AWS 的 S3 故障回顾和思考

2017年03月03日 (Http://coolshell.cn/articles/17737.html) 陈皓 (Http://coolshell.cn/articles/author/haoel) 15,574 人阅读

继Gitlab的误删除数据事件(http://coolshell.cn/articles/17680.html)没几天，“不沉航母”AWS S3（Simple Storage Service）几天前也“沉”了4个小时，墙外的半个互联网也跟着挂了。如约，按AWS惯例，AWS今天给出了一个简单的故障报告《Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region (https://aws.amazon.com/cn/message/41926/)》。这个故障和简单来说和Gitlab一样，也是人员误操作。先简单的说一下这份报中说了什么。

故障原因

简单来说，这天，有一个 AWS 工程师在调查 Northern Virginia (US-EAST-1) Region 上 S3 的一个和账务系统相关的问

题，这个问题是S3的账务系统变慢了（我估计这个故障在Amazon里可能是Sev2级，Sev2级的故障在Amazon算是比较大的故障，需要很快解决），Oncall的开发工程师（注：Amazon的运维都是由开发工程师来干的，所以Amazon内部嬉称SDE-Software Developer Engineer 为 Someone Do Everything）想移除一个账务系统里的一个子系统下的一些少量的服务器（估计这些服务器上有问题，所以想移掉后重新部署），结果呢，有一条命令搞错了，导致了移除了大量的S3的控制系统。包括两个很重要的子系统：

> 1）一个是S3的对象索引服务（Index），其中存储了S3对象的metadata和位置信息。这个服务也提供了所有的GET，LIST，PUT 和DELETE请求。
>
> 2）一个是S3的位置服务系统（Placement），这个服务提供对象的存储位置和索引服务的系统。这个系统主要是用于处理PUT新对象请求。

这就是为什么S3不可访问的原因。

在后面，AWS也说明了一下故障恢复的过程，其中重点提到了这点——

> 虽然整个S3的是做过充分的故障设计的（注：AWS的七大Design Principle 之一 Design for Failure）—— 就算是最核心的组件或服务出问题了，系统也能恢复。但是，可能是在过去的日子里 S3 太稳定了，所以，AWS 在很长很长一段时间内都没有重启过 S3 的核心服务，而过去这几年，S3 的数据对象存储级数级的成长（S3存了什么样数量级的对象，因为在Amazon工作过，所以多大概知道是个什么数量级，这里不能说，不过，老实说，很惊人的），所以，这两个核心服务在启动时要重建并校验对象索引元数据的完整性，这个过程没想到花了这么长的时候。而Placement服务系统依赖于Index 服务，所以花了更长的时间。

了解过系统底层的技术人员应该都知道这两个服务有多重要，简而言之，这两个系统就像是Unix/Linux文件系统中的inode，或是像HDFS里的node name，如果这些元数据丢失，那么，用户的所有数据基本上来说就等于全丢了。

而要恢复索引系统，就像你的操作系统从异常关机后启动，文件系统要做系统自检那样，硬盘越大，文件越多，这个过程就越慢。

另外，这次，AWS没有使用像以前那样 Outage 的故障名称，用的是 “Increased Error Rate” 这样的东西。我估计是没有把所有这两个服务删除完，估计有些用户是可以用的，有的用户是则不行了。

后续改进

在这篇故障简报中，AWS 也提到了下面的这些改进措施——

1）改进运维操作工具。对于此次故障的运维工具，有下面改进：

- 让删除服务这个操作变慢一些（陈皓注：这样错了也可以有时间反悔，相对于一个大规模的分布式系统，这招还是很不错的，至少在系统报警时有也可以挽救）
- 加上一个最小资源数限制的SafeGuard（陈皓注：就是说，任何服务在运行时都应该有一个最小资源数，分布式集群控制系统会强行维护服务正常运行的最小的一个资源数）
- 举一反三，Review所有和其它的运维工具，保证他们也相关的检查。

2）改进恢复过程。对于恢复时间过长的问题，有如下改进：

- 分解现有厚重的重要服务成更小的单元（在 AWS，Service是大服务，小服务被称之为 Cell），AWS 会把这几个重要的服务重构成 Cell服务。（陈皓注：这应该就是所谓的“微服务”了吧）。这样，服务粒度变小，重启也会快一些，而且还可以减少故障面（原文：blast radius – 爆炸半径）
- 今年内完成对 Index 索引服务的分区计划。

相关思考

下面是我对这一故障的相关思考——

0）太喜欢像Gitlab和AWS这样的故障公开了，那怕是一个自己人为的低级错误。不掩盖，不文过饰非，透明且诚恳。Cool!

