第 3 章

故障管理

即使再精心设计的系统，在运行过程中，由于一些无法预料的因素，也会遇到各种各样的故障。作为一名合格的系统运维人员，首先要对系统的架构、特征和弱点有所掌握；其次，“工欲善其事，必先利其器”，排查和消除故障，需要先搭建并且掌握先进顺手的工具软件；再者，需要通过一个完整的流程和制度规范对故障进行报告、解决和管理。

本章强调教给读者故障管理的通用方法和思路，使读者对运维工作的故障管理有一定掌握，并在后续的工作过程中起到一定帮助和指导作用。

3.1 集群结构

一个简单的大数据集群体系结构包括以下模块：系统部署和管理，数据存储，资源管理，处理引擎，安全、数据管理，工具库以及访问接口。

集群服务器根据集群中节点所承载的任务性质分为管理节点和工作节点。工作节点一般用于部署各自的存储、容器或计算角色。管理节点一般用于部署各自的组建管理角色。集群功能配置如表 3-1 所示。根据业务类型不同，集群具体配置也有所区别，以实时流处理服务集群为例：Hadoop 实时流处理性能对节点内存和 CPU 有较高要求，基于 Spark Streaming 的流处理消息吞吐量可随着节点数量增加而线性增长。硬件配置如表 3-2 所示。

表 3-1　集群功能配置

模　　块	组　　件	管 理 角 色	工 作 角 色
系统部署	Ambari		
数据存储	HDFS	NameNode	DataNode
		Secondary NameNode	
		JournalNode	
		FailoberController	
	HBase	HBase Master	RegionServer

续表

模　块	组　件	管 理 角 色	工 作 角 色
资源管理	YARN	ResourceManager	NodeManager
		Job HistoryServer	
处理引擎	Spark	History Server	
	Impala	Impala Catalog Server	Impala Daemon
		Impala StateStore	
	Search		Solr Server
安全、数据管理	Sentry	Sentry Server	
工具库	Hive	Hive Metastore	
		Hive Server2	

表 3-2　硬件配置

	管 理 节 点	工 作 节 点
处理器	两路 Intel® 至强处理器，可选用 E5-2650 处理器	两路 Intel® 至强处理器，可选用 E5-2660 处理器
内核数	6 核/CPU（或者可选用 8 核/CPU），主频 2.5 GHz 或以上	6 核/CPU（或者可选用 8 核/CPU），主频 2.0 GHz 或以上
内存	64 GB ECC DDR5	64 GB ECC DDR5
硬盘	两个 2 TB 的 SAS 硬盘（5.5 寸），7200 RPM，RAID1	4～12 个 4 TB 的 SAS 硬盘（5.5 寸），7200 r/min，不使用 RAID
网络	至少两个 1 Gb/s 以太网电口，推荐使用光口提高性能。 使用两个网口链路聚合提供更高带宽	至少两个 1 Gb/s 以太网电口，推荐使用光口提高性能。 使用两个网口链路聚合提供更高带宽
硬件尺寸	1U 或 2U	1U 或 2U
接入交换机	48 口千兆交换机，要求全千兆，可堆叠	
聚合交换机（可选）	4 口 SFP+万兆光纤核心交换机，一般用于 50 节点以上大规模集群	

一个中等规模的集群的节点数一般为 30～200，通常的数据存储可以规划到几百太字节，适用于一个中型企业的数据平台。结构本身也可以通过细分管理节点、主节点、工具节点和工作节点的方式，进一步降低节点复用程度。

3.2　故障报告

3.2.1　故障发现

在运维过程中，发现故障的方式一般分为用户报告、监控告警和人工检查 3 种。随着运维成熟度的逐步提高，用户报告故障的比例会越来越低，呈现反比趋势。主要原因是大部分故障都通过运维自检提前发现提前解决。通过监控系统配置的监控策略叫监控告警，自动根据监控资源发现异常，并通过预先配置的一种或多种告警方式通知管理人员。最后的人工检查是对上述告警的补充，对于监控无法覆盖的指标项，定期人为地进行巡检，能够更全面地评估系统的健康状态。

