40 | 故障管理：故障应急和故障复盘

2018-03-21 赵成

赵成的运维体系管理课 进入课程 >

讲述：黄洲君
时长 13:09 大小 6.02M

上周我们分享了故障管理中，应该如何对待故障，怎样做好故障定级和定责方面的管理工作。今天我就和你分享当故障真正发生后，我们在故障通报和故障复盘方面的实践经验。

故障应急

当故障真实发生后，带来的影响不仅仅是技术层面的，更多的是业务层面的，比如用户和商家的批量投诉，交易量下跌，广告资损等等。而这些影响又会产生巨大的外部压力，并传递到技术团队，这时如果没有很好的故障应对机制，技术团队就很容易陷入慌乱，不知所措。

我们能否有效应对这种突发且高压的状况，我觉得有两个方面十分关键。

第一方面，业务恢复预案。

这也是我们在故障应急状态下一定要坚守的**第一原则：优先恢复业务，而不是定位问题。**这就需要我们事先有充足的预案准备以及故障模拟演练，这就跟我们前面介绍的各种稳定性保障措施相关，通过稳定性平台的建设，与我们能够预见到的，以及我们经历过的故障场景相结合，当发生故障时能够第一时间执行对应的恢复预案。

同时，预案的执行不能仅仅在故障发生时才执行，而是应该把故障模拟和恢复演练放在平时。我在团队中经常传递的一个理念就是：**凡是没有演练过的预案，都是耍流氓。**也就是如果我们在日常系统稳定的状态下都不敢执行预案，或者执行了没效果，那真到了故障发生后，在更为复杂的状况下，预案 100% 也是不敢做的，因为这种异常状态下，我们还要考虑执行了预案是否会导致次生故障。

关于故障模拟，可以分为不同层面来梳理，比如：

IDC 层面，如电力切换、UPS 切换、核心网络设备切换，单设备故障等，这些故障是可以通过人为破坏进行模拟的，模拟手段相对简单，但是破坏力和影响面会很大，所以做之前一定要准备充分。我们会定期 1~2 个月做一次类似的模拟演练，涉及机房配合的，也会提前跟运营商约定好时间；

系统层面，如 CPU、磁盘 IO、网络 IO、网络时延、丢包等异常场景，这些都有开源或 Linux 系统自带的工具支持，比如 Stress 工具模拟 CPU 升高，dd 模拟磁盘 IO，tc 模拟网络问题；

应用层面，最典型的就是 RT 升高，抛出异常，返回错误码等等，这里还是会用 Spring 的注解功能，在运行时模拟异常状况，然后有针对性地看各种限流降级和开关预案策略能否生效。

关于故障模拟，我再次向你推荐 Netflix 的 Chaos Engineering，介绍得非常全面。

第二方面，有效的组织协调。

故障发生后的排障和恢复操作，往往需要多个技术团队协作完成，这时就需要有一定的应急机制，来确保相关人员能够快速响应和高效协作。同时，因为对业务造成的影响导致业务团队会承受很多外部压力，这时也需要有统一的口径对外反馈，比如大致原因（对外不用详细），影响面以及预估恢复时长等等，从而确保信息的透明，避免各种不着边际的猜测对公司信誉造成的影响。

这时，我们前面介绍到的技术支持这个角色就起到了非常关键的作用。对内，要有效组织技术团队的集中和协作；对外，负责对接业务部门同步信息，同时屏蔽各方对技术团队和故障处理人员的干扰。

出现一个严重故障后，技术支持通常要做如下几个关键事项。

确定故障影响面及等级。故障会通过监控、告警、业务反馈或用户商家投诉几个渠道反馈过来，这时技术支持会根据故障定级标准，快速做出初步判断，确认影响面，以及故障等级。

组织应急小组。对于无法马上恢复或仍需要定位排查的故障，会直接将相关技术团队的主管和骨干开发人员召集到一起，通常是专用的会议室，并确认故障处理主要指挥者，通常是受影响业务的技术负责人，比如商品出现故障就由商品的技术团队主管指挥排障，交易出现故障就由交易技术团队主管指挥，如果是全站性的故障通常会由技术总监直接介入负责指挥。

信息通报。完成上述第一步后，通常会给相关技术和业务团队通报故障初步信息，包括登记、影响面、故障简述以及主要处理团队和责任人。完成第二步，组织起应急小组之后，每隔一定时间，如 15~30 分钟要对进展做一次信息同步。同时，如果等级和故障信息有变，也要同步出来，直至故障排除，业务恢复。为了保证沟通的顺畅，技术支持并不与处理故障的人员直接沟通，而是通过指挥者沟通，这样确保高效沟通，同时也确保处理故障的人员能够相对地专注在故障处理上，而不是响应来自各方的询问，甚至是质问。

所以，整体总结下来，故障应急过程就是：**功夫要下在平时，注意建设各种工具和平台，同时要尽可能地考虑和模拟各种故障场景。**这就像一支军队在平时一定要做各种军事演习一样，然后就是临场发挥。当故障真正出现时，要有完善的应急机制，马上能够有效运转起来，而不是慌乱无措。

故障复盘

上面介绍了故障应急，那接下来我们再看故障管理的下一个阶段，也就是故障发生之后的复盘。

首先，我们一定要先搞清楚复盘的目的。**复盘的目的是为了从故障中学习，找到我们技术和管理上的不足，然后不断改进。**虽然我们不愿意故障发生，但是从故障中学习，反而是提升团队和员工能力的最佳手段，所以我们一定要辩证地看待故障这件事情。

