参考：<http://www.cyzone.cn/a/20160325/292702.html>

**大道至简的数据治理方法论——如何处理你手中的各种“脏数据”？**

永洪科技副总裁王桐2016-03-25

标签: [数据分析](http://www.cyzone.cn/tag/%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90/)[数据治理](http://www.cyzone.cn/tag/%E6%95%B0%E6%8D%AE%E6%B2%BB%E7%90%86/)[产品运营](http://www.cyzone.cn/tag/%E4%BA%A7%E5%93%81%E8%BF%90%E8%90%A5/)



如果你是一位大厨，刚刚眉飞色舞地给客人描绘了如何搭配一道色香味俱佳的大菜，甚至连炒菜的手法都一一交代了，当你备好了各种为这道菜增鲜增色的调料后准备烹饪时，才发现所需的主要原料有问题。

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，笔者有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。

第一篇[《大道至简的数据分析方法论》](http://www.cyzone.cn/a/20160304/291405.html)

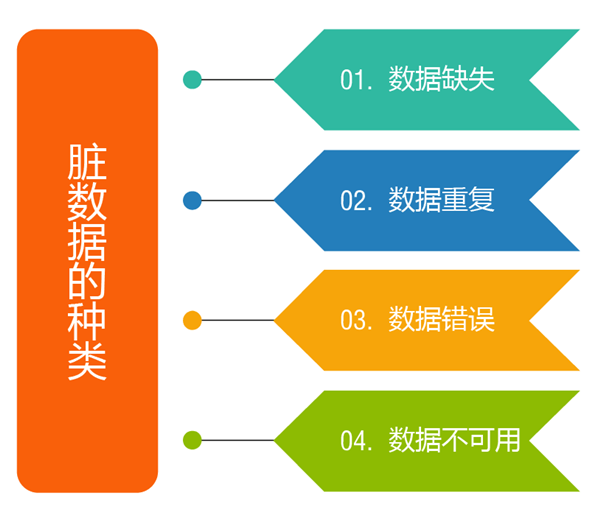
第二篇[《大道至简的数据体系构建方法论》](http://www.cyzone.cn/a/20160313/291400.html)

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。

*由于时间关系和水平有限，文中或有不妥之处还请读者多多谅解。如果希望一起探讨数据分析方法论和数据体系构建方法论等相关话题，欢迎读者通过tylerwang@yonghongtech.com或微信号tyler\_wangtong与本文作者永洪科技副总裁王桐取得联络。*