电商数据整合的阶段、难题与解决方案

讲电商数据前，不得不先说到网站分析。其实我们一直说的网站分析最起码需要包括以下三块：

- 数据收集
- 数据报告
- 数据分析

电商数据整合的演进肯定是从数据收集开始的，而且数据整合碰到的主要问题都是从数据收集产生的。

数据收集

说到网站分析这个话题，避免不了提起数据收集，也避免不了说到谷歌的Google Analytics (GA)，或者是百度统计这些网站分析工具(web analytics / wa)，甚至另一个国内的CNZZ统计，它们都是基于JavaScript收集数据。

我写过的一本书里就有提供过200多个网站分析工具的工具大全：

http://cn.analyticsbook.org/the-big-list-of-analytics-tools/

比如以下就是Google Analytics的Universal Analytics版本的监测代码：

```
<script>
(function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
(i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
m=s.getElementsByTagName(o)
[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
})(window,document,'script','//www.google-analytics.com/analytics.js','ga');
ga('create', 'UA-XXXXXXXX-Y', 'auto');
ga('send', 'pageview');
</script>
```

顺便一提：我是GA最早一批使用者，当时Google收购了一个叫Urchin的网站分析工具后，在2005年年底推出了GA免费版让大众使用，我在2006年开了个独立域名博客/网站（网址是 http://www.gordonchoi.com ），并在同一年安装了GA。所以我最早的一个GA帐号是一个6位数的帐号，你会发现很多国内的网站的GA帐号（如果有安装GA）一般是8位数甚至9位数。今年再新开的GA帐号都进入10位数了。

GA“看来”是最早期的wa工具，但并不如此。比起GA还有早出现很多年的，其实有网站日志数据。如果我们将互联网算作是1994/1995年左右出现的，那网站日志数据也是在1994/1995年开始有的。

这里先讲网站分析里最小单位的数据指标。

- 基于JavaScript原理的wa工具（比如GA），监测（汇报）的最小数据指标是单个pageview（网页浏览次）。
- 而网站日志数据的最小单位数据指标是单个hit。

网站日志是这样的：每次用户访问你网站的一张网页，网站服务器都会生成一条记录，甚至（在很多情况下是会生成多条记录）。我们举个相对简单的例子，一般来说一张网页是由多个文件组成的，html部分是一个文件，css是一个文件，JavaScript是一个文件，网页上有3张图片就会有三个文件。而访问一个网页时，html文件会产生一条记录，css会产生一条记录，如此类推，所以这个例子的网页被访问时，网站日志会产生一共6条记录。每一条记录就被计算做一个hit。

久远时代的网站/网页并不复杂，很多网页就只有一个html文件，所以当时用hits去计算网站的浏览次是可以的。但随着整个网站和单个网页的结构变得越来越复杂，继续用hits计算就太奇葩了。

以下是一个典型的网站日志记录（即我们说的hit）。用户使用IP：192.168.22.10地址、成功访问了网站首页（/）（即HTTP的返回码是200）、流量来源是谷歌（www.google.com）、用户使用的是火狐浏览器。


```
192.168.22.10 - - [21/Nov/2003:11:17:55 -0400] "GET / HTTP/1.1"
200 10801 "http://www.google.com/search?q=china+seo&ie=utf-
8&oe=utf-8 &aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7)
Gecko/20070914 Firefox/2.0.0.7"
```

我写过的一本书里也写过网站日志分析：

http://cn.analyticsbook.org/web-server-log-analytics/

接着基于JavaScript记录网站访问数据的工具出现了，就顺势用上pageview，用户也比较容易理解。

网站日志数据监测有优势也有短板

网站日志分析的优势：

无埋点：你是真的不需要预先为了日后的数据监测去埋点，网站服务器里的日志记录是服务器自带的，所以也并不需要在网站/网页里添加任何监测代码。

会让人产生混乱的概念：最近两、三年国内会有一些JavaScript原理的wa工具会通过某个“卖点”比如“无埋点”，去推广自己的wa工具，但这些工具并不是真正无埋点，它们其实是要做预先的“全埋点”。

