大道至简的数据分析方法论三部曲

从入门理论到深度案例剖析，
只需三步！让你快速成为数据分析师！

导读

大道至简的数据分析方法论三部曲，作者通过深入浅出的文笔总结出了一套易学易用的数据分析方法论，让你快速掌握数据分析方法中最核心、最常用的要点，满足90%的日常需求。

作者简介

AUTHOR

王桐，北京航空航天大学工学硕士，拥有8年商业智能领域的产品销售、市场营销经验，此前效力于甲骨文和IBM，均在咨询、销售岗位担任重要职位，曾成功推进多个大型项目的实施，在电商、政府、金融、互联网等行业积累了丰富经验。王桐目前主要负责产品销售和渠道拓展，已为上百家企业用户提供了完善的数据可视化分析解决方案，这些企业既有宝宝树等电商领域的明星公司，也有中国移动等传统巨头。

大道至简的
数据分析方法论

→

01

学习对大多数人而言是一件痛苦的事情，尤其看着厚厚的专业书籍、各种难以理解又缺乏解释说明的术语定义，会让这种痛苦加剧。但是有些书或文章能将复杂的理论用非常通俗、口语化的方式讲来，让读者不费劲，一下就能明白。这些内容实在是读书人的一种福音。说到底，互联网思维中的用户思维谈了这么久，教育、培训类内容的创作者们也应该好好改变一下，站在读者的角度说话了。

本文谈的是数据分析方法。根据笔者对众多企业的接触和了解，虽然现在大部分企业都对数据越来越重视，但目前仍有相当多的企业和从业者还没有摸清数据分析的门道，不知道自己的数据该怎么分析，希望得专业人员的到帮助。

▼ | 数据分析方法一点也不神秘

笔者以前学习数据分析方法时也很痛苦，看了不少书，内容很多，但难以记全，更难以运用，后来加入永洪科技给众多企业做数据分析系统，通过大量的项目实践，才慢慢能谈得上入门。

好的方法论应该是易学易用的。现在，本文就努力尝试用最简单易懂的文笔，让初学数据分析的人看完就能理解并掌握数据分析方法中最核心、最常用的要点，至少能满足90%的日常需求。做到这一点，必须将博大精深的数据分析方法提炼成人们能记得住的3点，而不是30点，再浓缩到一篇文章的篇幅，而不是一本书的厚度。

01 数据分两种，维度和度量，分析就是维度和度量的组合

下面是一个最简单的消费者购物的数据例子。

订单 ID	用户 ID	地区	年龄	订单金额	订单商品	订单时间
1	99	北京	19	126	T 恤衫	2014/10/8
2	1008	北京	14	80	牛仔裤	2014/9/1
3	27	上海	24	309	衬衫	2014/3/14
4	67	北京	22	286	衬衫	2013/5/25
5	983	北京	21	222	毛衣	2013/12/14
6	266	上海	31	560	西服	2014/1/8
7	54	上海	25	313	衬衫	2012/6/6
8	498	广州	22	275	衬衫	2012/11/9
9	1209	北京	24	299	牛仔裤	2013/4/1
10	709	北京	18	120	T 恤衫	2014/8/10

先不管这个数据表是存在excel里还是数据库里，只关注数据本身。表里涉及到的数据项（或者叫字段）有“订单ID”、“用户ID”、“地区”、“年龄”、“订单金额”、“订单商品”、“订单时间”。

这些数据项有什么差异呢？总体而言，数据分两种，一种叫维度，一种叫度量（或者叫指标）。上面这个例子里，“订单金额”是度量，其余数据项都是维度。

可以看出，度量是具体的计算用的量化数值，而维度是描述事物的各种属性信息。我们在做数据分析时，归根结底就是在不停的做各种维度和度量的组合，比如北京地区的订单金额总和，21到30岁用户的订单金额平均数；或者单独对维度和度量进行数学公式计算，比如所有的订单金额总和，用户数（用户ID的不重复计数）等等。

从数据类型上看，度量都是数值，但是数值不一定是度量，比如订单ID，虽然是数值，但是不是度量而是维度，而时间、文本类的数据都是维度。

有一点需要格外注意，维度和度量是可以转换的。比如要看“年龄”的平均数，这里的“年龄”就是度量，要看19岁用户的订单情况，这里的“年龄”就是维度。对于一个数据

项而言，到底它是维度还是度量，是根据用户的需求而定的，很像量子效应，状态只有需求确定后才会随之确定。

另外，维度可以衍生出新的维度和度量，比如用“地区”维度衍生出一个大区维度，“北京”、“天津”都对应“华北大区”，或者用“年龄”维度衍生出一个年龄范围维度，20到29岁=“青年人”，30到39岁=“中年人”，40到49岁=“资深中年人”。再比如上述的平均年龄，就是用“年龄”维度衍生出一个度量。

度量也可以衍生出新的维度和度量，比如用“订单金额”度量衍生出一个金额范围维度，100元以下对应“小额订单”，500元以上对应“大额订单”等等。再比如用“收入”度量和“成本”度量相减，可以得到一个“利润”度量。

