如何收集和存储服务器运营的数据

时间：2015-02-09 文章来源：马海祥博客 访问次数：1824 收藏到： 1

腾讯公司从2012年开始，通过对服务器运营流程、工具系统的建设，服务器从一线到三线的运营基本转入线上自动化，在服务器静态配置、动态的运行状态和生命周期各个节点的运营这几个方面，产生了大量的运营数据，这些信息像滚雪球一样，以几何量级快速增长，数据越来越多，该如何着手处理呢?

这就像刚入门的厨子一样，在农贸市场里面对堆积如小山般的食材，无从下手，到2013年，建立网平的大数据平台，把所有的基础架构运营数据统一接入和管理，从此，我们开始了在数据矿山中挖掘金矿的历程，下面马海祥就跟大家来讲讲他们是怎么收集和存储服务器运营数据的。

1、大数据的处理

经过长时间的实践和总结，我们发现服务器运营的大数据有以下四个特点，由浅入深，分别是：

（1）、Volume数据体量巨大，特别是腾讯有海量的服务器，综合起来，数据量可以到PB级别，需要大容量、高性能的存储技术，分析的算法也需要最优化。

（2）、Variety数据类型众多，涉及大量的运行日志、部件状态、生产链运营、环境变量等，经常要抽丝剥茧，才能找到有用的数据。

（3）、Value价值巨大，但并不是每个数据都有价值，需要经过清洗和加工处理后，其产生的效果才能显现，以机房环境温度告警为例，数百万条温度的信息，经过分析对比后，才有可能发现温度异常。

（4）、Velocity数据需要快速处理，特别是告警类的应用，时效性是非常重要的。

2、运营系统架构

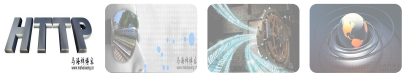
本月热点文章

1 关于大型网站架构的负载均衡技术详解
2 自然语言处理的单词嵌入及表征方法
3 基于高斯模糊原理的模糊图片的研究
4 基于贝叶斯推断应用原理的过滤垃圾...
5 如何收集和存储服务器运营的数据
6 详解内存数据库中的索引技术
7 深入解析互联网协议的原理
8 HTTP、SSL/TLS和HTTPS协议的区...
9 HTTPS建设使用的方案教程解析
10 基于眼球追踪技术对用户调研的探讨...

网站设计 交互设计
网站策划 网页制作
营销策划 营销案例
竞价技巧 数据分析
写作技巧 微信微博
自媒体 新媒体
内容营销 网站运营
O2O模式 App运营
产品运营 网赚教程
创新思维 电子商务
名人访谈 创业故事