只需要记住很简单的两点：

- 网站日志是网站分析的老祖宗。
- 网站日志是无埋点的老祖宗，并且是唯一的无埋点方式。

网站日志分析的短板：

漏计：某些场景里会有一定量的数据漏计。比如当网站使用文件缓存，一些多次访问网站的用户，会因为浏览器中有旧的缓存文件（比如图片，css，等），客户端并不需要发请求到服务器端去再次读取文件，导致该次的访问数据并没有被服务器记录。

我写过的另一本书里也写过浏览器缓存：

http://cn.mobilewebsitebook.com/implement-web-browser-caching/

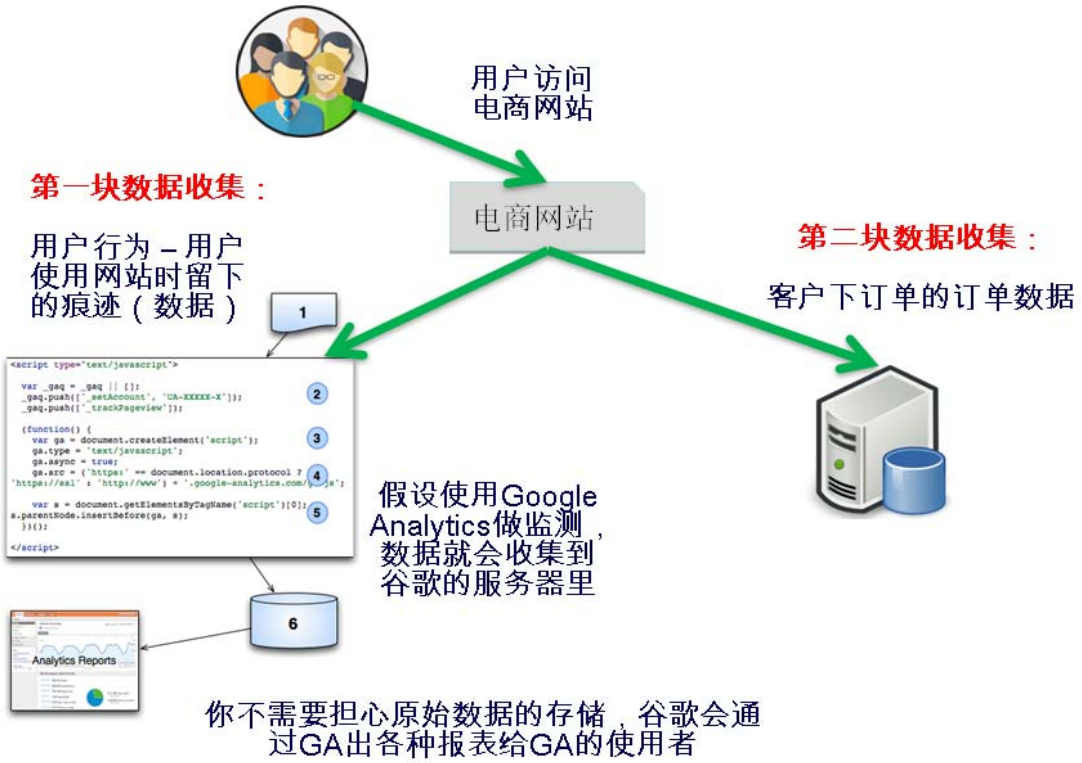
电商数据整合的第一个难点

当然一直以来的很多年，大部分网站都习惯了通过JavaScript原理去监测/收集网站数据。在只有PC端的世界里的时候，当时电商的数据其实主要有两块需要做整合。

第一块：用户行为数据。这个包括了用户访问网站后，我们会用比如sessions（会话次）、page views（浏览次）、new users/visitors（新访问者）去表达。