02 做判断用对比

下面提出一个问题：企业A今年收入8000万，是高还是低？大家看着这个问题，应该会感到无从判断，因为没有参照物，即没有对比。因此，拿到一个数据，要判断是好是坏是高是低，必须要进行对比。

首先，企业A可以跟自己比。如果前年收入2000万，去年收入4000万，那今年8000万算很好了。去年收入1个亿，今年8000万就是糟糕了。这叫纵向对比。

其次，企业A也可以跟其他人比。同行的几家竞争对手企业今年都收入几个亿，那企业A的8000万就不理想。这叫横向对比。

第三，企业A还可以对比不同的维度和度量。比如竞争对手都做全国市场，企业A只做山东市场。企业A在山东市场的收入比竞争对手在山东市场的收入高，那么就本地区而言，企业A做的更好，而放眼全国，企业A做的就有局限。比如如果竞争对手都做了十几年，而企业A刚做四五年，那企业A就算做的不错，但如果成立的时间相仿的竞争对手已经过亿了，那企业A就算做的不够好。这叫综合对比。

孩子考试考了95分，家长很高兴，因为知道满分是100分，有参照物。最近一次考试考了80分，家长会发火，因为过去的95分成了新参照物。后来一问，发现这次卷子出难了，孩子已经是班级第一了，就又转怒为喜，这里其他孩子就成了参（xi）照（sheng）物（pin）。

对比的参照物不同，得到的判断结论也就不同。为了避免结论片面、不客观，应该尽量多用综合对比。

03 找原因用细分

今年利润下降了，老板很生气，下令查找原因，缉拿“嫌犯”。原因怎么找呢？注意是找原因，不是找理由。很多人往往不知道如何查找原因，最后给出的都是理由。

先看一个示例的原因结论是什么——“因为四季度华南区域洗衣机的销量下降了，导致了今年利润的下降”。让我们分析一下这个原因有什么特点。

我们会发现，这个原因是由时间、区域、产品这三个维度和销量这一个度量组成的，于是我们可以知道，对于问题原因的查找定位，本质上就是在回答哪些维度下的哪些度量的下降或上升，导致了问题的发生。

这就是在做细分。

我们可以按维度细分，有多少维度，就可以有多少种细分的方向。比如看是去年所有月份都下降了，还是只有某几个月下降。如果是后者，那么就可以缩小查找的数据范围。聚焦到这几个月后，可以再看是哪些区域下降了，进一步细分。

入手的维度的先后顺序影响不大，问题原因涉及的维度也无法预知，因此可以从任意一个维度作为入口开始进行细分。

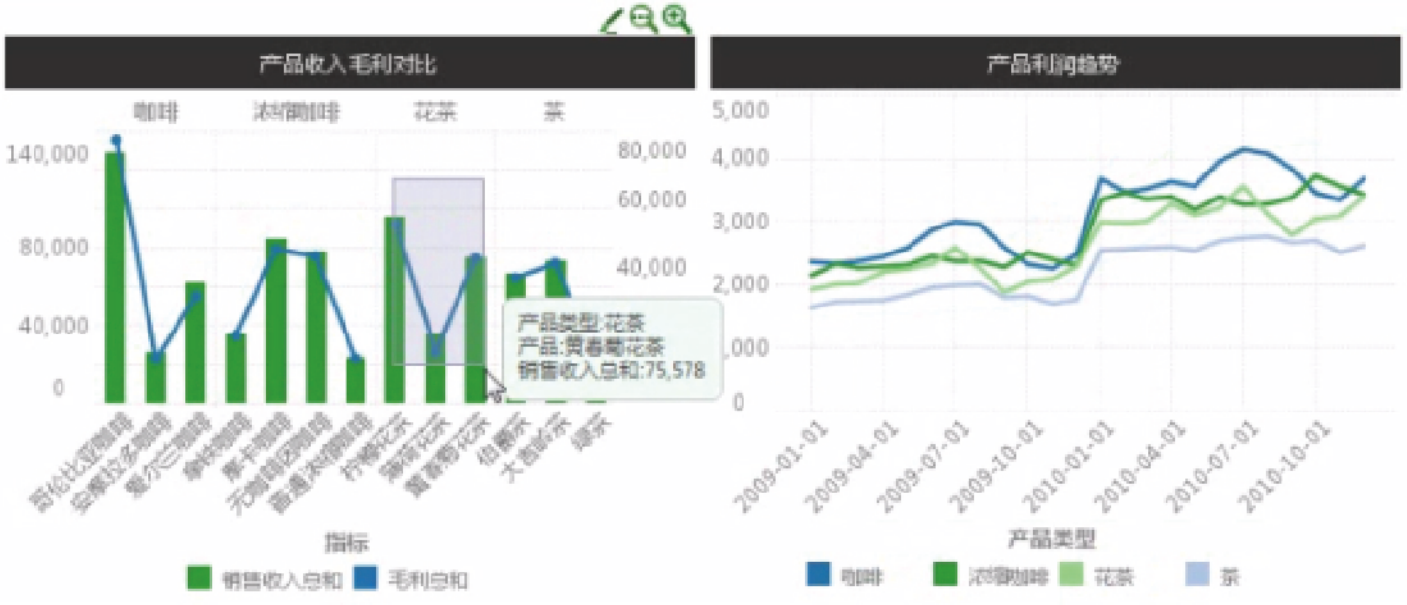
如果出问题的指标有相关的先导指标，则要想进一步挖掘问题原因，细分后还要看不同的度量，比如上述的原因结论示例是“因为四季度华南区域洗衣机的销量下降了，导致了今年利润的下降”，问题是“利润”而原因是“销量”，因为利润是通过别的度量计算衍生出来的。

