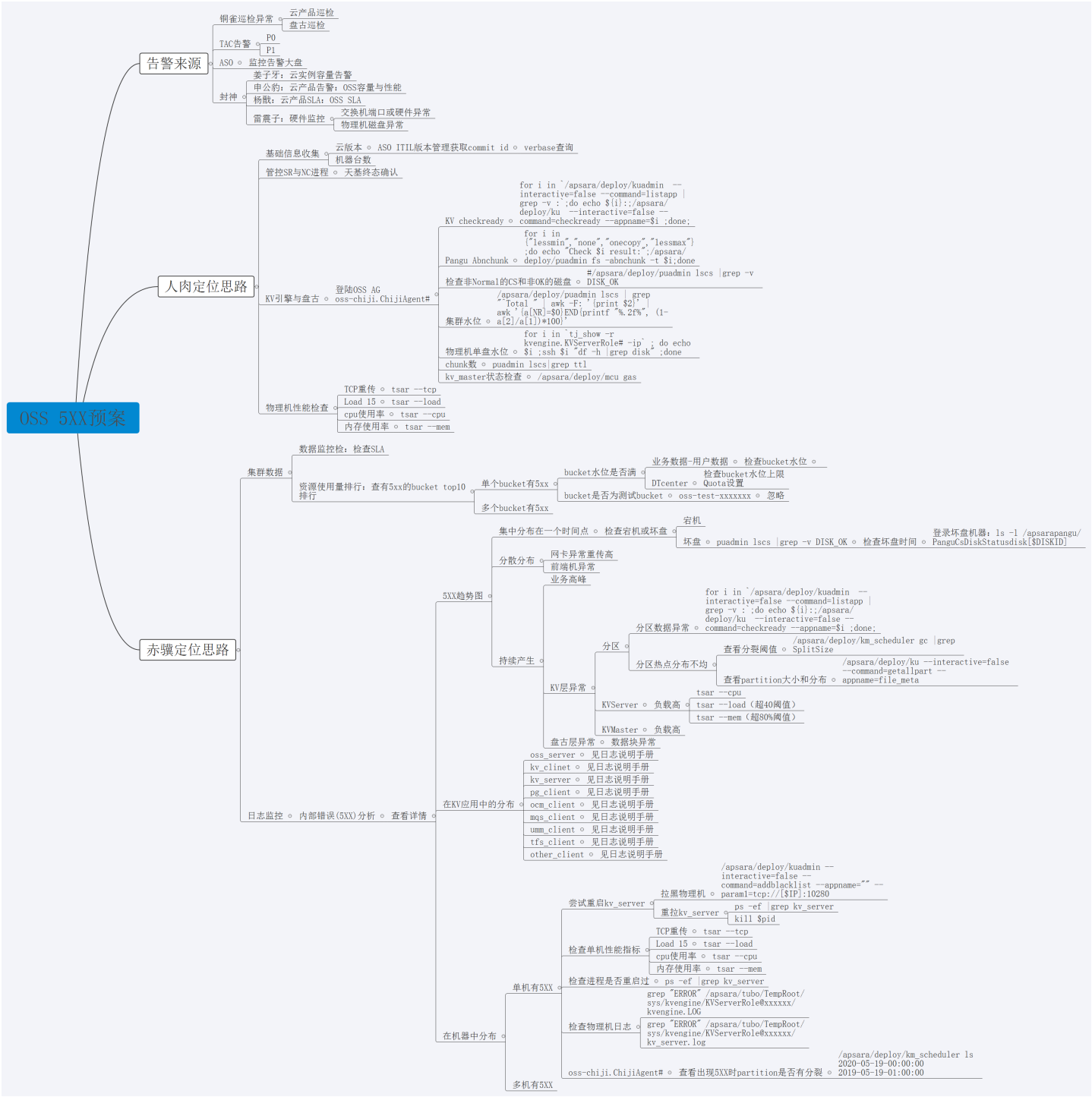

图5：5xx预案脑图

2.2 5xx错误的查看

在OSS赤骥页面选择业务数据>集群数据>集群概览，查看今日5xx总数是否为0，如果不为0，且该数字大于0.01%，会产生TAC告警。

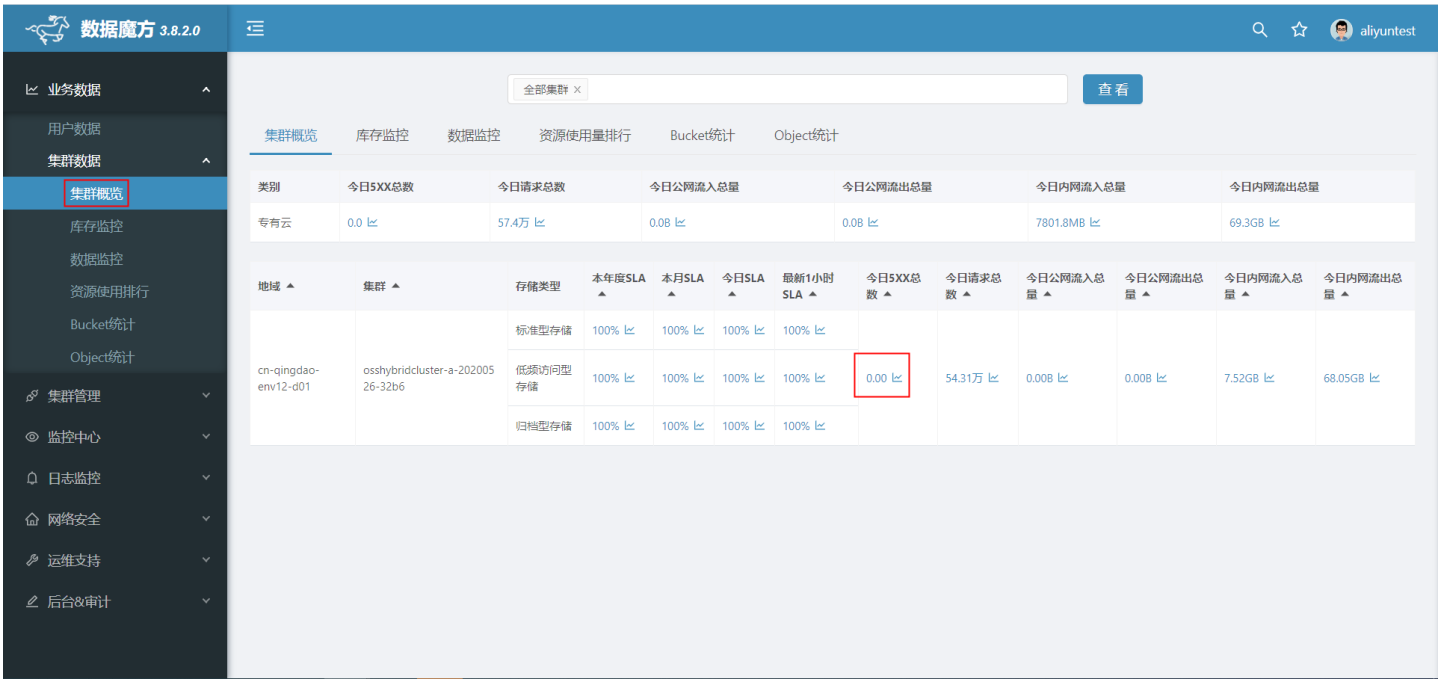

图6：集群概览

单击上图中的百分比，查看5xx错误的时间分布，并判断是持续出现还是间隔出现；

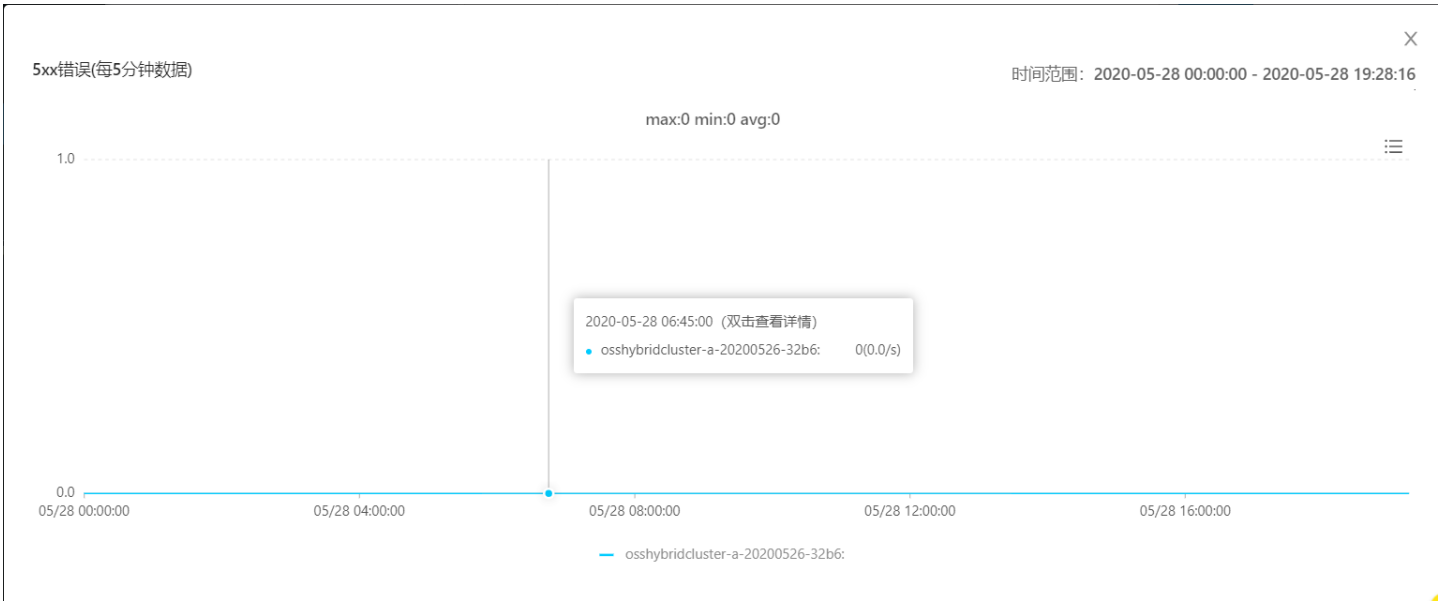

图7:5xx错误的时间分布

如果间隔出现，一般无需排查问题，例如在某时间点磁盘损坏，而数据刚好落在该坏盘上，造成数据无法访问，因此导致5xx错误；再