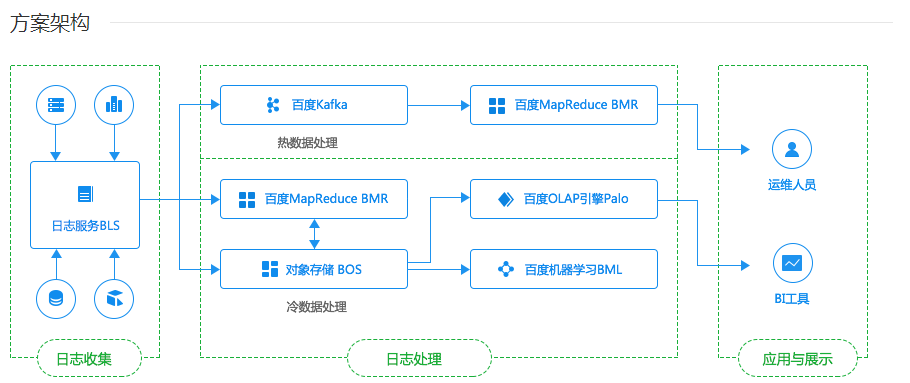
# 日志分析大数据项目

在大数据时代，企业的生产、销售、经营、管理逐渐由技术驱动转型为数据驱动，在这一过程中数据的重要性不言而喻。借助于云计算、大数据等技术的发展，我们通过收集、处理、分析、展示数据，重新认识了周围的世界，重新定义了规则与价值，数据驱动的企业能够更全面地认知现状、更精准地把握趋势。而在企业日常运行积累的海量数据中，日志是其中最易获得、覆盖面最广同时也是最有价值的数据之一。

日志指用户、服务器、网络设备、操作系统、数据库、应用软件等产生的各类数据，它的核心概念是时序递增的事件序列。基于时间与事件这两个基本要素，我们可以在日志中追溯过去的记录、记录现在的状态、探寻未来的趋势，这就是日志所蕴含的价值。根据日志的不同来源与类型，典型的日志应用场景可以划分为以下三类：  
· 运维日志：记录设备、系统的运行情况，用于监控服务器状态、定位排查故障、分析性能瓶颈等；  
· 用户日志：记录业务信息，例如页面的PV、UV、停留时间等，用于了解业务运营状况、分析用户行为特征等；  
· 安全日志：记录服务器、防火墙等设备的日志，供安全审计系统使用，用于监控系统安全、进行可信取证等。

传统日志分析工具往往局限于单一的应用场景，功能较为简单，难以满足运维、产品、运营、管理者等不同角色的不同需求。且随着业务量的增长，日志文件在容量、类型、产生速度等方面都成倍地增长，对于系统在处理速度、并发量、分析维度等方面的要求也越来越高，传统的日志分析工具越来越难以适应实际应用的需求。

随着大数据技术的日渐发展成熟，新的日志分析工具应运而生。将大数据技术与云计算技术相结合，托管式日志分析服务综合了大数据的大体量、多样性、时效性等特性与云计算强可靠性、高可用性、即取即用与经济性等优势，克服了传统工具在功能与性能上的限制，解决了自行搭建系统门槛高且运维困难的问题，是一种经济高效的解决方案。



基于Lambda架构同时实现热数据处理与冷数据处理，包括日志收集、处理与分析、应用与展示三个模块。

在日志收集模块，日志收集服务BLS从日志源（如云服务器BCC）收集日志。BLS是托管式日志收集服务，用户只需配置源地址、目的地址、收集规则等简单信息即可实现日志的高可靠、高可用收集。

收集到的日志即可接入日志处理模块。一方面对于热数据处理场景，可以将日志接入消息服务Kafka作为消息队列，投递到百度MapReduce BMR提供的流式处理服务如Spark Streaming中进行实时计算。另一方面对于冷数据处理场景，可以将日志先写入对象存储BOS进行存储，或者直接写入HBase集群，之后接入Hive/Spark SQL集群进行分析处理。百度BMR是全托管的Hadoop/Spark集群，可以按需部署并弹性扩展集群，用户只需专注于大数据处理、分析、报告，由拥有多年大规模分布式计算技术积累的百度运维团队全权负责集群运维。经过BMR处理好的数据可以写入数据仓库，如Palo。同时也可直接由BOS中的数据结合百度机器学习BML进行用户行为预测等分析操作。

