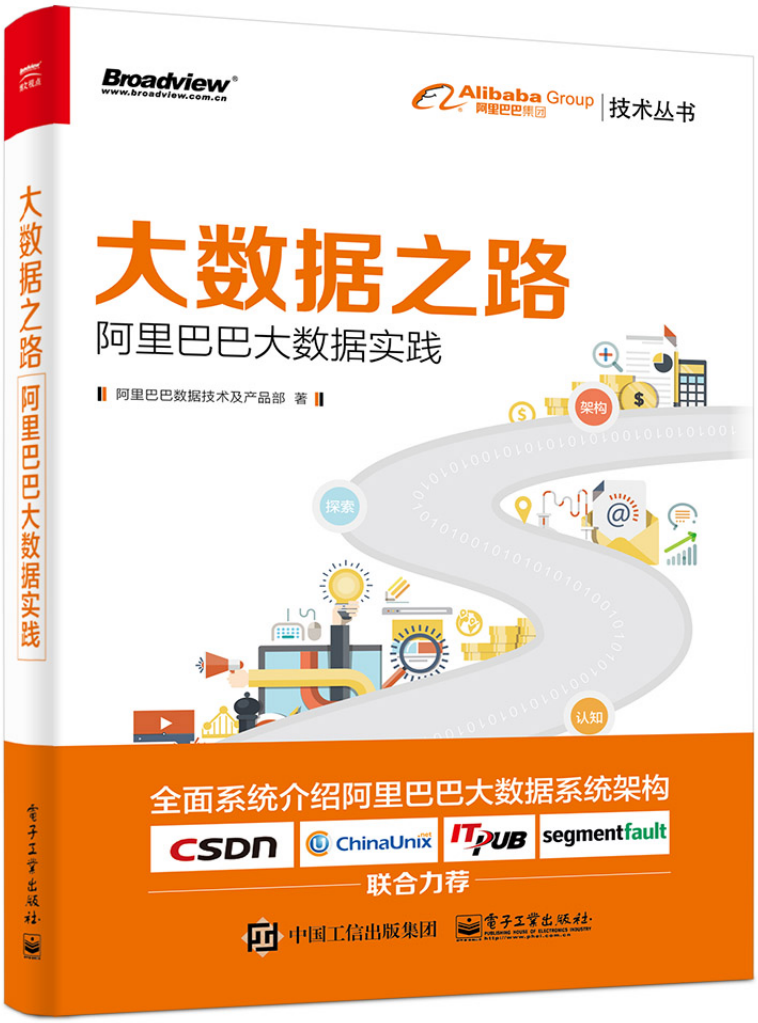
日志采集的挑战，留言免费获取《大数据之路：阿里巴巴大数据实践》

本文节选自《大数据之路：阿里巴巴大数据实践》，关注 iteblog_hadoop
公众号并在这篇文章里面
文末评论区留言（认真写评论，增加上榜的机会）。留言点赞数排名前5名的粉丝，各免费赠送一本《大数据之路：阿里巴巴大数据实践》，活动截止至08月11日18:00。

这篇文章评论区留言才有资格参加
送书活动：https://mp.weixin.qq.com/s/BR7M8Rty3dN1YBrXtpjmgQ

对于目前的互联网行业而言，互联网日志早已跨越初级的饥饿阶段（大型互联网企业的日均日志收集量均以亿为单位计量），反而面临海量日志的淹没风险。各类采集方案提供者所面临的主要挑战已不是日志采集技术本身，而是如何实现日志数据的结构化和规范化组织，实现更为高效的下游统计计算，提供符合业务特性的数据展现，以及为算法提供更便捷、灵活的支持等方面。



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

这里介绍两个最典型的场景和阿里巴巴所采用的解决方案。

日志分流与定制处理

大型互联网网站的日志类型和日志规模都呈现出高速增长的态势，而且往往会出现短时间的流量热点爆发。这一特点，使得在日志服务器端采用集中统一的解析处理方案变得不可能，其要求在日志解析和处理过程中必须考虑业务分流（相互之间不应存在明显的影响，爆发热点不应干扰定常业务日志的处理）、日志优先级控制，以及根据业务特点实现定制处理。例如，对于电商网站而言，数据分析人员对位于点击流前端的促销页面和位于后端的商品页面的关注点是不一样的，而这两类页面的流量又往往同等重要且庞大，如果采用统一的解析处理方案，则往往需要在资源浪费（尽可能多地进行预处理）和需求覆盖不全（仅对最重要的内容进行预处理）两个选择之间进行取舍。这种取舍的结果一般不是最优的。