第二块：客户在网站上下订单的订单信息/数据。

电商网站数据收集



主要的问题是，这两块数据是分离的，比如一个客户他下单购买了3个商品（你从第二块数据里知道了他的订单信息），你并不清楚他在网站上的很多行为（第一块），比如：

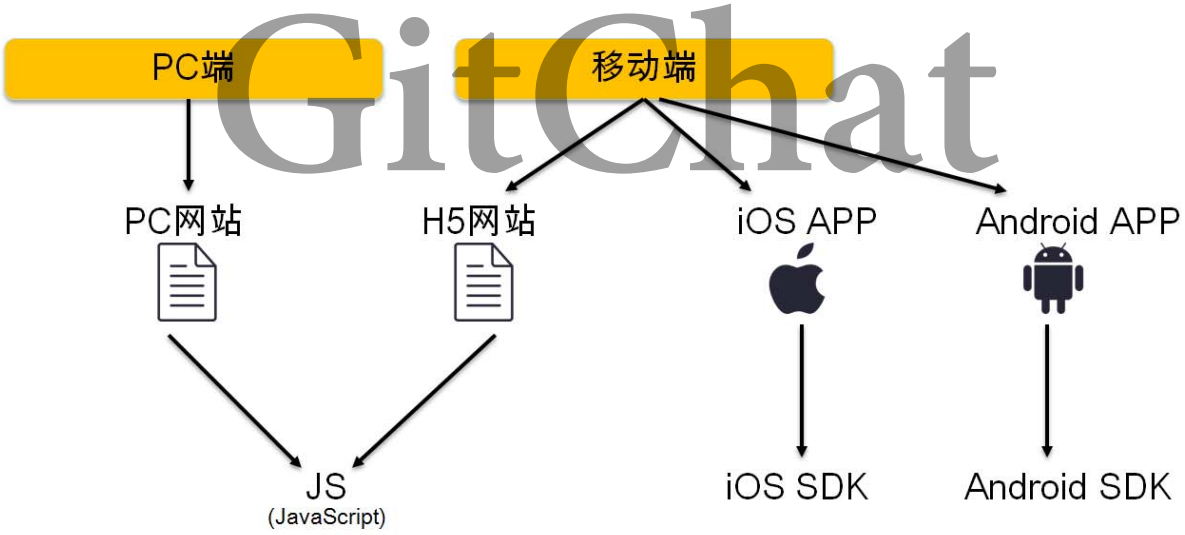
- 客户的外部渠道来源是什么？
- 客户访问网站后访问了那些网页？
- 客户一共访问过你的网站多少次后，才完成第一次订单？
- 其实还有很多其他的信息。

电商数据整合的第二个难点

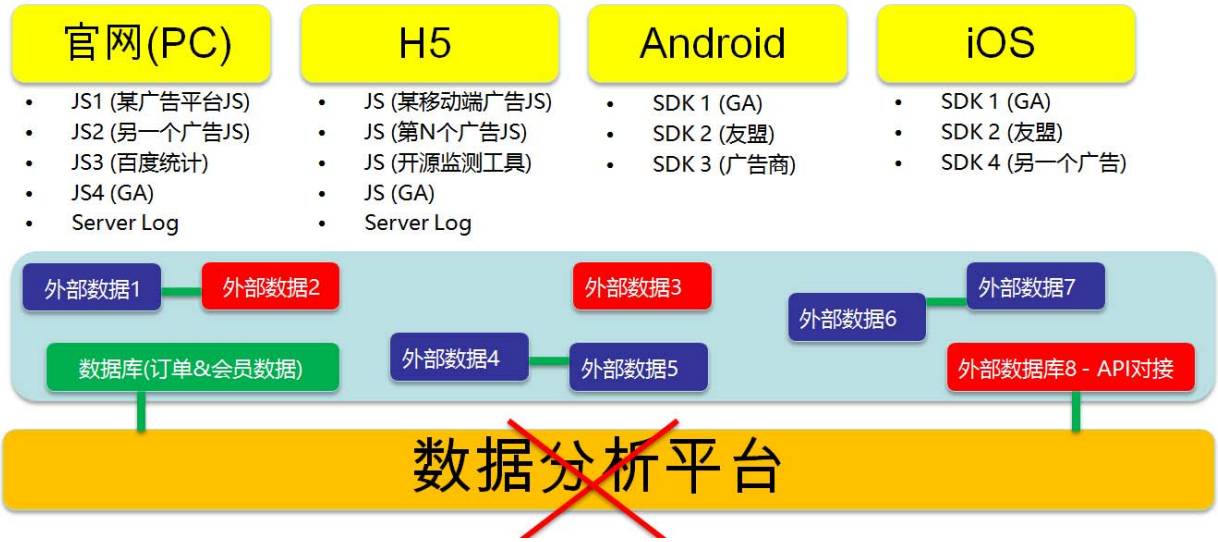
由于移动端（智能手机：iPhone、Android）的出现，后来大部分电商除了网站（作为客户购买商品平台/载体），还会有手机APP。这就让数据收集方面变得更复杂。