在应用与展示模块，热数据经过处理可以提供警报给运维人员；冷数据可通过BI工具连接Palo进行展示。

## 方案优势

托管服务

### 托管服务

提供托管式服务模式，用户无需关注技术细节即可搭建完整的服务系统，专业的运维团队提供全程的运维服务，使用户能够专注于业务本身。

灵活开放

### 灵活开放

可灵活搭配大数据分析服务，根据需要搭建系统；且对各主流架构及应用开放兼容，扩展性强、接入成本低。

经济高效

### 经济高效

提供按需部署的使用方式，按需计费并可弹性扩展，在高效完成任务的同时可有效较低人力、资源成本。

相比自行搭建开源产品实现日志分析，百度云提供了一站式服务平台，无需学习门槛或学习成本，也不用担心软件升级或系统扩容时的运维成本，让您专注于分析业务以获取洞察力，在大数据时代引领业务升级。

百度日志服务（Baidu Log Service）是百度云提供的日志收集与投递服务，依托云计算与大数据的技术优势，您只需简单地部署配置，面对海量日志亦可手到擒来，低成本、高效率地实现日志的采集、聚合与传输，轻松应对设备运维管理、商业趋势洞察、安全监控审计等业务场景。



百度MapReduce（BMR）是全托管的Hadoop/Spark集群，您可以按需部署并弹性扩展集群，只需专注于大数据处理、分析、报告，拥有多年大规模分布式计算技术积累的百度运维团队全权负责集群运维。





百度消息服务是分布式、高可扩展、高通量的托管消息队列服务，支持以下业务场景：

* 从网站、设备或应用程序采集海量的用户浏览、点击、搜索等数据以便实时分析。
* 汇总分布式应用的遥感数据方便系统运维。
* 对接Spark Streaming等服务以进行实时流数据分析。

百度消息服务是基于Apache Kafka的托管服务。Kafka是一个分布式、多分区、多副本的消息服务。通过消息队列，生产者和消费者异步交互，而不需要彼此等待。相对于传统的消息服务，Kafka有以下特点：

* 主题可以通过分区（Partition）来实现水平扩展。
* 分区分布在多个节点上以达到高数据可用性。
* 通过消费者组（Consumer Group）来支持单个消费者以队列或者Pub/Sub形式的消息消费，或者多个消费者集群顺序消费消息。

百度消息服务封装了Kafka集群细节，并以托管服务形式提供。您可以直接使用百度消息服务来集成大规模分布式应用，而无需考虑集群运维，仅按照使用量付费。

日志分析业务需求

|  |  |
| --- | --- |
| **Reports** | **Description** |
| Hosts Report | The hosts which have accessed the MS IIS Web server. The report contains the following details about the hosts: IP address of the client, Hits, Page views, Bytes sent by the host and Events of the host |
| Users Report | Users who have accessed the MS IIS Web server. The report contains the following details about the users: User name, Hits, Page views, Bytes sent by the user and Events of the user |
| File Type Report | The different file types accessed through the MS IIS Web server and the bandwidth consumed. The report contains the following details about the type of files: File type accessed, Hits, Percentage, Bytes sent by the file type and Events of the file type |
| Page URLs Report | The URLs accessed through the MS IIS Web server and the bandwidth consumed. The report contains the following details about the URLs accessed: URI Stem, Hits, Page views, Bytes sent by the URL access and Events for the URL accessed |
| Browser Usage Report | The list of browsers used to access the MS IIS Web server. The report contains the following details about the browser usage: Browser, Hits, Percentage and Events for the browser |
| OS Usage Report | The list of operating systems used to access the the MS IIS Web server. The report contains the following details about the OS usage: OS, Hits, Percentage and Events for the OS |
| HTTP Error Status Code Report | The list of error status codes occurred during the access of the MS IIS Web server. The report contains the following details about the HTTP error status codes: HTTP Status error codes and Events for the status code |
| Malicious URL Report | The list of malicious URLs accessed through the MS IIS Web server. The report contains the following details about the malicious URL access: Client IP address, User name and Events for the malicious URL access |
| Cross Site Scripting Attempts | The list of cross site scripting (XSS) attacks attempted through the MS IIS Web server. The report contains the following details about the XSS attacks: Client IP address, User name and Events for the XSS attempts |
| SQL Injection Attempts | The list of SQL injection attacks attempted through the MS IIS Web server. The report contains the following details about the SQL injection attacks: Client IP address, User name and Events for the SQL injection attempts |