细分无止境，细到什么地步才够呢？答案是，到可操作的区间才够。

比如就细分到“四季度利润下降，其它季度没有下降”，还是没有解决问题的办法，必须细到哪个时间段哪个区域哪条产品线，直到细到某一个最终责任人，才具有可操作性。需要注意的是，在真实情况中，问题往往不一定只有一个原因，而是多个原因综合起来形成的。

我司永洪科技主推的一站式大数据分析平台软件，为什么提供“缩放”和“笔刷”两种交互操作，就是为了满足“对比”和“细分”两种场景。

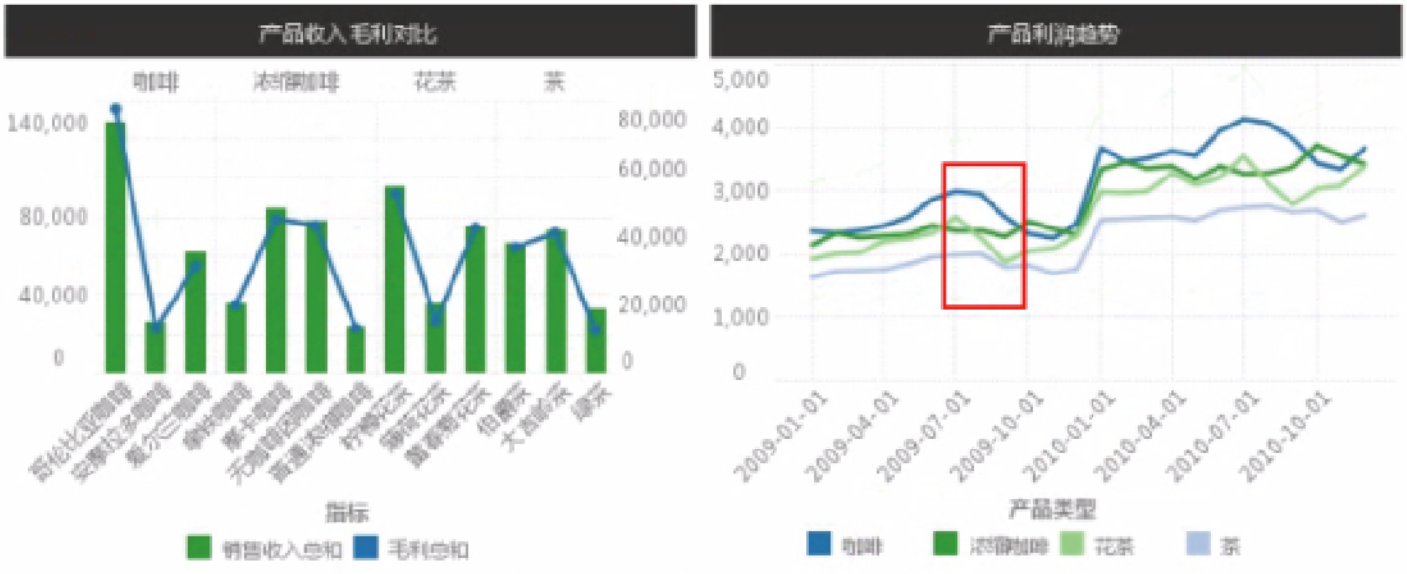
举一个例子，如下图，左图是各产品的收入毛利对比，右图是各品类利润趋势，现在用户想聚焦到“花茶”品类下的三种产品上，看看它们的利润如何。