就网站而言，为了符合手机端用户在手机浏览器上访问网站的需求，很多电商也会另外新开发移动端兼容的手机网站 – 国内很多企业会把它叫做H5网站。H5网站的数据收集原理还是跟PC网站的一样，通过JavaScript监测代码。

移动端APP本身也分裂出不同的手机品牌/操作系统（OS），最普遍的是苹果iOS和谷歌Android。数据收集的原理跟网站彻底不一样了，APP需要通过SDK去收集数据。



这些数据收集后，大部分都保存到不一样的服务器/数据库里，所以后续就会出现整合的任务。



Web上的网站对用户身份的唯一标识是cookies，而手机端APP对用户身份的唯一标识是一些ID，比如iOS上是IDFA，Android上是AID，当然也有一些其他的ID，比如IMEI。提到唯一身份ID（标识）是因为你需要用某个“数据点”去把保存在多个地方的原始数据整合，而唯一身份标识是能把多处数据源里的原始数据“关联”起来的一个重要因素。

我们把移动端的场景放进去我们之前给网站列过出来的问题里。

- 客户的外部渠道来源是什么（从什么地方下载APP，并激活APP）？
- 客户访问APP后访问了那些APP里的网页（screens）？
- 客户一共打开过你的APP多少次后，才完成第一次订单？
- 以上客户，是否也有访问过你的网站（PC网站、H5网站）？
- 还是有其他很多问题。