在故障发现之后，详细、精确记录包括故障起因（如果是用户，要保留用户的联系方式）、故障的现象、故障发生的时间点、故障暂时的影响等。故障描述的详细程度决定了后续故障处理与故障排查的效率，可以帮助管理员快速定位问题原因。一个典型的故障记录单如表 3-3 所示。

表 3-3　故障记录单

分　　类	记　　录
单号	20170511000328
状态	已指派
等待代码	等待管理员接单
记录人员	张三
分析员	李四
报告时间	2017-05-11　11:18:20
客户	王五
客户组织	业务一部
客户电话	×××
客户邮箱	×××
VIP 属性	VIP
故障来源	用户报告
摘要	大数据分析系统×无法登录
详细信息	今天 10:00，李四使用 Chrome 浏览器访问×系统时，在输入用户名和密码之后，页面出现错误信息“服务器内部故障 308，请联系管理员”，截图如附件所示
故障分类	大数据分析系统/×系统/用户登录故障
故障级别	低

3.2.2　影响分析

在运维体系下，一般会划分一、二、三线的人员层级：一线人员指的是直接面向客户处理日常运维问题的前台运维人员；二线人员一般是负责跟进复杂故障问题的专业系统管理员或业务资深运维顾问。三线人员主要是处理深层次故障以及严重问题的研发人员、服务供应商。例如，当架构体系引起组件冲突、软件代码异常性报错等问题时，会由一线、二线逐级上报给三线人员进行排查和后续跟进，在完全修复后逐级向下回溯反馈。

当故障发生之后，一线人员会通过故障记录单记录下故障的详细内容，对故障进行初步归类与判断，划分故障的性质与所属模块的重要性，通过这两个初核信息加上用户故障记录单的反馈数量可以判定故障的影响范围。

判断故障的影响程度对后续处理至关重要，应运用合适的处置手段应对不同层级影响程度的故障。在运维工作中，既不能过度耗费重要资源去处理微小故障问题，也不能按部就班地用常规方式应对可能对系统可用性造成严重打击的致命故障问题。前者可能过度消耗企业的生产资源，且无法让真正重要的事项得到及时支持，后者则会造成核心功能数据的污染甚至造成直接经济损失。对故障的影响程度进行分级，安排合适的资源，

给定合适的预期时间适配同等层级的问题是一般运维工作的重要经验。故障影响分析如表 3-4 所示。

表 3-4 故障影响分析

类 别	识 别 标 准	处 理 方 法
致命	核心系统整体功能或者核心功能失效	立即上报部门或者组织管理层； 协调所有相关资源参与处置
高	核心系统的非核心功能失效； 非核心系统的整体功能失效	协调二线立即参与处置
中	非核心系统的部分功能失效	协调二线参与处置
低	个别用户反馈无法使用； 尚未导致功能受影响的故障	一线参与处置和进一步分析
微小	不对可用性造成影响，暂时不处理也没关系	记录

3.3 故障处理

3.3.1 故障诊断

从故障的发生所属层面来看，可以细化为应用层故障、网络层故障、硬件层故障、系统层故障、客户端故障、机房环境故障等。而如果从故障原因角度出发，则可以参照表 3-5 所示的故障描述。

表 3-5 常见故障

故 障 原 因	描 述
人为操作失误	由于人为操作失误造成的故障，例如误删了系统重要资源
性能容量问题	由于访问量增加、运行时间的累积，JVM HEAP 内存空间、磁盘空间、线程数、网络连接数、打开文件数等超限
软件缺陷	软件在研发过程中遗留的技术债务，临时解决方案，常常在升级变更之后出现问题
硬件故障	服务器因为长时间运行所导致的元部件老化、损坏等故障
兼容性问题	由于应用、服务器、组件、网络等配置参数的冲突，或是组件应用服务与组件本身的软件冲突，在同一个集群环境运行时产生了故障。例如，在应用服务升级过程中发现应用本身依赖的服务 jar 包对高版本上层应用不兼容，从而引起的服务报错

在故障诊断中，有如下几个重要因素。

1．故障的完整描述

如本节前文所述，运维人员对故障的快速定位以及故障范围的准确预估，依赖于故障记录人员准确翔实的故障描述。详尽的故障描述应该尽可能包括下列几个信息：问题的报错码、报错时间段、是否首次发生、可能涉及的业务范围等。通过对上述几个方面的仔细核实，可以避免运维人员把大量的时间成本浪费在资源排查上面。