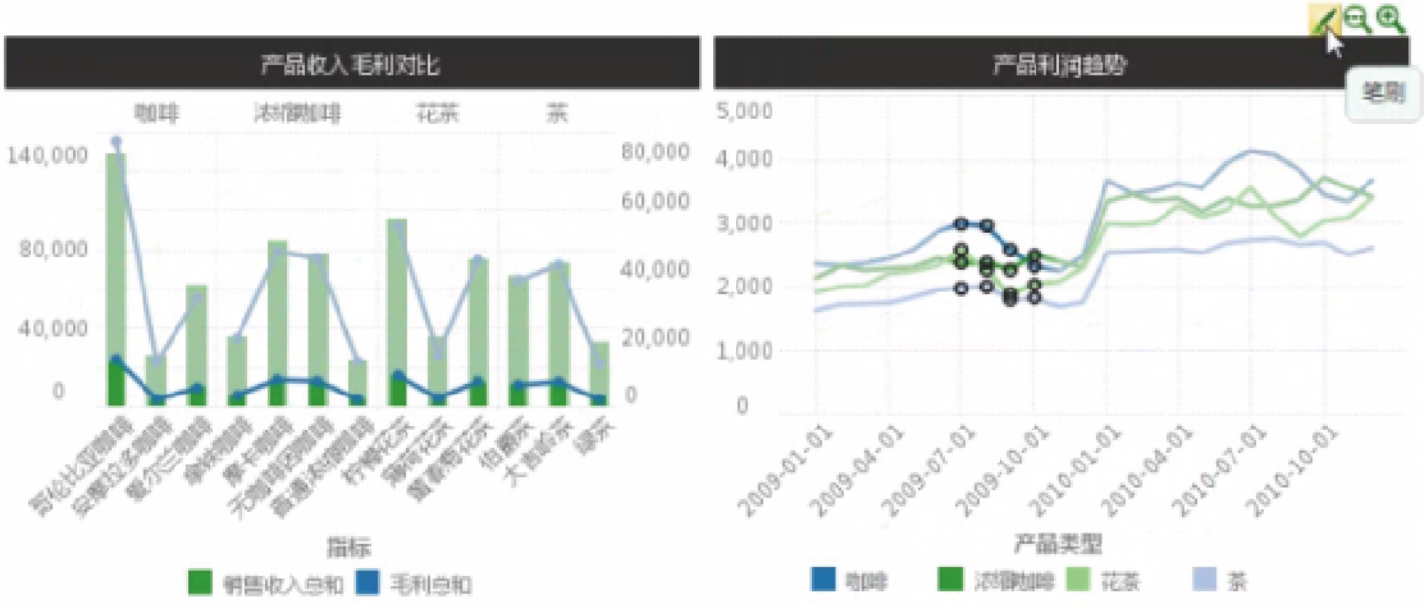
这时用户就可以使用“缩放”功能，圈选代表这3种产品的3根柱子，点击“缩放”按钮，这时左边图表只剩下这3种产品，而右边的利润趋势则显示这3个产品的利润总和趋势。这就是在做“细分”。

有人可能会问，这个效果很类似筛选，为什么不在旁边放一些筛选器来实现呢？筛选器可以有，但现实情况中，当我们在一个图表上发现问题，不一定就能很容易地找到与其对应的筛选条件，尤其是散点图。因此，直接在图表上选择会非常方便高效。

再举一个例子，下图是产品利润趋势分析，用户发现从2009年7月开始，利润有连续4个月的下滑（如红框所示），用户想知道为什么。



这时用户就可以使用“笔刷”功能，在趋势图上选中这4个月的点，点击“笔刷”按钮，同一报告页面的其他图表就会淡化，然后突出显示用户选中的7到10月在这个图表上的占比，所以下图中左边的图表高亮显示出的矮的绿柱子，就是这些产品在这4个月的销售收入。



与“缩放”不同，“笔刷”方便用户将局部数据和整体数据进行对比。因为在上面这个例子中，单纯看哪些产品这4个月销售收入的绝对值低，并不能说明什么，有些产品本来卖的就少，一定要看哪些产品在这4个月相对表现不好。

先判断数据好不好，再分析原因是什么，数据分析的环节链条基本就算完整了。

▼ | 怎么看待机器学习/数据挖掘等这类高大上的内容?

什么时候去碰机器学习/数据挖掘这样高大上的内容？一句话，先把上述的数据发分析方法做到游刃有余，再搞那些高大上的。不要迷信复杂的算法，很多企业内部数据分析的大拿，往往都是深度理解业务，用的都是普通的计算方法，就能完成很精彩实用的分析过程。

机器学习/数据挖掘等什么时候会用到？简单而言，数据项多到人眼看不过来的时候会用到。如果总共就十来个数据项，每个拿出来单独出张图看一眼就看出端倪了，其实就不太需要用挖掘算法。如果总共几百个数据项，想看某一个数据项是受哪几个数据项影响最大，人眼看不过来，用挖掘算法就比较合适。

大道至简的
数据体系构建方法论

——两步就让你打造出数据化运营的核心支柱！

→

02

引言：

本文是“数据化运营方法论系列”文章的第二篇。第一篇《大道至简的数据分析方法论》之后的讲的是“不知道该怎么分析”的问题，本文讲的是“不知道该分析什么”的问题。

与“不知道该怎么分析”一样，“不知道该分析什么”同样是很多人常问的问题之一。事实上，如果知道了方法，虽然不能做到没有一蹴而就，但是也能明晰如何一步步坚实地打造属于自己的数据体系路径。