数据报告

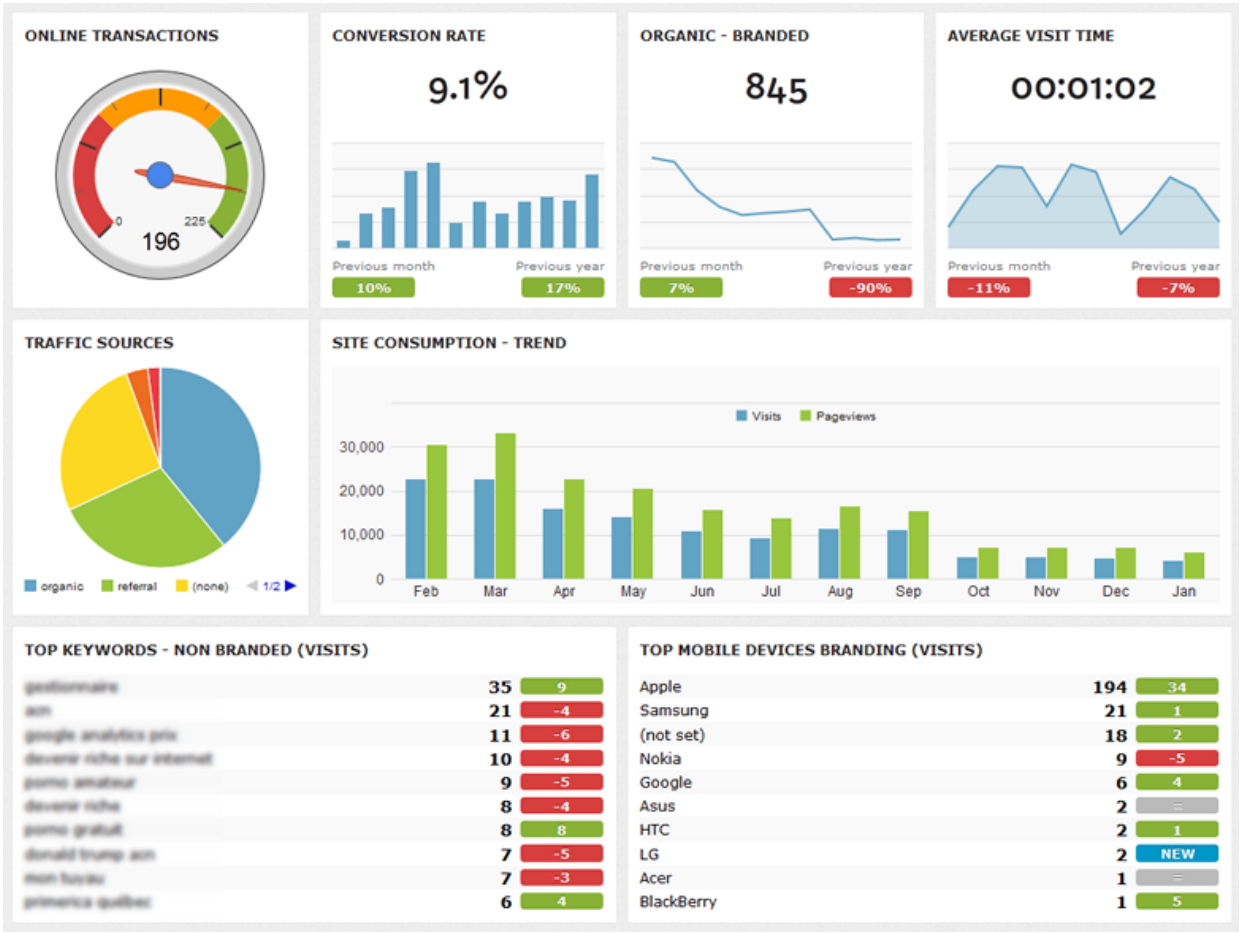
数据收集后，接下来就是抽取出网站（和终端）的数据。在第一阶段收集的原始数据需要转化成报告，有两个主要目标：

- 创建常规性数据报告
- 创建临时性数据报告

常规性数据报告：

这类报告需要每天一次，每周一次或者每月一次的规律性的被收到。基于报告的接收者，这类报告被分门别类到不同的水平。一个行政管理人员（比如公司的CEO）需要高层级的报告显示公司每一个主要部门主要的收入数据，比如一个dashboard数据报告。运营经理将需要中层级的数据报告，这类报告允许他们追踪每个独立团队负责的产品中的“潜在问题”。