热点推荐

HTTP服务的七层架构技术解析及运用

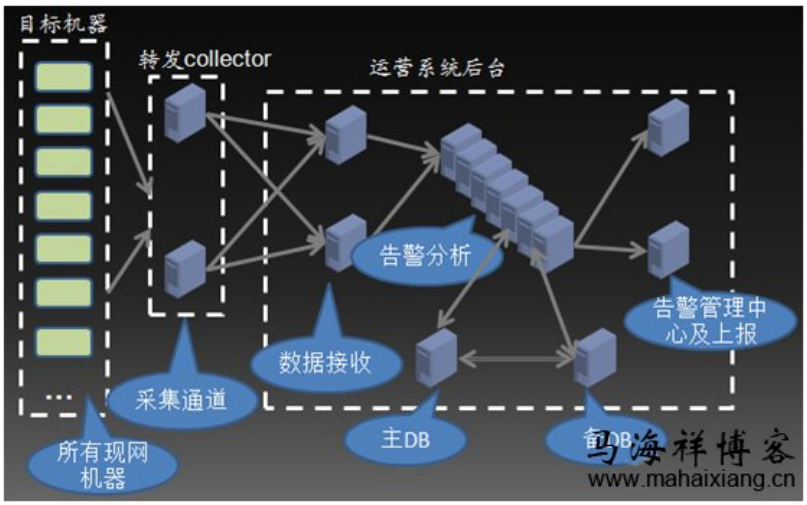


运营思维 更多>>

自媒体运营的规范准则
立即访问

教你写出提高客户转化率的6个文案策略
立即访问

对于海量服务器的管理，我们建立了一套功能强大的运营分析系统，从服务器的带内和带外收集了全面的静态属性和动态运行数据，对服务器的每个关节进行的全方位的数据采集和监控，犹如我们平时体检，把心、肝、脾、肺、肾，甚至每个毛孔，都进行了检查，系统架构如下图所示：



3、存储和分析

数据收集起来后，除了一部分实时的数据存在本地数据库，几乎全部的历史数据都会存储在公司级的数据平台中，这个数据平台提供了丰富的工具系统，功能全面，涵盖了数据存储、分析、实时计算等。

例如，TPG是基于postgreSQL的数据库，用于存放TDW（Tencent distributed Data Warehouse腾讯分布式数据仓库）离线分析后的结果数据，便于系统调用（如服务器利用率分析，故障分析、服务器生命周期等生产数据）；Hbase基于No SQL，万亿级的分布式、有序数据存储，用于存放分析后的结果数据（如温度功耗分析结果数据），整体的架构如下图所示：

4、大数据的四个实践

大数据的规划分析，决策者和开发者首先要从业务驱动的角度，选择数据生产的业务场景，即要预计数据分析得到的结果能带来哪些效益，根据公司服务器运营的特点，我们

在以下四个场景做了大数据的分析和应用，给实际的运营带来的实实在在的好处。

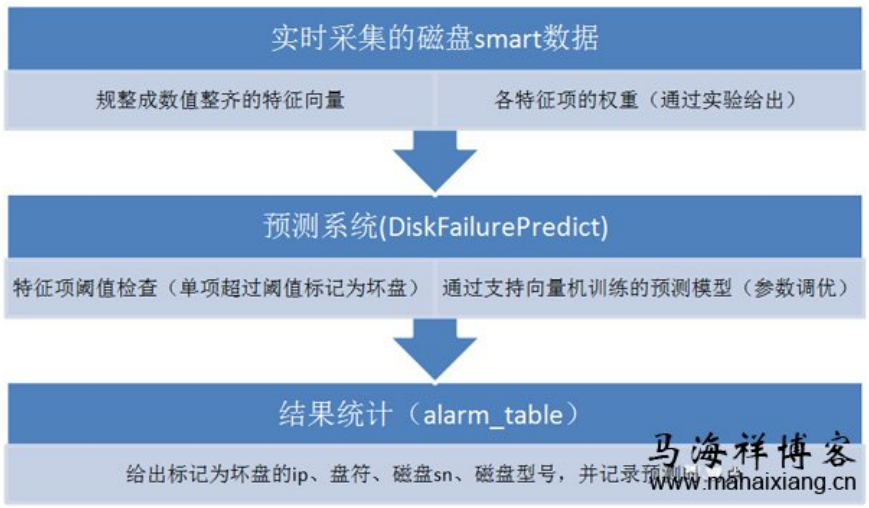
（1）、硬盘故障预测

硬盘是服务器硬件故障率最高的一个部件，如果能提前预测到硬盘故障，对业务体验、完善备件管理都有莫大的收益，这也是基础架构运营在经历自动化、流程化后，需要进一步提升运营效率、降低运营成本的天然要求。

涉及硬盘的运营数据包括业务IO数据、硬盘内部的SMART和硬盘运行的环境变量数据(温度和湿度)。

目前，运营系统对IO数据是每小时采集一次，SMART数据每三小时采集一次，温度和湿度每半小时采集一次，这些数据合计起来每天的记录数上亿条。

硬盘故障预测，适合使用分类算法，我们使用了目前较为流行的SVM分类算法，辅以合适的核函数来加快学习计算的效率。



经过了一年多时间的实践，走了不少弯路，也碰到了很多坑，在硬盘故障标准确定、业务IO分类定义等方面吃了不少的亏，我们在基于SMART数据做的故障预测，达到了令人满意的效果，在实际运营环境中验证的结果如下：准确率precision达到98%，预测时间leadtime的整体偏差不超过2天。