* IIS网络服务器报表
  + 用户
  + 方法
  + 顶页
  + 热门查询
  + 图片
  + 文件类型
  + 访问者
  + 浏览器
  + 操作系统的排行
  + 错误的排行
  + 窗口错误
  + 主机的排行
  + IIS服务器趋势
  + IIS服务器概览
* IIS网络服务器错误报表
  + HTTP状态成功
  + 失败的用户认证
  + HTTP错误请求
  + HTTP付费要求
  + 站点访问被拒绝
  + 更改密码
  + HTTP请求URI太大
  + HTTP请求实体太大
  + 失败的HTTP预期
  + HTTP不支持的媒体类型
  + HTTP锁定错误
  + HTTP无效网关
  + IP地址被拒绝
  + 读禁止访问
  + 写禁止访问
  + 服务不可用
  + 网关超时
  + UNC授权失败
  + 拒绝Global.asa直接请求
  + IO操作中止
  + Web服务器重启
  + Web服务器繁忙
  + 信息报表
  + 成功的报表
  + 重定向报表
  + 客户端错误报表
  + 服务器错误报表
* IIS网络服务器攻击报表
  + SQL攻击报表
  + 跨站脚本报表
  + 恶意URL请求
  + 恶意文件执行
  + cmd.exe和root.exe文件执行
  + xp\_cmdshell执行
  + 管理资源访问
  + 拒绝目录清单
  + DoS攻击
  + 目录遍历
  + 垃圾邮件头
* Apache web服务器错误报表
  + HTTP状态成功
  + HTTP错误网关
  + HTTP内部服务器错误
  + HTTP网关超时
  + HTTP请求URI太大
  + HTTP不支持的媒体类型
  + HTTP请求实体太大
  + HTTP禁止
  + HTTP服务器未找到
  + HTTP请求超时
  + HTTP错误请求
  + HTTP未经授权
  + 信息报表
  + 成功的报表
  + 重定向的报表
  + 客户端错误报表
  + 服务器错误报表
* Apache网络服务器报表
  + 访客
  + 用户
  + URL
  + 浏览器
  + 错误
  + 反向链接
  + Apache服务器趋势
  + Apache报表概览
* Apache web服务器的攻击报表
  + SQL攻击报表
  + 跨站脚本报表
  + D目录遍历
  + 恶意URL请求

用户活动报表：

* User Logons
* User Logoffs
* Failed Logons
* Successful User Account Validation
* Failed User Account Validation
* Audit Logs Cleared
* Audit Policy Changes
* Objects Accessed
* User Account Changes
* User Group Changes

## 1. Web日志分析概述

Web日志由Web服务器产生，可能是Nginx, Apache, Tomcat等。从Web日志中，我们可以获取网站每类页面的PV值（PageView，页面访问量）、独立IP数；稍微复杂一些的，可以计算得出用户所检索的关键词排行榜、用户停留时间最高的页面等；更复杂的，构建广告点击模型、分析用户行为特征等等。

在Web日志中，每条日志通常代表着用户的一次访问行为，例如下面就是一条nginx日志：

222.68.172.190 - - [18/Sep/2013:06:49:57 +0000] "GET /images/my.jpg HTTP/1.1" 200 19939

"http://www.angularjs.cn/A00n" "Mozilla/5.0 (Windows NT 6.1)

AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.66 Safari/537.36"

**拆解为以下8个变量**

* remote\_addr: 记录客户端的ip地址, 222.68.172.190
* remote\_user: 记录客户端用户名称, –
* time\_local: 记录访问时间与时区, [18/Sep/2013:06:49:57 +0000]
* request: 记录请求的url与http协议, “GET /images/my.jpg HTTP/1.1”
* status: 记录请求状态,成功是200, 200
* body\_bytes\_sent: 记录发送给客户端文件主体内容大小, 19939
* http\_referer: 用来记录从那个页面链接访问过来的, “http://www.angularjs.cn/A00n”
* http\_user\_agent: 记录客户浏览器的相关信息, “Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.66 Safari/537.36”