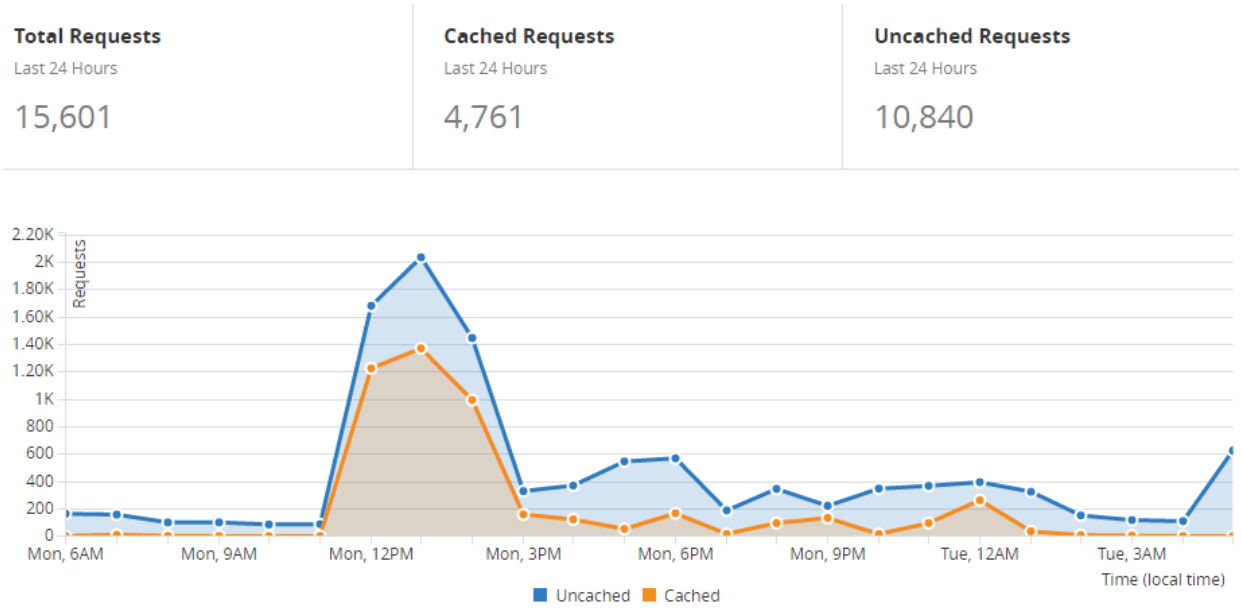
以下是一个dashboard报告的例子：



临时性数据报告：

这类报告不会以任何固定的周期运行。通常情况下，临时性报告是为了查看某个一次性的线上活动目的而产生的。当需要深入挖掘数据发现问题时也会需要临时性报告，这些问题可能是某个既定的KPI数字在过去两周内下降，运营经理需要比较详细的临时性报告去找到藏在下降背后的原因。

以下是流量（访问次）的波动：

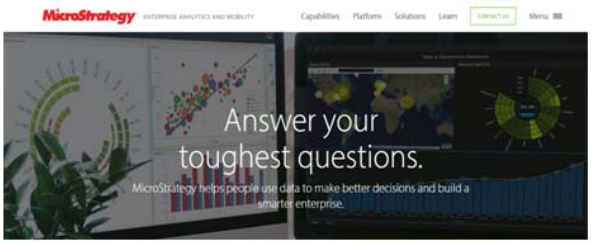


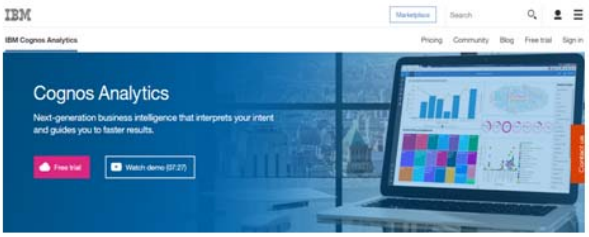
工具

工具方面，一种方式是可以使用最传统的Excel去创建和更新表格数据报表，甚至制作可视化报告。

另一种方式是使用一些开源的或付费的可视化工具，去创建并更新数据报告，比如：Microstrategy、IBM Cognos、Tableau、Pentaho、等。

以下是一些可视化工具：





数据分析

当你的电商网站已经有几十万，甚至几百网几千万的用户时，数据的量级已经比较巨大，用户在你网站上的行为数据就会有更多。这时候，通过大数据的方式当然是可以在某程度上帮助到你找到一些问题，比如某一群用户都喜欢购买某几类商品，她们购买完后再多少天内会回来网站做一些什么，等。通过大数据理解她们的行为后，你就能搭建一套给她们推荐商品的体系。最终目的是希望能提升订单转化率，替网站增加更多收入。

假如还没做到能用大数据的境界，其实也是有很多平时“人工”能够分析的地方。比如，你的网站一般正常的跳出率（bounce rate）都在50%左右，某几天网站某部分的跳出率特别高，达到80%以上。这就是一个方向，根据这个方向你继续往数据里挖，就会开始陆续理解到为什么，然后再想对应的办法去解决具体的问题。

最后

最后，我会把电商数据整合分成主要三大块：

- 电商数据整合的演进阶段
- 电商数据整合的主要难处
- 电商数据整合的解决方案