需要重点指出的是，我们做的预测结果，除了training阶段用历史数据外，验证的过程是用现网的实时数据来进行的，就是说，经过SVM算法得到的预测模型后，我们是用最新采集的实时数据输入到模型中，得到的ok和fail两种预测结果，在3天、7天、14天后再对预测的结果进行验证。

这个比传统的预测方式（训练和验证都是使用历史数据），对现网应用的价值大大提高了，目前在现网环境中，主要的落地场景包括：

①、预测出来的结果，经过运营流程，对BG业务提前发出预警，以提高业务运维效率。

②、根据预测出来的大规模硬盘故障，对备件进行有效管理。

（2）、服务器利用率分析

一般来说，大网络公司的业务类型和机型都相当多，机器分配给业务后，使用的情况如何？我们需要跟踪服务器的利用率情况，下图是某业务某机型磁盘IO的利用率统计分析图：

分析过程如下：存储类机型，看到一段时间统计出来的IO的利用率并不高，并且是写少读多的应用，是否可以考虑使用IOPS相对不高的廉价硬盘？还是业务的架构存在优化的空间?

服务器利用率分析给运营带来的好处在于：

①、结合业务模型，发现业务应用服务器的短板，在发现并修复系统架构缺陷的同时，提高整体利用率。

②、对机型选型的优化，例如对于磁盘容量使用率不高的机型，在后续的机型定制中减少硬盘的数量。

（3）、故障率分析

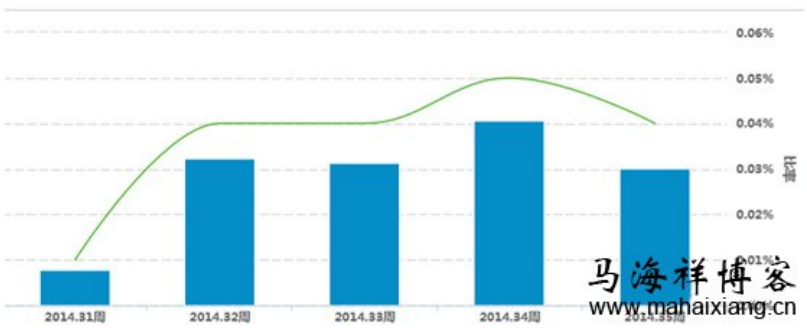
服务器故障分析对服务器的各个部件的故障率都做了分析和监控，包括：

①、生成月度故障率报表。

②、故障率异常的实时监控和自动告警。

③、分析外部条件与故障率的关系。

④、与OS的软件告警信息联动起来，及时发现服务器的亚健康状态。



上图是某服务器硬件最近几周的故障率统计信息，按部件给出各个机型的故障率情况，及时发现批次性故障并给出告警。

（4）、环境监控

2013年8月，华东地区遭遇罕见的高温天气，很多机房空调制冷扛不住了，频繁发生服务器高温重启的事件，如果能把机房环境温度有效的监控起来，我们就能在发现异常时发出高温告警，提前采取措施，对服务器入风口温度进行采集和监控是一个较为有效的方案。

本月热点文章

1 关于大型网站架构的负载均衡技术详解
2 自然语言处理的单词嵌入及表征方法
3 基于高斯模糊原理的模糊图片的研究
4 基于贝叶斯推断应用原理的过滤垃圾...
5 如何收集和存储服务器运营的数据
6 详解内存数据库中的索引技术
7 深入解析互联网协议的原理
8 HTTP、SSL/TLS和HTTPS协议的区...
9 HTTPS建设使用的方案教程解析
10 基于眼球追踪技术对用户调研的探讨...


