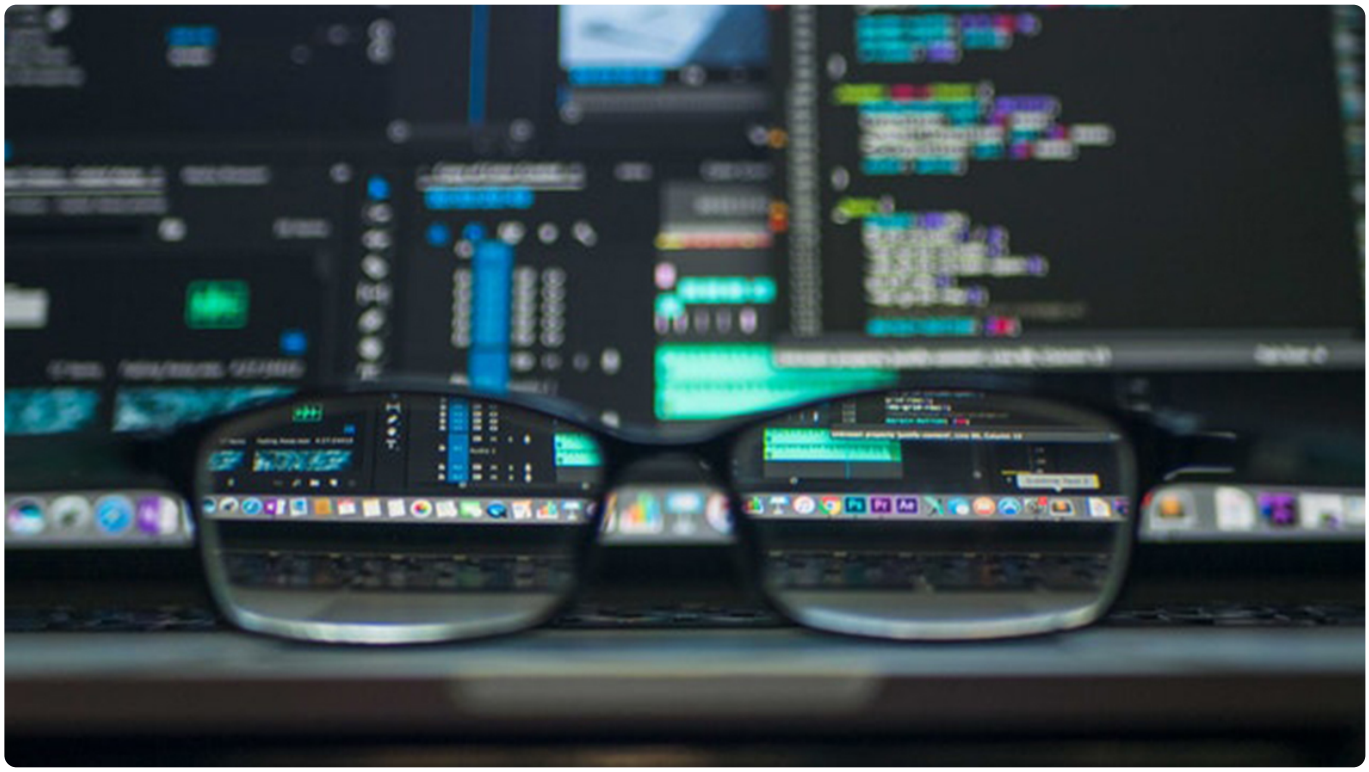
18 | 故障处理最佳实践：故障改进

2017-11-30 陈皓

左耳听风 进入课程 >



讲述：柴巍
时长 08:27 大小 3.88M

在上篇文章中，我跟你分享了在故障发生时，我们该怎样做，以及在故障前该做些什么准备。只要做到我提到的那几点，你基本上就能游刃有余地处理好故障了。然而，在故障排除后，如何做故障复盘及整改优化则更为重要。在这篇文章中，我就跟你聊聊这几个方面的内容。