例如一台机器损坏，也可能导致某时间点产生大量的5xx错误，此时会触发OSS中自动拉黑已坏机器的机制，从而不再让请求访问该机器，5xx问题便会自动消退； 如果持续产生，可能涉及盘古和KV深层次的问题，需要研发介入处理恢复，此处不再赘述。

当OSS出现5xx问题时，首先要做一些基础检查，然后按照赤骥定位和人肉定位的标准检查流程逐步分析排查问题。具体涉及哪些基础检查，又该如何去逐步排查呢？请看下文描述。

3. 赤骥定位思路

3.1 检查SLA曲线是否有波动

在OSS赤骥页面选择业务数据>集群数据>数据监控，检查一天SLA曲线。值得注意的是5xx问题出现不一定有SLA波动，有SLA波动一般都会有5xx问题。如果SLA持续波动或不满100%，5XX持续产生，可能的原因有：集群压力大（机器cpu、mem、硬盘使用率高等）、机器硬件故障（如网络设备故障导致的重传高等）、服务异常（oss_server异常、kv_master异常、kv_server hang等情况）、ycf文件损坏等。

图8：SLA曲线图

3.2 检查资源使用排行top 10

在OSS赤骥页面选择业务数据>集群数据>资源使用排行，检查有5xx的bucket top10排行，其中有类似osstestcn的测试bucket，测试bucket在测试之后易被删除一些信息，再次访问可导致5xx错误，该情况不会对真正的服务产生影响，无需关注。如果正式使用的bucket产生大量5xx，可能该bucket中某个文件被损坏，或者该bucket中设置了quota值，容量超限（quota值可在dt上查看）。