在做SEO优化的过程中，网页代码中的Meta标签可以.....


网站页面标题的SEO优化及布局要点
对于一个刚入行的站长或SEO来说，首先要搞明白.....

标签云

入风口温度(℃)				功率(W)			CPU1(℃)
当前	昨高	偏移	昨低	当前	昨高	昨低	
37	20	85.00%	19	-	-	-	62
37	20	85.00%	19	22	336	288	64
37	20	85.00%	19	16	256	240	71
37	20	85.00%	20	21	328	288	52
35	19	84.21%	19	23	360	328	54
35	19	84.21%	18	20	336	288	57

上图显示服务器入风口温度变化的异常情况，经过数据的规整和误差修正，产生了高温告警，通过自动化流程，及时知会到机房现场负责人。

5、数据质量的把控

数据的质量和字段规范性对后面分析效果的影响很大，但业务开发所设计的数据不是为了运营分析而服务的，很多情况下都是为了功能开发而存在，如果可以在系统构建初期进行介入，其实可用避免很多清洗工作，数据可直接投入分析使用。

这里开发人员和数据分析的人员存在一个gap，如果对数据在系统设计中遇上各种约束的话，开发人员会觉得很痛苦，开发效率非常低；而数据分析人员却觉得如果数据能做到工具级定制，就是连数据的表字段的名称、注释、连内部关系，都是由系统统一生成，这样采集完美的。

后来，我们内部经过一段时间的讨论和磨合，形成的共识，我们做的是运营系统，归根到底是为运营服务的，而数据分析是运营的一个重要功能，所以没有办法，这个问题还是需要开发阶段来解决，开发人员只能克服了。

6、精细化的传感器

对于服务器上传感器的设计，互联网企业有特殊的需求，对上游硬件厂商的依赖是比较高的，腾讯有大量的服务器运营数据，非常希望可以跟业界一起在数据、资源、算法等各个维度可以共享，寻求更多提高运营效率的途径。

这里的传感器也可以从广义上来展开，除了服务器物理上的sensor越来越多，在服务器各个运营环节都可以在流程中加入各种采集代码，把服务器部署、搬迁、退役等每个细小的步骤都如实的记录下来，运营系统的不断优化将使“传感器”体积微型化，它将出现在生产的每一个角落，为运营决策提供更科学的数据支撑。

7、不要被数据误导

人们很容易被大数据忽悠，在很多场合我们都谈了大数据强大的功能和美好的未来，认为可以解决许多社会问题，甚至预测未来。

但在马海祥看来，无论大数据如何神奇，若试图用大数据引领未来只会误入歧途，因为大数据背后本就存在着“先天不足”：从本质上看，大数据最大的缺陷就在于试图以确定去“颠覆”混沌与不确定性。

之前我们做硬盘故障预测，直观的认为硬盘的读写压力对硬盘老化和故障是有直接关系的，但经过分析，发现业务使用硬盘的随机性太大了，硬盘响应IO的模式也很多变，对于业务的IO读写比例、块大小等，有太多的不确定性，就是前面说的混沌，导致前面基于IO做的预测结果非常糟糕。

其实这里要说的就是，目前这个阶段，依靠大数据来指导服务器运营，不靠谱，服务器运营智能化远远没有达到，这里还是要靠运营和开发人员的思维和头脑，把自动化运营

本月热点文章
1 关于大型网站架构的负载均衡技术详解
2 自然语言处理的单词嵌入及表征方法
3 基于高斯模糊原理的模糊图片的研究
4 基于贝叶斯推断应用原理的过滤垃圾...
5 如何收集和存储服务器运营的数据
6 详解内存数据库中的索引技术
7 深入解析互联网协议的原理
8 HTTP、SSL/TLS和HTTPS协议的区...
9 HTTPS建设使用的方案教程解析
10 基于眼球追踪技术对用户调研的探讨...