2．监控信息、dump 文件、日志等现场快照

故障发生时的现场信息是排查故障的关键，把日志、监控信息、dump 文件、网络抓包情况等现场内容汇总获取，可以完成对故障的复现与定位。应用开发时预留的日志输出点显得尤为重要，大多数故障其实都可以通过故障现场的日志数据发现端倪，一些复杂的故障则需要依靠多块日志记录或者监控手段才能定位原因。需要注意的是，这种预留日志的输出需要遵循以下 3 个原则：日志的输出并非越多越好，无用冗余的业务日志甚至可能影响关键信息的获取。日志关键位置输出。合理安排日志输出点的位置，尽力做到以最小输出的代价包含模块的定位。故障现场的保留。可以在异常捕获时多输出一部分故障当场的参数信息，各环节执行结果，等等。遵循上述 3 个原则的日志信息可以极大增加故障解决的进度，减少运维及开发人员的无效排查工作。

3．文档、经验和知识

通过现场快照发现错误的具体信息后，还要根据系统本身的文档、知识库或者管理员的经验更深入地分析。例如，输出日志显示用户授权失败，表明用户的权限信息没有被正常赋权获取。这类常见的问题场景其实可以通过以往的问题检索快速找到解决方案。建立运维体系的知识库和文档资源有助于运维人员迅速提升自身运维经验，运维经验的提升也将极大减少资源诊断排查的时间。当然经验的积累其实并不局限于企业或者公司，互联网开源软件的帮助文档、论坛、搜索引擎检索到的相关问题记录和解决方案，都是故障排查处理的有效手段。

3.3.2　故障排除

故障排除通常有两种做法：变通解决和根本解决。变通解决是当服务故障导致系统不可用时，服务恢复的时效成为第一要素的情况下，通过其他替代方案或是临时方案进行短期内的服务快速恢复。根本解决是指找到并解决引起故障的直接深层原因。例如，我们常见的系统蓝屏，此时通过重启计算机就可以完成蓝屏的变通解决，而根据蓝屏的报错码找到蓝屏的最终原因并予以解决，就是根本解决。

不同种类的故障有不同的排除方法，如表 3-6 所示[1]。

表 3-6　故障排除方法

排 除 方 法	适 应 场 景
重启服务	软件或者硬件产生不明原因的故障时，可通过重启相关模块恢复服务，但要注意的是，复杂系统尤其是分布式系统包含多台服务器、多个应用模块，按照怎样的顺序重启，重启哪些模块也都是可以注意的点
性能调度	当访问量激增时，系统会出现卡顿，一些模块可能会由于资源耗尽而无法再服务，可以通过扩充系统性能来解决。如果系统部署在云上，可以通过云管理平台动态地增加 CPU、内存，甚至整个服务器等来解决性能问题
修补数据	当故障造成数据错误、丢失、重复时，故障的处理就会变得异常烦琐。如果数据特别重要，一定可以修复，则可以安排资源对数据进行逐笔核对，识别错误的地方。这个工作量通常非常大

续表

排 除 方 法	适 应 场 景
升级变更	如果是硬件故障，可以通过升级变更更换硬件；如果是软件问题，可以通过升级变更修复缺陷
隔离、重置等其他应急操作	当系统存在冗余的模块时，为了避免流量仍然导向故障模块，可以彻底手工隔离故障模块；一些系统可能由于自身结构的原因，有一些常发性故障，例如用户登录状态错误，对此可以将重置用户登录状态做成一个功能，以方便在排除故障时使用
自动化	在有了一定故障处理经验和原则之后，对于固定场景的故障，可以考虑开发成自动处理，在捕获到异常之后，由系统管理模块对故障进程自动隔离、自动重启、自动重置、自动扩容等

3.4 故障后期管理

3.4.1 建立和更新知识库

在 3.2.1 节中已经介绍过，在发现故障之后，可以通过单据记录故障的信息，故障的分析和处理过程也可以通过单据记录，保证整个故障处理过程都可以被查阅跟踪。如果是用户反馈的问题，还可以在故障完全解决并验证完成后，由一线运维人员回访用户，完成故障处理的整个业务闭环。一般的机构会遵循 ITIL 的事件和问题流程对故障进行流程化管理。故障处理过程中的单据也应该由运维人员进行收集整理，形成知识库故障处理样例，以供后续处理类似运维问题时借鉴参考。