与第一篇文章一样，本文会用最简单质朴的语言来讲清楚数据体系构建的路径。简单来讲，就是先梳理出数据指标体系，再将其落地到BI（商业智能，其实叫业务智能更对味）系统里。

一、由上至下地梳理数据指标体系

1 确定目标

这是第一个应该问自己的问题。花大力气做数据分析，最终为了什么呢？如果这都没想清楚，那数据体系肯定无从下手。

是想提高用户活跃度、增加用户、增加销量，还是别的什么目标？这么一想，好像我都想要。都想要没有问题，但是会让工作的边界无限蔓延，导致事情无法推进。所以，应该从最关心的那个目标/KPI入手。

那么，什么问题才是我们最需要关心的目标呢？

对于不同领域、不同阶段的公司和不同角色的用户而言，这个问题的答案都不一样：对于很多公司老板来说，利润就是他们最关心的目标；对于非售卖产品/服务的公司或政府而言，也许客户满意度是最关心的目标；对于交易平台类公司或早期电商公司而言，利润不是重点，交易量是最关心的目标。

对于单人而言，无论是老板还是执行层，同时关注的目标/KPI都不宜过多。同时看几十个KPI，想象一下也知道会很晕，且耗费时间。但是，对企业而言确实有很多KPI都是非常重要的。这该怎么办？可以分解到多人，即不同角色一起协作，每个角色关注自己的目标，所有角色合在一起是公司所有目标/KPI的全集。

假设老板最关注的目标是利润，利润=收入-成本，可以将这个目标分解为由销售总监来关注收入，运营总监来关注成本。当然，并不是说老板不能看收入，而是把常规性的关注目标锁定在一个可行的范围之内。

2 分解指标

目标确定了，下一步是分解出相关的指标。

针对目标，需要哪些指标来监控或分析能达成目标呢？比如利润，相关指标就是收入和成本，当然这太粗了，收入有哪几类，成本有哪几类，都应该考虑进去。比如对于零售行业的销售额，可以分解为客流量、进店率、购买率、客单价和复购率等。

所以，分解的方式有很多种，需要遵循MECE原则（完全穷举，相互独立）。

3 细化字段

针对指标的计算公式，涉及到哪些字段，分别在哪些库的哪些表里，是否需要数据清洗，清洗规则是什么等。

比如购买率，是通过公式“购买人数/进店人数”算出来的，购买人数又是对“客户ID”进行计数计算得出来的，这些指标涉及到的字段对应到数据库里哪张表的哪个字段，需要梳理清楚，这部分就需要IT人员或数据库管理员的介入和配合了。