1）这次事件，还好没有丢失这么重要的数据，不然的话，将是灾难性的。

2）另外，面对在 US-EASE-1 这个老牌 Region 上的海量的对象，而且能在几个小时内恢复，很不容易了。

3）这个事件，再次映证了我在《关于高可用的系统 (http://coolshell.cn/articles/17459.html)》中提到的观点：一个系统的高可用的因素很多，不仅仅只是系统架构，更重要的是——高可用运维。

4）对于高可用的运维，平时的故障演习是很重要的。AWS 平时应该没有相应的故障演习，所以导致要么长期不出故障，一出就出个大的让你措手不及。这点，Facebook就好一些，他们每个季度扔个骰子，随机关掉一个IDC一天。Netflix 也有相关的 Chaos Monkey，我以前在的路透每年也会做一次大规模的故障演练——灾难演习。

5）AWS对于后续的改进可以看出他的技术范儿。可以看到其改进方案是用技术让自己的系统更为的高可用。然后，对比国内的公司对于这样的故障，基本上会是下面这样的画风：

a）加上更多更为严格的变更和审批流程，

b）使用限制更多的权限系统和审批系统

c）使用更多的人来干活（一个人干事，另一个人在旁边看）

d）使用更为厚重的测试和发布过程

e）惩罚故障人，用价值观教育工程师。

这还是我老生长谈的那句话——如果你是一个技术公司，你就会更多的相信技术而不是管理。相信技术会用技术来解决问题，相信管理，那就只会有制度、流程和价值观来解决问题。（注意：这里我并没有隔离技术和管理，只是更为倾向于用技术解决问题）

最后，你是要建一个“高可用的技术系统”，还是一个“高用的管理系统”？;-)

（全文完）

关注CoolShell微信公众账号可以在手机端搜索文章

——=== 访问 酷壳404页面 (http://coolshell.cn/404/) 寻找遗失儿童。 ===——

业界新闻 (Http://coolshell.cn/category/itnews), 杂项资源 (Http://coolshell.cn/category/misc), 程序设计 (Http://coolshell.cn/category/progdesign)
Amazon S3 (Http://coolshell.cn/tag/amazon-s3), AWS (Http://coolshell.cn/tag/aws), Design (Http://coolshell.cn/tag/design), High Availability (Http://coolshell.cn/tag/high-availability)

相关文章

- 2011年04月27日 关于Amazon云宕机的网贴收集 (http://coolshell.cn/articles/4601.html)
- 2017年02月02日 从Gitlab误删除数据库想到的 (http://coolshell.cn/articles/17680.html)
- 2016年08月21日 关于高可用的系统 (http://coolshell.cn/articles/17459.html)
- 2009年04月12日 9个强大免费的PHP库 (http://coolshell.cn/articles/455.html)
- 2010年10月18日 一些非常不错的资料 (http://coolshell.cn/articles/3192.html)
- 2012年03月09日 Bret Victor – Inventing on Principle (http://coolshell.cn/articles/6775.html)
- 2011年09月08日 千万不要把 bool 设计成函数参数 (http://coolshell.cn/articles/5444.html)
- 2012年03月13日 多版本并发控制(MVCC)在分布式系统中的应用 (http://coolshell.cn/articles/6790.html)

《AWS 的 S3 故障回顾和思考》的相关评论


哈士奇说道：

2017年03月03日 14:44 (http://coolshell.cn/articles/17737.html#comment-1913343)
沙发


康斯坦丁说道：

2017年03月03日 14:47 (http://coolshell.cn/articles/17737.html#comment-1913344)
2楼


Chunyang (http://chunyang-wen.github.io)说道：

2017年03月03日 14:50 (http://coolshell.cn/articles/17737.html#comment-1913345)
赞公开解决过程。不掩饰自己的缺点。


Weizw说道：

2017年03月03日 14:51 (http://coolshell.cn/articles/17737.html#comment-1913346)
深刻


eason (http://www.fengyingsheng.com)说道：

2017年03月03日 14:56 (http://coolshell.cn/articles/17737.html#comment-1913347)
高可用赞！


观象士 (http://github.com/xusiwei)说道：

2017年03月03日 14:58 (http://coolshell.cn/articles/17737.html#comment-1913348)
最后一句亮了