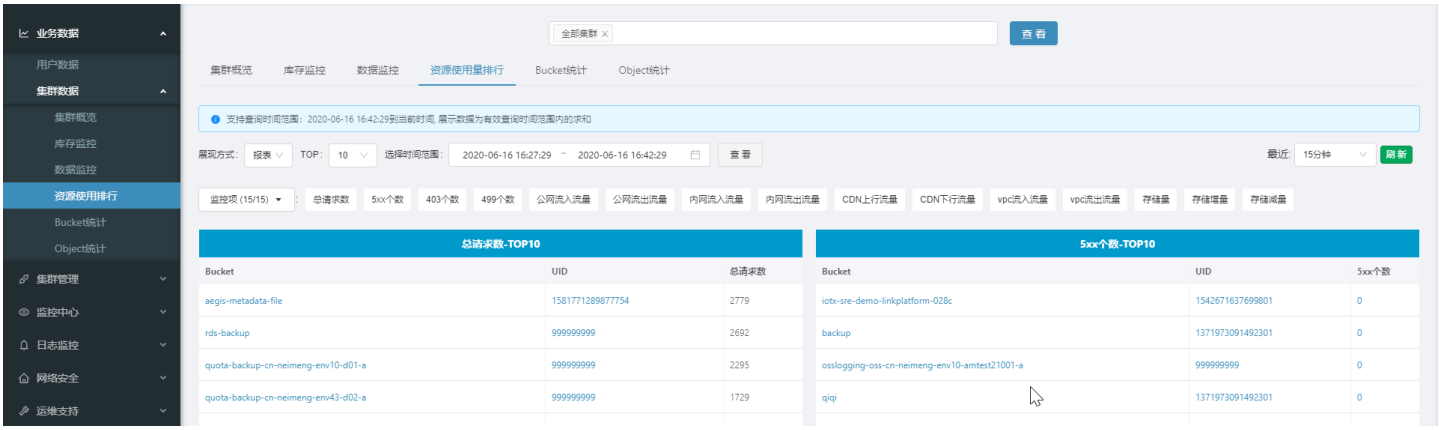

图9：资源使用排行

3.3 查看内部错误（5xx）分析详情

在OSS赤骥页面选择日志监控>内部错误（5xx）分析>查看详情，查看详细说明，包括报错的服务、机器、bucket分布等情况，可查看不同应用错误数占比及相关应用的具体日志分析，例如：
oss_server是前端机，可查看5xx的数量及其在bucket上的分布情况，确认是否为网络问题；
kv_client可查看5xx的数量及其在机器上的分布情况，然后确认是否为Youchaofile的问题；
kv_server可查看5xx在partition上的分布情况。如果5xx集中出现在一台物理机上，可以尝试拉黑该物理机，然后重启kv_server。如果不再产生，需排查该机器问题，确认该机器上的partition是否损坏亦或存在硬件问题等。拉黑操作命令如下：

```
#/apsara/deploy/kuadmin --interactive=false --command=addblacklist --appname="" --param1=tcp://[$IP]:10280
```