4 非功能需求

上述第3步完成之后，我们其实已经算是梳理完了指标体系，可以落地了，但为了让最终形成的数据系统更加完备、友好、可用，还需要一些非功能需求的梳理。

UI：偏好什么样的展示风格，这点看着无关紧要，但实际上用户每天都会与数据系统打交道，美观、体验好的系统UI会让用户更加喜欢。

页面流：哪些相关指标摆放到同一个报告页面上，页面之间的层次关系如何，用户可以在页面之间如何跳转。

权限：谁能看哪些数据范围，谁能看哪些字段和指标，需要有统一的权限控制，避免出现数据安全问题。

ETL：数据从数据源同步到分析系统的频率如何，规则如何。

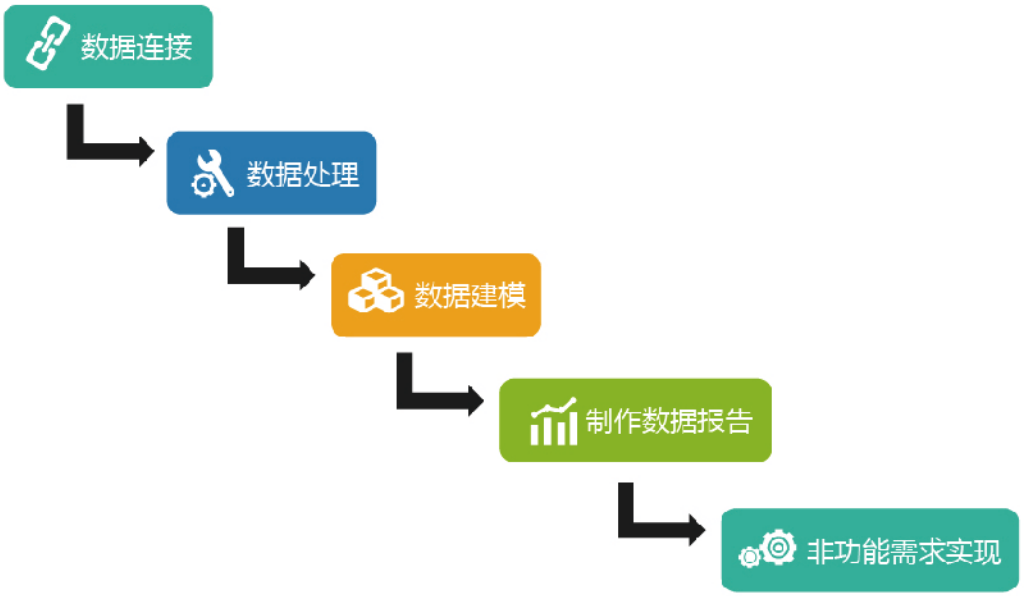
集成：是否需要在界面、预警消息等层面与其它系统进行集成。

性能：看不见摸不着，但是直接决定系统可用性。如果数据量大时需要几分钟甚至几十分钟才能看到结果，相信这个系统就不会有人愿意用了。

5 系统实施

上述4项完成之后，我们就形成了《数据运营系统需求文档/实施方案》，即可落地到数据运营系统里，然后，再根据报告页面数量、数据准备复杂度等确定工作量和时间计划。

二. 由下至上地实施落地到BI系统



1 连接数据

根据需求文档/实施方案，一步步进行系统搭建工作。这个系统有的企业称之为大数据平台，有的企业称之为BI系统。大数据平台的范畴会更广一些，但对企业数据化运营而言，BI一定是核心构成。

那么，无论是开发还是基于像永洪科技一样的第三方工具快速实施，系统搭建的第一步都是连接各个数据源，打通和各个数据源之间的通路。

在企业里，数据环境往往是异构的，数据源可能包括数据库、Hadoop系列平台、Excel文件、日志文件、NoSQL数据库、第三方接口等，需要对每种数据源都有快速友好的对接方式。

最终，我们在系统里能看到所需要的各个数据源中所有的表格和字段。

2 数据处理

数据源里的数据往往是有或多或少的不规范性存在的，比如有重复记录，比如有遗漏的空值，比如有明显不合理的异常值（比如有2020年的成交订单），还可能有同一个事物在系统中存在多个名称的情况。

这些数据如果不做一些处理或称之为清洗的工作，是会对分析的准确性产生很大影响的，所以需要做些预处理。这个过程往往是最耗时、最枯燥的，但也是十分重要的。

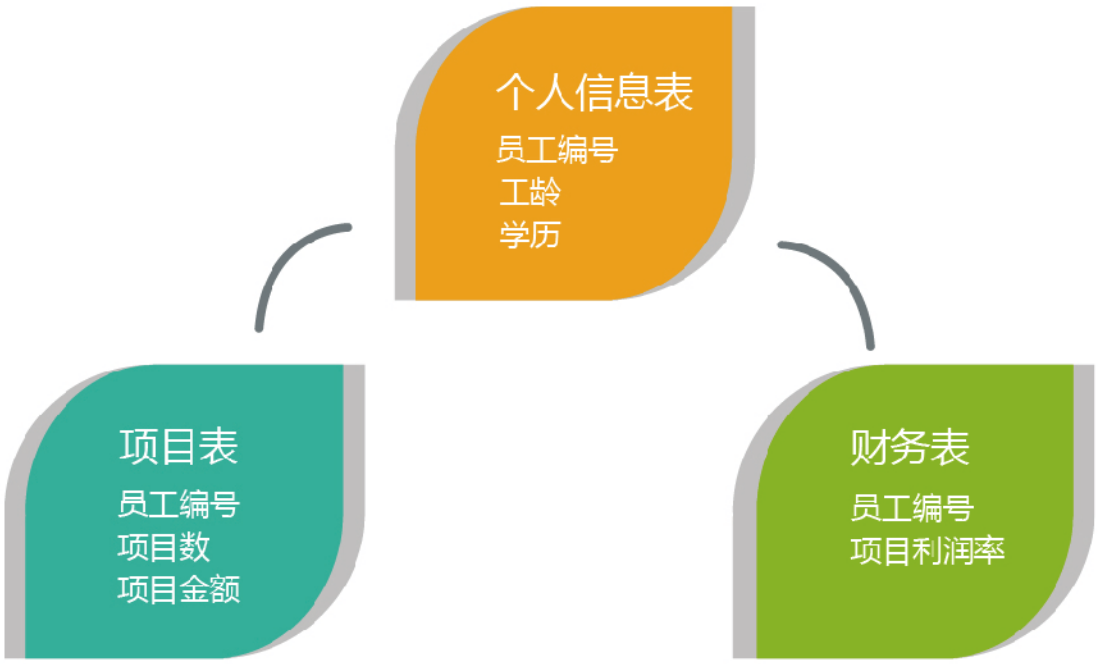
作者提醒：这个环节的问题将在下一篇《大道至简的数据治理方法论》文章中再深入探讨。