先做好。

马海祥博客点评：

随着数据的逐步完善和开放，互联网和企业都将建立起完善的大数据服务基础架构及商业化模式，从数据的存储、挖掘、管理、计算等方面提供一站式服务，将各行各业的数据孤岛打通互联。

而且数据应用的生态系统也将变得非常成熟，甚至出现用户与数据服务商之间的算法提供商，他们有专业领域内的精英人才，通过数据挖掘的方式，寻找事物间的联系，用户只需将其原始数据导入，提供商很快的就能在线的将分析结果返回，如水和电一样，即开即用。

打赏

相关标签搜索： 数据收集 数据存储 服务器运营

上一篇：详解内存数据库中的索引技术
下一篇：HTTP与HTTPS的区别

相关文章推荐：

1 HTTP与HTTPS的区别
2 深入解析互联网协议的原理
3 基于眼球追踪技术对用户调研的探讨研究
4 自然语言处理的单词嵌入及表征方法
5 HTTPS建设使用的方案教程解析
6 如何开启苹果系统的两步验证机制，避免
7 今日头条的个性化推荐算法
8 关于大型网站架构的负载均衡技术详解
9 HTTP、SSL/TLS和HTTPS协议的区别与联系
10 HTTP服务的七层架构技术解析及运用

您可能还会对以下这些文章感兴趣！

详解大型网站系统的特点和架构演化发展历程

大型网站的挑战主要来自庞大的用户，高并发的访问和海量数据，任何简单的业务一旦需要处理数以P计的数据和面对数以亿计的用户，问题就会变得棘手，大型网站架构主要就是解决这类问题。大型网站不是从无到有一步就搭建好一个大型网站，而是能够伴随小型网站业务的渐进发……【查看全文】

阅读：853 关键词： 大型网站 网站架构 网站系统 日期：2017-03-02

计算机的开机启动原理

计算机从打开电源到开始操作，整个启动可以说是一个非常复杂的过程。总体来说，计算机的整个启动过程分成四个阶段：第一阶段：BIOS；第二阶段：主引导记录；第三阶段：硬盘启动；第四阶段：操作系统；直至执行/bin/login程序，跳出登录界面，等待用户输入用户名和密码。……【查看全文】

阅读：3039 关键词： 计算机 计算机启动 计算机原理 开机启动原理 日期：2014-01-16

HTTP服务的七层架构技术解析及运用

本月热点文章

1 关于大型网站架构的负载均衡技术详解
2 自然语言处理的单词嵌入及表征方法
3 基于高斯模糊原理的模糊图片的研究
4 基于贝叶斯推断应用原理的过滤垃圾...
5 如何收集和存储服务器运营的数据
6 详解内存数据库中的索引技术
7 深入解析互联网协议的原理
8 HTTP、SSL/TLS和HTTPS协议的区...
9 HTTPS建设使用的方案教程解析
10 基于眼球追踪技术对用户调研的探讨...

一般来说，计算机领域的体系结构普遍采用了分层的方式，从最底层的硬件往高层依次有：操作系统->驱动程序->运行库->系统程序->应用程序等等。从网络分层模型OSI来讲，由上至下为：应用层->表示层->会话层->传输层->网络层->数据链路层->物理层。当然实际应用的TCP/IP协……【查看全文】

阅读：4386 关键词： 七层架构解析 七层架构运用 七层架构技术 http服务 日期：2014-09-

如何开启苹果系统的两步验证机制，避免iCloud帐号遭到攻击

首先，你需要登录至苹果的网页版Apple ID管理系统，你需要点击“管理你的Apple ID”，随后输入帐号密码信息。在登录之后，你需要从左侧导航栏中选择“密码和安全”选项，在这里，你将需要验证安全问题，随后下拉至“两步验证”区域，点击蓝色的“开始”链接并阅读其中的……【查看全文】