Jim说道：

2017年03月03日 15:01 (http://coolshell.cn/articles/17737.html#comment-1913349)
技术管理人。这里有个错别字。
所以导致要么长其不出故障


陈皓 (http://coolshell.cn)说道：

2017年03月03日 19:06 (http://coolshell.cn/articles/17737.html#comment-1913361)
谢谢！马上更正！


MT说道：

2017年03月03日 15:24 (http://coolshell.cn/articles/17737.html#comment-1913350)
关于第五点，你确定AWS内部不是按照你说的五点执行的？

亚马逊在美帝的口碑可是比阿里在中国的口碑差很多很多很多的。。。


Malloc说道：

2017年03月03日 16:00 (http://coolshell.cn/articles/17737.html#comment-1913351)
Amazon内部会写COE（Correction of Error），写COE的时候都是对事不对人，所有做错的人都称为The engineer。COE会对整个大组甚至全公司开放，COE里面的Action Item也是对全组来说的。现在S3整个大组氛围不错，应该跟前几年不一样了...

另外，我实在想不出公司还有什么办法惩罚人，能比连续一周oncall，天天半夜三点被page醒更折磨人了。


Malloc说道：

2017年03月03日 16:33 (http://coolshell.cn/articles/17737.html#comment-1913352)
其实还想提一句，虽然不清楚AWS总体，但是org内部普遍认为惩罚事故责任人基本完全没有意义。系统的错误往往来自于团队的工程错误，比如一个设计糟糕的没有design for failure的系统，过短的开发周期导致的short cuts，为什么运维没有足够的自动化措施而需要人肉运维导致犯错；如果是新人那么oncall training是不是没有做好。即使是这个工程师就是个蠢蛋，也要想想hiring bar怎么搞的，为什么招聘的时候没有按照高标准来招聘。其实事故责任人反而是最无辜的。

我觉得喜欢问责事故责任人的公司最后会变成“大家谁也不做事怕犯错，一旦出问题就开始责备别人”的文化。这样的科技公司最后注定会走下坡路的吧。


SizeOf说道：

2017年03月03日 17:51 (http://coolshell.cn/articles/17737.html#comment-1913356)
不能评论了。


SizeOf说道：

2017年03月03日 17:53 (http://coolshell.cn/articles/17737.html#comment-1913358)
不会写代码，就可以去亚马，亚马的招聘门槛那么低，毕竟美帝著名xue han工厂。前一段时间出现了线上笔试两轮直接发offer的情况。题目就是9选3，所以哪怕不会编程，你把9题背下来也是有机会进去的。。。


Malloc说道：

2017年03月04日 02:01 (http://coolshell.cn/articles/17737.html#comment-1913370)
两轮OA给offer现在已经叫停了... 确实是去年的bar太奇怪

最近中文论坛有人给HR举报了群面泄题的事，好像offer也收回了


猪猪侠 (http://YsY5)说道：

2017年03月04日 23:17 (http://coolshell.cn/articles/17737.html#comment-1913377)
亚马逊现在都有35W员工了，需要招很多人做人肉automation啊，所以要求肯定低，很多人都想进亚马逊做云计算，但是95%的人写了主业务代码，而是负责onCall的devops了。


SizeOf说道：

2017年03月03日 17:54 (http://coolshell.cn/articles/17737.html#comment-1913359)
评论不能出现 亚X马X逊关键词？


。说道：

2017年03月03日 16:55 (http://coolshell.cn/articles/17737.html#comment-1913353)
赞！
“我觉得喜欢问责事故责任人的公司最后会变成“大家谁也不做事怕犯错，一旦出问题就开始责备别人”的文化。这样的科技公司最后注定会走下坡路的吧。”
特别认同这点。
而且，文末对国内公司的一般做法，真是说的很到位，我老板也是这样，出了问题从来不从技术上来规避，反而限制大家发布，或者发布流程更繁琐。


YellowTree (http://blog.fungenomics.com)说道：

2017年03月03日 19:29 (http://coolshell.cn/articles/17737.html#comment-1913363)
哈哈


Solomon说道：

2017年03月03日 19:37 (http://coolshell.cn/articles/17737.html#comment-1913364)
只有这样的心态才能接受住考验，也给其他宝贵的经验分享。


michael说道：

2017年03月03日 22:30 (http://coolshell.cn/articles/17737.html#comment-1913367)
耗子哥的酷壳开始接广告了？


陈皓 (http://coolshell.cn)说道：

2017年03月03日 22:55 (http://coolshell.cn/articles/17737.html#comment-1913368)
广告费捐Wikipedia