3 数据建模

数据处理好了，下一步就该做数据建模了。

一提到建模，非技术背景的用户就生畏，觉得高深不可理解。其实建出的模是个什么东西呢？简单来讲，把多张表关联到一起，就是一个数据模型。

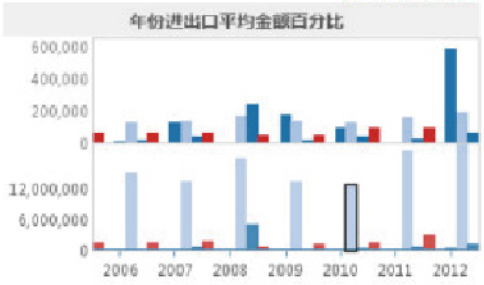
比如，公司要做绩效分析，需要员工的工龄、学历、项目数、项目金额、项目利润率等指标，其中工龄、学历在个人信息表里，项目数、项目金额在项目表里，项目利润率在财务表里，这三张表有个共同字段“员工编号”，通过这个字段把这三张表关联起来，这就是一个数据模型，一个绩效分析主题的数据模型。

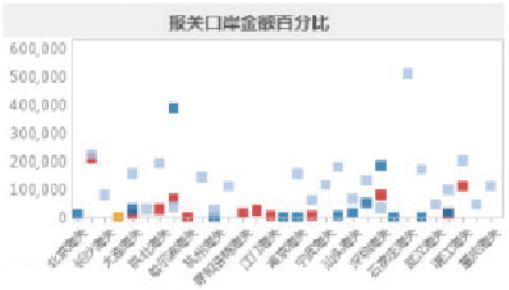


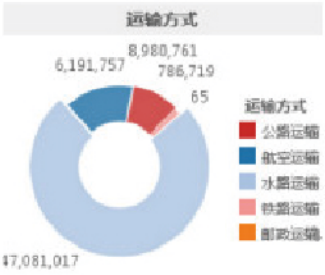
4 制作数据报告

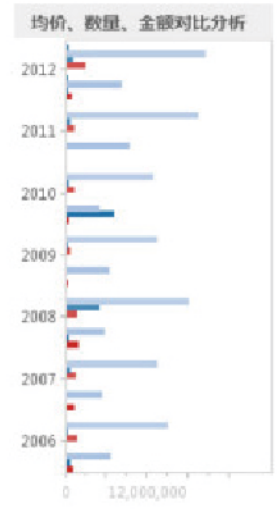
基于建好的数据模型，我们就可以开始制作数据报告了。

数据模型提供了基础数据和字段，按照需求将它们以公式进行组合，用合适的图表类型进行展示，将相关指标摆放到同一个报告页面上，配置好页面之间的层次关系和跳转关系。以下是基于永洪科技一站式大数据分析平台制作的Demo。









通关演示数据

企业名称	企业编码	总金额
慈溪徐龙进出口有限公司	3320910005	4,658,483
辽宁万融贸易有限公司	2102913129	3,162,484
烟台水星食品有限公司	3706238993	2,686,920
山东东方海洋科技股份有限公司	3706968894	2,558,995
辽宁成大股份有限公司	2102912020	2,450,558
青岛国星食品有限公司	3702948401	1,601,524
浙江省粮油食品进出口股份有限公司	3301910115	1,578,456
烟台广源食品有限公司	3706942017	1,419,264
盛龄家纺(安徽)有限公司	3412940045	1,396,885
烟台伟成食品有限公司	3706938527	1,368,916
中粮发展有限公司	1101940173	1,364,228
唐山永泰水产有限公司	1302940063	1,315,010
青岛罐头食品厂有限公司	3702941244	1,309,361
烟台天山食品有限公司	3706938114	1,209,140

5 非功能需求实现

经过第4步之后，我们的数据系统已基本成型，剩下的就是实现上述的各个非功能需求了。这样，一个完备、友好、可用的数据运营系统就上线了。

上线并不是工作的终点，业务需求时刻都会变化或新增，需要能够快速迭代调整，数据处理、建模、制作数据报告等操作需要高度工具化，以保证灵活可配置。第三方工具对比自开发的优势也在这点上体现尤为明显。

归根结底，做数据的目的要么是为了提升管理（节流），要么是业务创新（开源）。一个系统化的数据体系将是数据化运营的核心支柱。

大道至简的
数据治理方法论

→

03

引言：

数据分析师的角色犹如一位大厨，原料有问题，大厨肯定烹饪不出色香味俱佳的大菜，数据有问题，数据分析师得出的结论自然也就不可靠，再好的数据分析方法论也只是建立在失真的数据基础上，苦心构建的数据体系当然也被白白浪费了。

过往的项目中，笔者也时常遇到这样的情况，客户用永洪科技的产品做了一些精美专业的数据报告，却因数据不准而影响了报告的使用价值。

前两篇文章笔者分别探讨了面对数据指标如何分析，以及如何构建系统化的数据体系，本文是“数据化运营方法论系列”文章的第三篇，重点探讨的核心话题是——数据治理。

数据治理是一项基础工作，在很多人眼中是一项苦活儿累活儿，但是越是这样的工作越是不能忽视，基础打扎实了，上层建筑才会更稳固。

下面，笔者先从脏数据的种类及处理方法谈起。

一. 脏数据的种类及处理方法

首先，我们来了解一下脏数据的种类，明白我们可能会面对哪些问题。

1 数据缺失

缺一些记录，或者一条记录里缺一些值(空值)，或者两者都缺。原因可能有很多种，系统导致的或人为导致的可能性都存在。如果有空值，为了不影响分析的准确性，要么不将空值纳入分析范围，要么进行补值。前者会减少分析的样本量，后者需要根据分析的计算逻辑，选择用平均数、零、或者等比例随机数等来填补。如果是缺一些记录，若业务系统中还有这些记录，则通过系统再次导入，若业务系统也没有这些记录了，只能手工补录或者放弃。