同时，**切忌将复盘过程和目的搞成追究责任或实施惩罚，这对于团队氛围和员工积极性的打击是非常大的。**这一点在前面的内容中已经详细介绍过，这里就不再重复了。

在复盘过程中，技术支持仍然要起到关键作用。

召集复盘会议。会提前将故障信息发给故障处理的参与方，准备复盘过程中需要讨论的问题，视情况决定是否邀请业务方人员参会；

组织会议流程。协调和控制会议中的讨论，也就是俗称的控场；

对故障定级定责。起到类似“法官”的判决作用，根据前面讲到的标准执行；

明确后续改进行动及责任人，录入系统并定期跟踪。

复盘会议中，通常会有哪些关键环节呢？

第一，故障简单回顾。主要针对故障发生时间点，故障影响面，恢复时长，主要处理人或团队做简要说明。

第二，故障处理时间线回顾。技术支持在故障处理过程中会简要记录处理过程，比如每个操作的时间点，责任人，操作结果，甚至是中间的沟通和协作过程，比如几点几分给谁打了电话，多长时间上线的等等，这个过程要求客观真实即可。业务恢复后，会发给处理人进行核对和补充。这个时间线的作用非常关键，它可以相对真实地再现整个故障处理过程。

第三，针对时间线进行讨论。回顾完上述时间线之后，我们会提出过程中存在的疑问，这一点会对主要处理人产生一定的压力，所以一定要保持对事不对人。通常我们会针对处理时长过长、不合理的环节提出质疑，比如为什么告警没有发现问题，而是用户投诉反馈的？为什么从发生故障，到有人上线响应拖了很长时间？为什么对应的场景没有限流、降级和开关等预案？为什么预案执行了没有生效？为什么没有做灰度发布和验证等等？通过这些问题和细节的讨论，我们会找出明显的不足，记录下过程中的改进点。

第四，确定故障根因。通过讨论细节，我们对故障根因进行判断，并再次对故障根因的改进措施进行讨论。在这个环节和上个环节中，通常会有很多讨论甚至是争论，技术支持要发挥的作用就是控制好场面，就事论事，一定不要让讨论失控，演变成相互指责和批斗会，一旦有这种苗头，技术支持一定要及时干预并给出警告。

第五，故障定级定责。根因确定后，结合前面已经确认的故障影响面，就可以对故障定级定责了，这里还要依赖前面我们介绍到的故障标准。不过，定责时，我们会让责任方团队和相关处理人员在场，小范围告知，这样做主要是考虑责任人的个人感受。如果无异议，就形成故障完结报告；如果有异议，则可以向上级主管反馈，直至技术团队负责人（CTO 或技术 VP）为止。

第六，发出故障完结报告。故障完结报告的主要内容包括故障详细信息，如时间点、影响面、时间线、根因、责任团队（这里不暴露责任人）、后续改进措施，以及通过本次故障总结出来的共性问题和建议。这样做的主要目的是保证信息透明，同时引以为戒，期望其它团队也能够查漏补缺，不要犯同样的错误。

定期总结故障案例

除了例行的故障应急和故障复盘，我们还会定期对一个时期内的故障案例进行总结。比如按照一个季度、半年和全年的周期，这样可以更容易地发现一些共性问题，以便于研发团队在稳定性建设方面的规划。

举个例子，2017 年年底，我们整体总结了全年的故障案例，对 P0~P2 严重级别的故障进行分类汇总，就发现全年第三方原因的故障，以及数据类的故障占了很大比例。

我们再往细节分析，发现第三方原因的故障，多数是机房 IDC 的电力、网络切换，单台服务器硬件故障导致的。这些在单次故障复盘时，很容易归因于第三方，但是从全年来看，我们认为根因上，还是我们的系统健壮性不够，在限流降级以及日常的故障模拟演练上，还有很大的提升空间。所以，我们就拉上研发团队的主管和骨干员工，重新看这些故障，重新制定出稳定性提升的改进措施。

同时，在故障定级定责方面，由第三方原因导致的故障，后续不再作为故障根因，而只作为触发因素。所以，在故障复盘时一定要制定出我们自身需要改进的措施。

针对数据类故障，我们总结后发现大多集中在“有状态业务”发布过程中。代码和配置发布可以走发布系统，有完善的流程支持，但数据的变更却更多地依赖人工操作，且流程和周边部件的配合上也不成熟。所以，我们就明确下来，要加大对有状态业务的发布和数据变更工具的支持，将经验固化下来，而不是靠人。

总结

上述这些经验，同时又可以推广到整个研发团队，在不断总结的过程中，整个系统的稳定性不断提升，技术架构也不断完善。

到这里，我们整个故障管理的内容就介绍完了。

总结一下，我们首先要对故障有一个正确和理性的认识，既不能放任不管，也不要谈之色变；同时我们也需要科学的管理方式，跟业务结合，制定出对应的故障等级和定级定责制度。

其次，结合我们前面介绍的稳定性保障体系，在日常要做好各类预案和模拟演练，当故障真实发生时，能够做到冷静处理和高效地组织协调。

最后，在故障复盘中总结出我们的不足，然后不断地改进。

关于故障管理的内容，你还有哪些问题想和我交流，欢迎你留言讨论。

如果今天的内容对你有帮助，也欢迎你分享给身边的朋友，我们下期见！

上一篇　39 | 故障管理：鼓励做事，而不是处罚错误

下一篇　41 | 唇亡齿寒，运维与安全

精选留言 (1)

写留言

孙志宇
2019-04-01

这一讲收获很大呀，重新刷新了我对故障的认知

展开