故障复盘过程

对于故障，复盘是一件非常重要的事情，因为我们的成长基本上就是从故障中总结各种经验教训，从而可以获得最大的提升。在亚马逊和阿里，面对故障的复盘有不一样的流程，虽然在内容上差不多，但细节上有很多不同。

亚马逊内部面对 S1 和 S2 的故障复盘，需要那个团队的经理写一个叫 COE（Correction of Errors）的文档。这个 COE 文档，基本上包括以下几方面的内容。

故障处理的整个过程。就像一个 log 一样，需要详细地记录几点几分干了什么事，把故障从发生到解决的所有细节过程都记录下来。

故障原因分析。需要说明故障的原因和分析报告。

Ask 5 Whys。需要反思并反问至少 5 个为什么，并为这些“为什么”找到答案。

故障后续整改计划。需要针对上述的“Ask 5 Whys”说明后续如何举一反三地从根本上解决所有的问题。

然后，这个文档要提交到管理层，向公司的 VP 级的负责人进行汇报，并由他们来审查。

阿里的故障复盘会会把所有的相关人员都叫到现场进行复盘。我比较喜欢这样的方式，而不是亚马逊的由经理来操作这个事的方式。虽然阿里的故障复盘会会开很长时间，但是把大家叫在一起复盘的确是一个很好的方式。一方面信息是透明的，另一方面，也是对大家的一次教育。

阿里的故障处理内容和亚马逊的很相似，只是没有“Ask 5 Whys”，但是加入了“故障等级”和“故障责任人”。对于比较大的故障，责任人基本上都是由 P9/M4 的人来承担。而且对于引发故障的直接工程师，阿里是会有相关的惩罚机制的，比如，全年无加薪无升职，或者罚款。

老实说，我对惩罚故障责任人的方式非常不认同。

首先，惩罚故障责任人对于解决故障完全没有任何帮助。因为它们之间没有因果关系，既不是充分条件，也不是必要条件，更不是充要条件。这是逻辑上的错误。

其次，做得越多，错得越多。如果不想出错，最好什么也不要做。所以，惩罚故障责任人只会让大家都很保守，也会让大家都学会保守，而且开始推诿，营造一种恐怖的气氛。

说个小插曲。有一次和一个同学一起开发一个系统，我们两个人的代码在同一个代码库中，而且也会运行在同一个进程里。这个系统中有一个线程池模型，我想直接用了。结果因为这个线程池是那个同学写的，他死活不让我用，说是各用各的分开写，以免出了问题后，说不清楚，引起不必要的麻烦。最后，在一个代码库中实现了两个线程池模型，我也是很无语。

另外，亚马逊和阿里的故障整改内容不太一样。亚马逊更多的是通过技术手段来解决问题，几乎没有增加更复杂的流程或是把现有的系统复杂化。

阿里的故障整改中会有一些复杂化问题的整改项，比如，对于误操作的处理方式是，以后线上操作需要由两个人来完成，其中一个人操作，另一个人检查操作过程。或是对于什么样的流程需要有审批环节。再比如：不去把原有的系统改好，而是加入一个新的系统来看（kān，第一声）着原来的那个不好的系统。当然，也有一些整改措施是好的，比如，通过灰度发布系统来减少故障面积。

故障整改方法

就故障整改来说，我比较喜欢亚马逊的那个 Ask 5 Whys 玩法，这个对后面的整改会有非常大的帮助。最近一次，在帮一家公司做一个慢 SQL 的故障复盘时，我一共问了近 9 个为什么。

1. 为什么从故障发生到系统报警花了 27 分钟？为什么只发邮件，没有短信？
2. 为什么花了 15 分钟，开发的同学才知道是慢 SQL 问题？
3. 为什么监控系统没有监测到 Nginx 499 错误，以及 Nginx 的 upstream_response_time 和 request_time？
4. 为什么在一开始按 DDoS 处理？
5. 为什么要重启数据库？
6. 为什么这个故障之前没有发生？因为以前没有上首页，最近上的。
7. 为什么上首页时没有做性能测试？
8. 为什么使用这个高危的 SQL 语句？
9. 上线过程中为什么没有 DBA 评审？