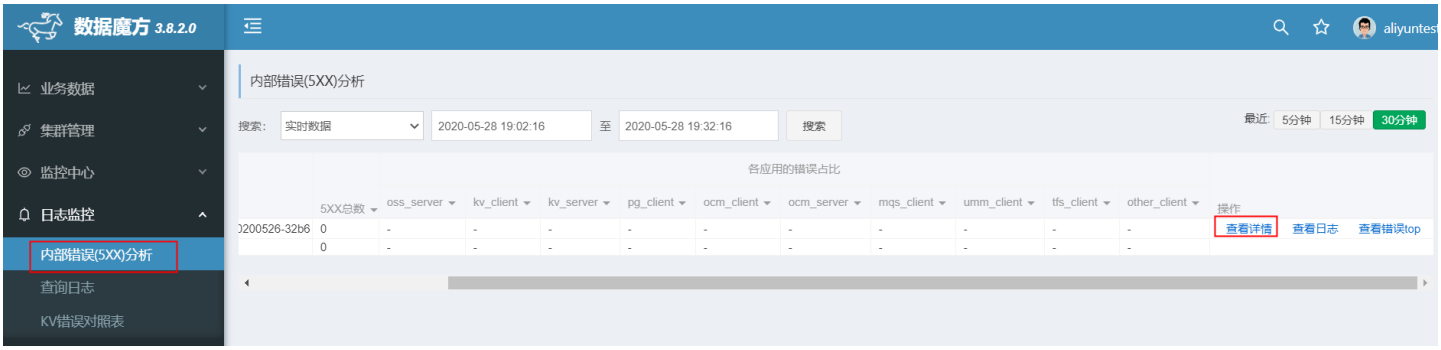

图10：5xx内部错误分析

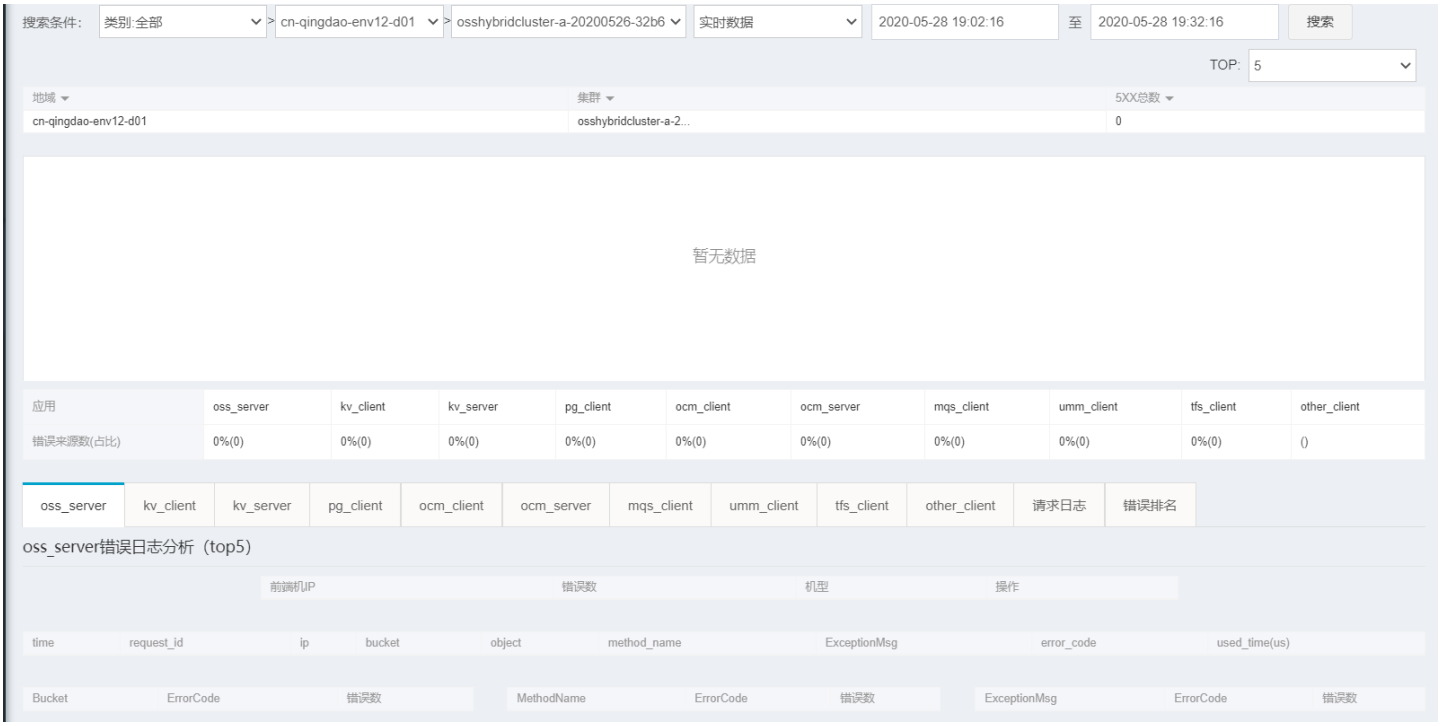

图11：5xx内部错误分析详情

3.4 业务数据中查看有5xx错误的bucket信息

在OSS赤骥页面选择业务数据>用户数据，输入存在5xx错误的bucket，查看该bucket的基础信息和数据监控。

图12：bucket基础信息和数据监控

4. 人肉定位思路

4.1 天基中查看服务终态

在天基中检查oss-server、kvengine、pangu、fuxi、nuwa服务是否到达终态。如果未达终态需要具体检查服务是否可用等。

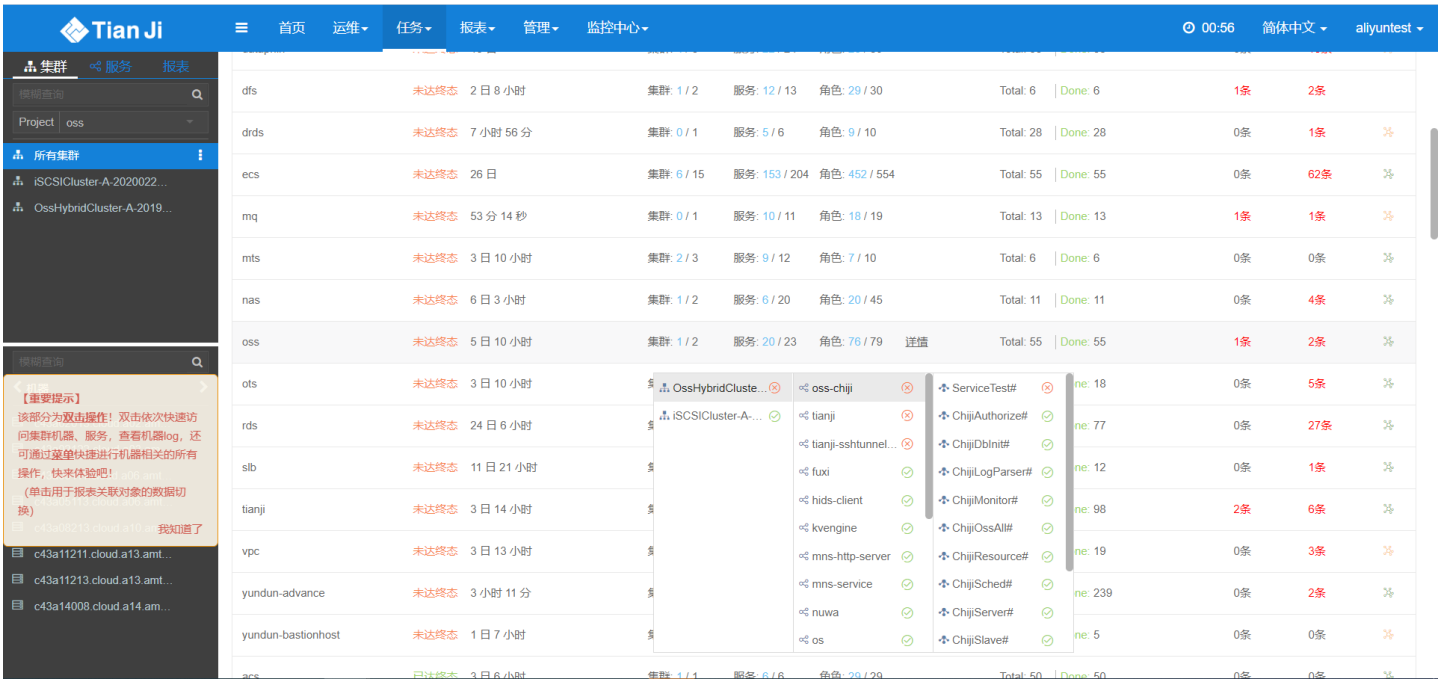

图13：OSS集群天基终态检查

4.2 检查盘古水位和物理机上的单盘水位是否超过90%

登录OSS集群的oss-chiji-agent.ChijiAgent#所在vm，以admin用户的身份执行以下命令检查盘古水位和物理机上的单盘水位，查看是否超过90%，如果超过会触发禁写，从而导致数据无法写入，产生5xx错误。

```
# /apsara/deploy/puadmin lscs | grep "^Total " | awk -F: '{print $2}' | awk '{a[NR]=$0}END{printf "%.2f%", (1-a[2]/a[1])*100}’
```

```
# for i in tj_show -r kvengine.KVServerRole# -ip ; do echo $i ;ssh $i "df -h |grep disk" ;done
```