注：要更多的信息，则要用其它手段去获取，通过js代码单独发送请求，使用cookies记录用户的访问信息。

利用这些日志信息，我们可以深入挖掘网站的秘密了。

**少量数据的情况**

少量数据的情况(10Mb,100Mb,10G)，在单机处理尚能忍受的时候，我可以直接利用各种Unix/Linux工具，awk、grep、sort、join等都是日志分析的利器，再配合perl, python，正则表达工，基本就可以解决所有的问题。

例如，我们想从上面提到的nginx日志中得到访问量最高前10个IP，实现很简单：

~ cat access.log.10 | awk '{a[$1]++} END {for(b in a) print b"\t"a[b]}' | sort -k2 -r | head -n 10

163.177.71.12 972

101.226.68.137 972

183.195.232.138 971

50.116.27.194 97

14.17.29.86 96

61.135.216.104 94

61.135.216.105 91

61.186.190.41 9

59.39.192.108 9

220.181.51.212 9

**海量数据的情况**

当数据量每天以10G、100G增长的时候，单机处理能力已经不能满足需求。我们就需要增加系统的复杂性，用计算机集群，存储阵列来解决。在Hadoop出现之前，海量数据存储，和海量日志分析都是非常困难的。只有少数一些公司，掌握着高效的并行计算，分步式计算，分步式存储的核心技术。

Hadoop的出现，大幅度的降低了海量数据处理的门槛，让小公司甚至是个人都能力，搞定海量数据。并且，Hadoop非常适用于日志分析系统。

## 2.需求分析：KPI指标设计

下面我们将从一个公司案例出发来全面的解释，如何用进行**海量Web日志分析，提取KPI数据**。

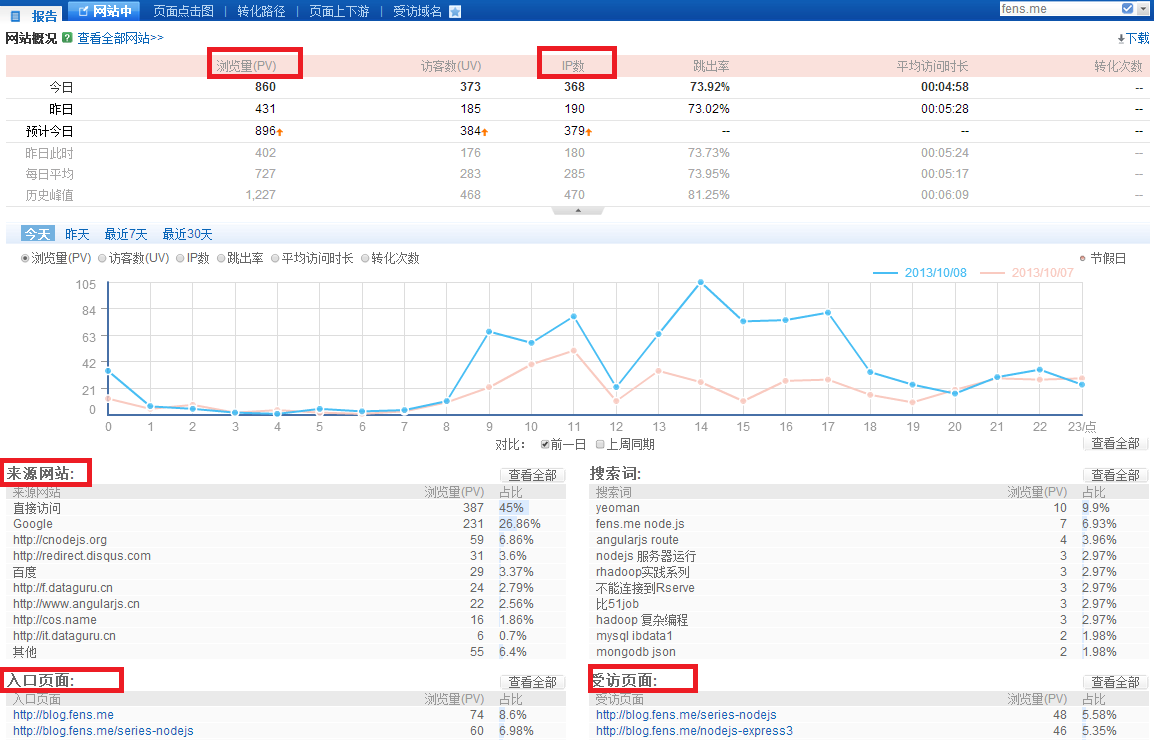
**案例介绍**  
某电子商务网站，在线团购业务。每日PV数100w，独立IP数5w。用户通常在工作日上午10:00-12:00和下午15:00-18:00访问量最大。日间主要是通过PC端浏览器访问，休息日及夜间通过移动设备访问较多。网站搜索浏量占整个网站的80%，PC用户不足1%的用户会消费，移动用户有5%会消费。