阅读：1407 关键词： 苹果系统 验证机制 icloud攻击 icloud帐号 icloud 日期：2014-09-

关于大型网站架构的负载均衡技术详解

负载均衡是将负载（工作任务，访问请求）进行平衡、分摊到多个操作单元（服务器，组件）上进行执行，是解决高性能，单点故障（高可用），扩展性（水平伸缩）的终极解决方案。面对大量用户访问、高并发请求，海量数据，可以使用高性能的服务器、大型数据库，存储设备，高性能W……【查看全文】

阅读：809 关键词： 大型网站 网站架构 负载均衡 日期：2016-08-05

今日头条的个性化推荐算法

互联网给用户带来了大量的信息，满足了用户在信息时代对信息的需求，但也使得用户在面对大量信息时无法从中获得对自己真正有用的那部分信息，对信息的使用效率反而降低了，而通常解决这个问题最常规的办法是推荐系统。推荐系统能有效帮助用户快速发现感兴趣和高质量的信……【查看全文】

阅读：12908 关键词： 今日头条 日期：2016-01-20

基于贝叶斯推断应用原理的过滤垃圾邮件研究

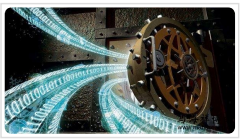
随着电子邮件的应用与普及，垃圾邮件的泛滥也越来越多地受到人们的关注。而目前正确识别垃圾邮件的技术难度非常大。传统的垃圾邮件过滤方法，主要有关键词法和校验码法等。前者的过滤依据是特定的词语；后者则是计算邮件文本的校验码，再与已知的垃圾邮件进行对比。它们……【查看全文】

阅读：855 关键词： 贝叶斯推断 贝叶斯应用 贝叶斯原理 过滤垃圾邮件 垃圾邮件 日期：

HTTPS建设使用的方案教程解析

百度已对部分地区开放HTTPS加密搜索服务，随后，百度实行全站化HTTPS安全加密服务，百度HTTPS安全加密已覆盖主流浏览器，旨在用户打造了一个更隐私化的互联网空间、加速了国内互联网的HTTPS化。同时也希望更多网站加入到HTTPS的队伍中来，为网络安……【查看全文】

阅读：42 关键词： seo https 日期：2018-02-01



详解内存数据库中的索引技术

传统的数据库管理系统把所有数据都放在磁盘上进行管理，所以称作磁盘数据库（DRDB:Disk-Resident Database），磁盘数据库需要频繁地访问磁盘来进行数据的操作，磁盘的读写速度远远小于CPU处理数据的速度，所以磁盘数据库的瓶颈出现在磁盘读写上，基于此，内存数据库的概……【查看全文】

阅读：3257 关键词： 内存数据库 索引技术 数据库 日期：2015-01-09

HTTP、SSL/TLS和HTTPS协议的区别与联系

本月热点文章

HTTPS是为了安全性而设置的，要验证很多的信息，相对应http请求的速度肯定有点慢，如果使用HTTPS的话很麻烦的，无意给服务器和客户端增加了很大的压力，所以平时最好不要使用HTTPS，如果牵扯到个人隐私或者是其他的什么重要信息就一定要这么做了，很多的时候你感觉有点问题，……【查看全文】

阅读：14035 关键词： http ssl https https协议 日期：2016-05-13

↓ 点击查看更多 ↓

本月热点文章

1. 关于大型网站架构的负载均衡技术详解
2. 自然语言处理的单词嵌入及表征方法
3. 基于高斯模糊原理的模糊图片的研究
4. 基于贝叶斯推断应用原理的过滤垃圾...
5. 如何收集和存储服务器运营的数据
6. 详解内存数据库中的索引技术
7. 深入解析互联网协议的原理
8. HTTP、SSL/TLS和HTTPS协议的区...
9. HTTPS建设使用的方案教程解析
10. 基于眼球追踪技术对用户调研的探讨...