4.3 检查cs和磁盘状态是否有非NORMAL和非DISK_OK状态

执行以下命令检查cs和磁盘的状态，判断是否存在非NORMAL的机器和非DISK_OK的磁盘，该情况危害相对较小，因其只在挂掉的时间段产生5xx。

```
# /apsara/deploy/puadmin lscs |grep -v DISK_OK
```

4.4 检查网络重传值及网卡丢包情况

执行以下命令检查网络状态，包括近24小时和1小时的重传值以及网卡丢包情况。如果重传值大于0.8，此时需至对应的盘古机器确认是否持续大于0.8；如果不是，无需关注；如果是，可能网卡、交换机等设备异常，此时需要网络小组协助排查处理。

```
#for i in tj_show -r kvengine.KVServerRole# -ip ; do echo $i ; ssh $i "tsar --tcp |tail -n 3 " |awk '{print $9}' ; done
#for i in tj_show -r kvengine.KVServerRole# -ip ; do echo $i ; ssh $i "tsar --tcp --watch 60 |tail -n 3 " |awk '{print $9}' ; done
# for i in tj_show -r kvengine.KVServerRole# -ip ; do echo $i ; ssh $i "ifconfig |grep -A 7 bond0 |grep errors"
```

4.5 检查abnchunk是否有nonecopy

执行以下命令做abnchunk检查，重点关注是否有nonecopy情况发生，如果有说明文件丢失。

```
# for i in {"lessmin","none","onecopy"};do echo "Check $i result:";/apsara/deploy/puadmin fs -abnchunk -t $i;done
```

```
Check lessmin result:
Master Address: nuwa://localcluster/sys/pangu/master
ChunkId Type  FoundTime

Check none result:
Master Address: nuwa://localcluster/sys/pangu/master
ChunkId Type  FoundTime

Check onecopy result:
Master Address: nuwa://localcluster/sys/pangu/master
ChunkId Type  FoundTime
```

图14：检查abnchunk

4.6 检查checkready中partition是否均为YES

执行以下命令做checkready检查，关注各个app的partition是否为YES，如果为YES则正常，如果为非YES，则会出现5xx的错误，此时大多数情况是kv_server出现问题，需要登录至tongque容器，执行脚本检查partition异常的物理机。

以下命令chijiagent#执行，检查checkready是否异常。

```
# for i in /apsara/deploy/kuadmin --interactive=false --command=listapp | grep -v : ;do echo ${i}:;/apsara/deploy/ku --interactive=false --command=checkready --appname=$i ;done;
```

如果存在异常，在安装了tpcmon的铜雀容器执行以下命令，确认part异常的机器。

```
#/home/tops/bin/python2.7 OssTacMonitor.py --opsnode OPS#2 --alarminfo AlarmInfoKeyKvReady300s.json
```

5. 结语

《OSS从入门到精通》一共四个章节，到此为止就告一段落了，感谢大家一路的陪伴，我们后续还会有其他产品的课程，欢迎大家关注我们的SRE学堂。

往期内容

《OSS从入门到精通》第三章：OSS深度巡检_03巡检异常处理案例解析
《OSS从入门到精通》第三章：OSS深度巡检_02快速吃透黑屏巡检那些事儿
《OSS从入门到精通》第三章：OSS深度巡检_01细说白屏巡检的方方面面
《OSS从入门到精通》第二章：OSS使用及常见操作
《OSS从入门到精通》第一章：OSS产品综述

我们是阿里云智能全球技术服务-SRE团队，我们致力成为一个以技术为基础、面向服务、保障业务系统高可用的工程师团队；提供专业、体系化的SRE服务，帮助广大客户更好地使用云、基于云构建更加稳定可靠的业务系统，提升业务稳定性。我们期望能够分享更多帮助企业客户上云、用好云，让客户云上业务运行更加稳定可靠的技术，您可用钉钉扫描下方二维码，加入阿里云SRE技术学院钉钉圈子，和更多云上人交流关于云平台的那些事。

原文链接：https://developer.aliyun.com/article/780507?