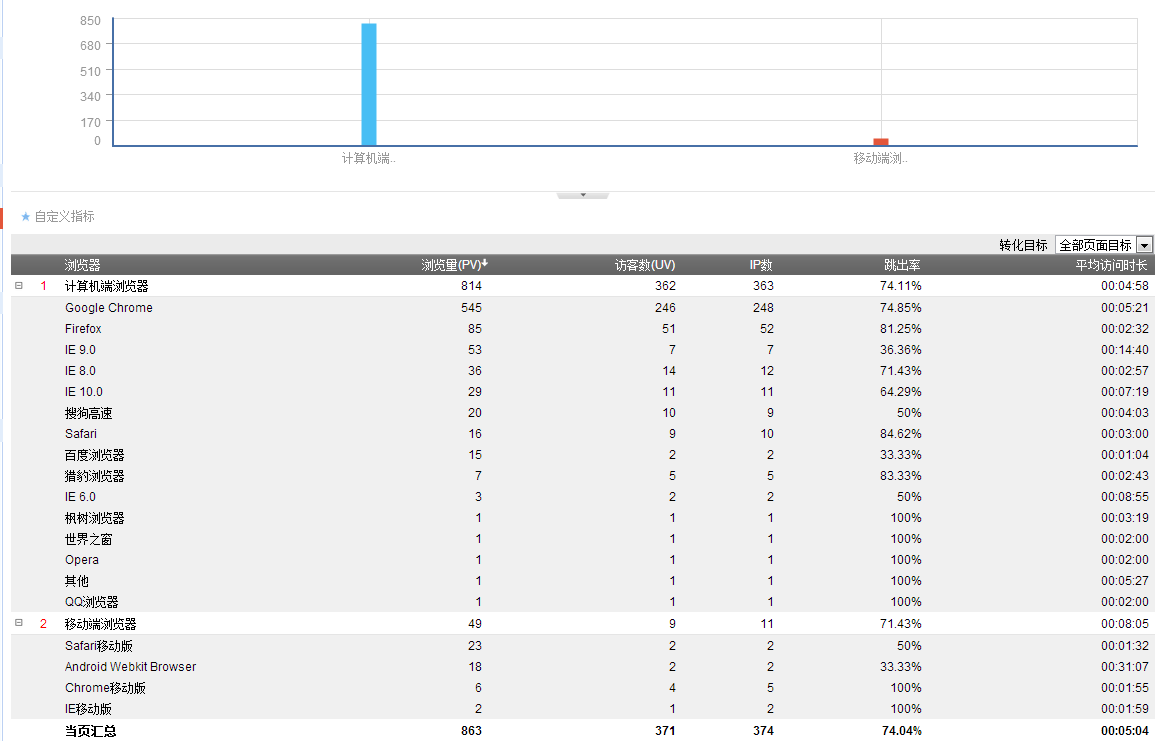
通过简短的描述，我们可以粗略地看出，这家电商网站的经营状况，并认识到愿意消费的用户从哪里来，有哪些潜在的用户可以挖掘，网站是否存在倒闭风险等。

**KPI指标设计**

* PV(PageView): 页面访问量统计
* IP: 页面独立IP的访问量统计
* Time: 用户每小时PV的统计
* Source: 用户来源域名的统计
* Browser: 用户的访问设备统计

下面的内容，将以我的个人网站为例提取数据进行分析。

[](http://blog.fens.me/wp-content/uploads/2013/10/hadoop-kpi-baidu.png)

**用户的访问设备统计指标：**  
[](http://blog.fens.me/wp-content/uploads/2013/10/hadoop-kpi-baidu2.png)

从商业的角度，个人网站的特征与电商网站不太一样，没有转化率，同时跳出率也比较高。从技术的角度，同样都关注KPI指标设计。