通过这 9 个为什么，我为这家公司整理出来很多不足的地方。提出这些问题的大致逻辑是这样的。

第一，优化故障获知和故障定位的时间。

从故障发生到我们知道的时间是否可以优化得更短？

定位故障的时间是否可以更短？

有哪些地方可以做到自动化？

第二，优化故障的处理方式。

故障处理时的判断和章法是否科学，是否正确？

故障处理时的信息是否全透明？

故障处理时人员是否安排得当？

第三，优化开发过程中的问题。

Code Review 和测试中的问题和优化点。

软件架构和设计是否可以更好？

对于技术欠债或是相关的隐患问题是否被记录下来，是否有风险计划？

第四，优化团队能力。

如何提高团队的技术能力？

如何让团队有严谨的工程意识？

具体采取什么样的整改方案会和这些为什么有很大关系。

总之还是那句话，解决一个故障可以通过技术和管理两方面的方法。如果你喜欢技术，是个技术范，你就更多地用技术手段；如果你喜欢管理，那么你就会使用更多的管理手段。**我是一个技术人员，我更愿意使用技术手段。**

根除问题的本质

最后，对于故障处理，我能感觉得到，**一个技术问题，后面隐藏的是工程能力问题，工程能力问题后面隐藏的是管理问题，管理问题后面隐藏的是一个公司文化的问题，公司文化的问题则隐藏着创始人的问题……**

所以，这里给出三条我工作这 20 年总结出来的原则（Principle），供你参考。

1. **举一反三解决当下的故障。**为自己赢得更多的时间。
2. **简化复杂、不合理的技术架构、流程和组织。**你不可能在一个复杂的环境下根本地解决问题。

3. **全面改善和优化整个系统，包括组织。**解决问题的根本方法是改善和调整整体结构。而只有简单优雅的东西才有被改善和优化的可能。

换句话说，我看到很多问题出了又出，换着花样地出，大多数情况下是因为这个公司的系统架构太过复杂和混乱，以至于你不可能在这样的环境下干干净净地解决所有的问题。

所以，你要先做大扫除，简化掉现有的复杂和混乱。如果你要从根本上改善一个事，那么首先得把它简化了。这就是这么多年来我得到的认知。

但是，很不幸，我们就是生活在这样一个复杂的世界，有太多的人喜欢把简单的问题复杂化。所以，要想做到简化，基本上来说是非常非常难的。（下面这个小视频很有意思，非常形象地说明了，想在一个烂摊子中解决问题，几乎是不可能的事儿。）

0:00 / 0:09

路漫漫其修远兮……

在这篇文章的末尾，我想发个邀请给你。请你来聊聊，在处理好故障之后，你所在的企业会采取什么样的复盘方式。

上一篇　17 | 故障处理最佳实践：应对故障

下一篇　19 | 答疑解惑：我们应该能够识别的表象和本质

精选留言 (27)

写留言

茎待佳阴
2018-01-06
30

我们公司比较奇葩，记得是今年7月的一个晚上，因为那段时间用户量涨得快，所以对服务扩分片，然后，由于需要GO那边的一哥们重启代理，没沟通好，导致，他把代理重启了，我服务还没启动，导致一半的用户无法登陆。CTO当时也在那坐着，起来就把键盘摔了，在那骂半天。之后线上故障了基本也都是这样，只要出问题，就用骂来解决问题，表示问题跟领导没关系。

展开

不会跑
2017-12-05
18

一般会在故障发生时一刀切强调止损，然后故障结束后强调事故报告，接着强调责任“划

分”，最后发现责任人过少或者事故太大，那简单 加上运维团队就好；最后的最后催一催故障报告以及美化故障报告. 对的我是运维

展开

bullboyin...
2017-12-05

13

故障分为自产软件类，第三方软硬件类，操作类，外部原因类共四类。
每起故障都会有技术复盘，由研发总监牵头处理。另外会有月度管理复盘，探讨有哪些管理改进措施。所有改进措施都要创建任务单跟踪，确保必须有个结果，是落实了或者是投入产出比不合适而取消了。
持续优化故障处理流程几年了，故障发生率和平均业务恢复时间都在持续下降中。