考虑到阿里日志体量的规模和复杂度，分治策略从一开始便是阿里互联网日志采集体系的基本原

则。这里以PV日志采集领域一个最浅显的例子来说明，与业界通用的第三方日志采集方案的日志请求路径几乎归一不同，阿里PV日志的请求位置（URL）是随着页面所在业务类型的不同而变化的。通过尽可能靠前地部置路由差异，就可以尽可能早地进行分流，降低日志处理过程中的分支判断消耗，并作为后续的计算资源调配的前提，提高资源利用效率。与业界方案的普遍情况相比，阿里的客户端日志采集代码的一个突出特点是实现了非常高的更新频次（业界大多以季度乃至年为单位更新代码，阿里则是以周/月为单位），并实现了更新的配置化。我们不仅考虑诸如日志分流处理之类的日志服务器端分布计算方案，而且将分类任务前置到客户端（从某种程度上讲，这才是真正的"分布式"！）以实现整个系统的效能最大化。最后可以在计算后端几乎无感知的情况下，承载更大的业务量并保证处理质量和效率。

采集与计算一体化设计

以PV日志为例，页面PV日志采集之后一个基础性操作是日志的归类与汇总。在早期的互联网日志分析实践中，是以URL路径，继而以URL（正则）规则集为依托来进行日志分类的。在网站规模较小时，这一策略还可以基本顺利地运转下去，但随着网站的大型化和开发人员的增加，URL规则集的维护和使用成本会很快增长到不现实的程度，同时失控的大规模正则适配甚至会将日志计算硬件集群彻底榨干。

这一状况要求日志采集方案必须将采集与计算作为一个系统来考量，进行一体化设计。阿里日志采集针对这一问题给出的答案是两套日志规范和与之对应的元数据中心。其中，对应于PV日志的解决方案是目前用户可直观感知的SPM规范（例如，在页面的URL内可以看见spm参数）和SPM元数据中心。通过SPM的注册和简单部署（仅需要在页面文件内声明一个或多个标签），用户即可将任意的页面流量进行聚类，不需要进行任何多余的配置就可以在相应的内部数据产品内查询聚合统计得到的流量、转化漏斗、引导交易等数据，以及页面各元素点击数据的可视化视图。对应于自定义日志的解决方案则是黄金令箭（Goldlog）/APP端的点击或其他日志规范及其配置中心。通过注册一个与所在页面完全独立的令箭实体/控件实体，用户可以一键获得对应的埋点代码，并自动获得实时统计数据和与之对应的可视化视图。通过简单的扩展配置，用户还可以自动获得自定义统计维度下的分量数据。

在当前的互联网环境下，互联网日志的规模化采集方案必须具备一个与终端设备的技术特点无关，具有高度扩展弹性和适应性，同时深入契合应用需求的业务逻辑模型，并基于此制定对应的采集规范交由产品开发人员执行。若非如此，则不足以保障采集—解析—处理—应用整个流程的通畅。目前阿里已成功实现规范制定—元数据注册—日志采集—自动化计算—可视化展现全流程的贯通。通过一体化设计，用户甚至可以在不理解规范的前提下，通过操作向导式界面，实现日志采集规范的自动落地和统计应用。日志本身不是日志采集的目的，服务于基于日志的后续应用，才是日志采集正确的着眼点。

活动规则

【1】关注 iteblog_hadoop 公众号，并在评论区留言获点赞数最高前5名将赠送；《大数据之路：阿里巴巴大数据实践》1本，共送出5本；
【2】活动时间：即日起至08月11日18:00点；
【3】活动结束后，收到中奖通知的用户请在公众号回复：微信号 + 姓名 + 地址+ 电话 + 邮编；
【4】本活动解释权归Hadoop技术博文所有。