企业知识库建立的初衷是由于运维工作中积累的大量故障处理经验和知识资源长期以来零散地存储在员工个人的存储介质中，未得到有效整合与共享。这样的情况导致 3 个主要问题：运维日常故障处理的过程完全依赖于特定关键人物的经验积累；运维体系下的全部人员在有限的沟通交流方式下很难做到经验知识的有效分享与积累；已存在的固有经验与处理方案随着环境组件版本的升级无法做到及时更新整理与版本拉平。

针对上述问题，建立知识管理系统，可以实现对大量有价值的案例、规范、手册、经验等知识内容的分类存储和管理，积累知识资产，避免流失；规范知识内容的分类与存储，以此为基础实现后续使用过程中的快捷检索；通过记录并分析故障的处理过程，促进故障处理经验的记录、共享、复用与传承，并与现有管理体系、流程系统进行嵌入，实现整个架构层面的多系统间的知识整合。

3.4.2 故障预防

对于重大故障，找到其根本原因有助于预防和消除同类故障。海恩法则是德国飞机涡轮机的发明者帕布斯·海恩提出的，是一个在航空界关于飞行安全的法则。海恩法则指出：每一起严重事故的背后，必然有 29 次轻微事故和 300 起未遂先兆以及 1000 起事故隐患。法则强调两点：一是事故的发生是量的积累的结果；二是再好的技术，再完美的规章，在实际操作层面，也无法取代人自身的素质和责任心。

海恩法则多被用于企业的生产管理，特别是安全管理中。海恩法则对企业来说是一种警示，它说明任何一起事故都是有原因的，并且是有征兆的；它同时说明安全生产是可以控制的，安全事故是可以避免的；它也给了企业管理者生产安全管理的一种方法，即发现并控制征兆。具体来说，利用海恩法则进行生产的安全管理主要步骤如下。

（1）首先任何生产过程都要进行程序化，这使整个生产过程都可以进行考量，是发现事故征兆的前提。

（2）对每一个程序都要划分相应的责任，可以找到相应的负责人，让他们认识到安全生产的重要性，以及安全事故带来的巨大危害性。

（3）根据生产程序列出每一个程序可能发生的事故，以及发生事故的先兆，培养员工对事故先兆的敏感性。

（4）在每一个程序上都要制定定期的检查制度，以便发现事故的征兆。

（5）在任何程序上一旦发现生产安全事故的隐患，要及时报告、及时排除。

（6）在生产过程中，即使有一些小事故发生，可能是避免不了的或者经常发生的，也应引起足够的重视，要及时排除。当事人即使不能排除，也应该向安全负责人报告，以便找出这些小事故的隐患，及时排除，避免安全事故发生。

许多企业在对安全事故的认识和态度上普遍存在一个“误区”：只重视对事故本身进行总结，甚至会按照总结得出的结论“有针对性”地开展安全大检查，却往往忽视了对事故征兆和事故苗头进行排查；而那些未被发现的征兆与苗头，就成为下一次事故的隐患，长此以往，安全事故的发生就呈现出“连锁反应”。一些企业会发生安全事故，甚至重/特大安全事故接连发生，问题就在于对事故征兆和事故苗头的忽视。

3.5　作业与练习

一、问答题

1．从故障的原因出发，故障可以分为哪些种类？
2．当发生故障时，需要记录哪些相关信息？
3．运维的一线、二线、三线人员的工作职责如何划分？

二、判断题

1．当故障发生时，每次都应该先排查原因，再解决问题。（　　）
2．所有的故障都需要立即协调所有资源进行处理。（　　）
3．重启服务是解决软件故障的唯一办法。（　　）
4．部分影响程度小的故障可以暂缓处理。（　　）

参考文献

[1] 赵川，赵明，路学刚，等．基于大数据的电力运维故障诊断及自动告警系统设计[J]．自动化与仪器仪表，2019（10）：222-226．