2 数据重复

相同的记录出现多条，这种情况相对好处理，去掉重复记录即可。但是怕就怕不完全重复，比如两条会员记录，其余值都一样，就是住址不一样，这就麻烦了，有时间属性的还能判断以新值为准，没有时间属性的就无从下手了，只能人工判断处理。

3 数据错误

数据没有严格按照规范记录。比如异常值，价格区间明明是100以内，偏偏有价格=200的记录;比如格式错误，日期格式录成了字符串;比如数据不统一，有的记录叫北京，有的叫BJ，有的叫beijing。对于异常值，可以通过区间限定来发现并排除;对于格式错误，需要从系统级别找原因;对于数据不统一，系统无能为力，因为它并不是真正的“错误”，系统并不知道BJ和beijing是同一事物，只能人工干预，做一张清洗规则表，给出匹配关系，第一列是原始值，第二列是清洗值，用规则表去关联原始表，用清洗值做分析，再好一些的通过近似值算法自动发现可能不统一的数据。

4 数据不可用

数据正确，但不可用。比如地址写成“北京海淀中关村”，想分析“区”级别的区域时还要把“海淀”拆出来才能用。这种情况最好从源头解决，即数据治理。事后补救只能通过关键词匹配，且不一定能全部解决。

二. BI对数据的要求

接下来，我们了解一下BI对数据的要求，结合上面脏数据的种类，中间的规避手段就是数据治理。

1 结构化

数据必须是结构化的。这可能是句废话，如果数据是大段的文本，比如微博，那就不能用BI做量化的分析，而是用分词技术做语义的分析，比如常说的舆情分析。语义分析不像BI的量化分析一样百分百计算准确，而是有概率的，人的语言千变万化，人自己都不能保证完全理解到位，系统就更不可能了，只能尽可能提高准确率。

2 规范性

数据足够规范。这么说比较含糊，简单来讲就是解决了上述各类脏数据的问题，把所有脏数据洗成“干净数据”。

3 可关联

如果想将两个维度/指标做关联分析，这两个维度/指标必须能关联上，要么在同一张表里，要么在两张有可关联字段的表里。

三. 数据治理的原则

前面讲了脏数据的处理方法，但那些都是治标不治本的应对方法，且需要长期耗费大量时间和人力来做这种痛苦的工作。要想从根本上改善脏数据的问题，还是需要做好数据治理的规范工作。

简单来讲，数据治理就是要约束输入，规范输出。

1 约束输入

你永远想不到用户会输入哪些值，所以别给用户太多发挥的空间，做好约束工作。该用户填写的，系统必须设置为“必填”;值有固定选项的，一定用列表让用户选，别再手工输入;系统在录入提交时就做好检查，格式不对，值不在正常范围内，直接报错的情况必须让用户重新输入;设计录入表单时尽量原子化字段，比如上面说的地址，设计时就分成国家、省、市、区、详细地址等多个字段，避免事后拆分;录入数据保存的数据表也尽量统一，不要产生有大量相同数据的表，造成数据重复隐患。

2 规范输出

老板看不同人做的报表，同一个“收益率”指标，每张报表的值都不一样，老板的内心一定是崩溃的，不知该骂谁，只能全骂。排除计算错误的情况，一般都是统计口径不一致造成的。所以要统一语义，做一个公司级别的语义字典(不是数据库的数据字典)。所有给人看的报告上的指标名称，都要在语义字典中备案，语义字典明确定义其统计口径和含义。不同统计口径的指标必须用不同的名词。如果发现一个词已经在语义字典中有了，就必须走流程申请注册一个新词到语义字典。

四. 数据治理的落地

脏数据的处理需要ETL工具，语义字典不一定要借助于系统。事实上，由于这类系统过于复杂，国内鲜见实施成功的案例，用Excel加制度就能达到很好的效果。

关于落地推广策略，说来也简单，老大拍板说必须实行，再用优先话语权吸引一个部门试点，再横向扩展。哪个部门先落地，哪个部门就能按最符合自己习惯的用词来命名指标，相当于占坑。后面的部门都要遵从前人的标准，重名但意义不同的指标需要另外找词儿命名。这样就不怕没人积极主动。

以上，就是精炼版的数据治理方法论。大家都知道这是个苦活，但是笔者还要提醒的是，越晚动手越苦。有了经验以后，做新业务系统设计时，大家就可以充分考虑数据治理的规范了。

微信公众号：永洪科技
电子邮件：public@yonghongtech.com
电话联络：400-900-2326
官方网站：www.yonghongtech.com