展开

z_sz
2018-03-03

12

做得越多，错得越多，隔壁组一个男生就是因为这个考核很差愤而离职了……

helloworld
2017-12-05

5

阿里做基础架构的是不是经常背锅

展开

Geek_fb3db...
2018-11-13

4

为什么网页版本 复制不了，想记录下笔记 都没法复制，这样不好吧。

Geek_fb3db...
2018-11-13

3

问题分析报告 总结原因 然后记录到oa 最后罚款或者绩效 最后该错还是错

亿光年
2018-11-05

2

在前公司也有故障复盘，我也比较喜欢这种模式，每次也都会定位问题，问多个为什么，

但没有亚马逊那么丰富，也会意思性惩罚主要的责任人。当时从领导学到重要一点就是遇到故障立即想办法恢复，而不是去定位问题，定位问题可能需要个很长时间!

展开

xpisme 1

2018-06-26

一：止损 (回滚)
二：事故通报(原因 解决的流程 TODO)
三：case study

展开

剃刀吗啡 1

2018-06-06

我司的处理方式和亚马逊的COE类似，要写这种东西，基本内容也一样。然后严重的故障P1 P2级别的要在公司级别的每个release review大会上复盘。。。另外我司是2B公司，客户很重要，基本上出了大问题都是会给客户造成million级别的损失，所以我司没有惩罚机制，直接fire。。。

展开

冰梨icePea... 1

2017-12-10

阿里内部应该不同bu有不同的处理方式吧，反正支付宝这里比较像你描述的亚马逊的方式，需要回溯过程，分析问题，提出问题以及解决方法，最后action给相关人，在限定时间内给出action 的结果

西北偏北

2019-05-13

复盘整个过程，系统的，全局的去思考问题，并解决，不要赶工被动的，临时的解决问题。

abners

2019-04-02

我们公司会有COE复盘，之前执行的挺好的。会深层次剖析问题根源，并加以解决，到现在我感觉越来越流于形式了，团队拆分，都是回避自己的责任了😔😔

展开

Geek_9ed47...

2019-04-01

耗子哥，亚马逊的工程师是不是更偏向全栈？我了解公司都有故障责任人惩罚的条款。同意您的观点，复盘主要是总结经验教训，避免类似问题再次发生，而惩罚并不能产生这种效果。
我们复盘，也是大家一起讨论，但过程比较简单，没有具体的流程，以后得多向耗子哥学习。

展开

UioSun

2019-03-04

支持“不从物质上惩罚工程师”。

如果觉得无法掌控员工的生产力盈余，可以要求团队写周记甚至日报；如果觉得员工工作不合适，要么谈话，要么开除。惩罚工程师看起来很解气，但对这个人能否反省和进步，意义不大。不再犯错不等于反省，或许就如同文中说的，只是等于“不再触碰”。...

展开

小思绪

2019-02-17

线上出问题之后第一要务是及时恢复线上，但是如何及时找到问题根本原因，不是简单的事，我们就经常在这个上面吃亏。

针对线上问题，会有定期的质量回溯，质量回溯也分几个层次，分别是小组内回溯，系统部门级别回溯，公司级别回溯。

展开

Anker

2019-01-12

复盘过程和AWS类似，不同的是由责任人来写报告

展开

The one ...

2018-12-20

确实，上家公司做的话很多，也出了不少错，但是那些不做事的在一边看笑话，这就有点不爽了

艾尔欧唯伊

2018-09-25

看来我还是太渣，呆过的公司就没有复盘的，解决了就过去了。。要有人愿意口口相传已经万幸

FeiFei Ji...

2018-09-05

在技术债的包袱下，
在混乱的基础架构里，
面对不确定是否可靠的服务，
根本不可能降低故障发生率。