数据质量问题分析及对应的解决办法

原创：冯海文 数据工匠俱乐部 4月17日

↑点击蓝字，轻松关注

导读

数据的重要性已经在业内已达成普遍共识，数据是企业最具竞争力的战略资产。但其实数据的价值是应该乘上一个质量系数的，数据质量好则价值高，数据质量差则价值大打折扣，针对烂数据进行分析、应用，是消耗企业宝贵资源，不仅仅是时间、还有钱财、甚至发展时机！

本文着重探讨一下大数据平台、商业智能系统的数据质量问题。

话在当下

2019年3月10日，一架埃塞俄比亚航空公司的波音737MAX型客机起飞后6分钟坠毁，机上的149名乘客和8名机组人员不幸全部遇难。然而之前的2018年10月29日，印尼狮航一架由波音737MAX执飞的飞机，在从雅加达起飞大约13分钟之后失联、坠毁，机上189人不幸全部罹难。

初步的调查表明：狮航空难是飞机信号系统接收到一个假信号，信号显示飞机“抬头”，所以飞机自动失速控制系统持续给出了“低头”的指令，机组人员与飞机自动失速控制系统搏斗很长时间，但最终还是发生了坠机悲剧。

这个“假信号”及基于这个“假信号”进行“自动失速控制”导致的悲剧，促使我们要进一步重视数据质量！

引发数据质量的原因

有各种各样的因素导致数据质量问题，但归纳起来就两个原因：

1)源系统数据质量本身不可控

2)本系统的ETL程序对异常考虑不足引发数据质量问题

导致数据质量的两个原因归根结底是一个，都是由于程序不完善导致的，但源系统的程序不在控制范围内，只能客观面对数据本身。

笔者对数据质量的几个观点

数据的范畴太大，我想限定在信息系统的范围内说说自己的理解。
由于数据质量，笔者先说一下自己的几个观点，暂不浪费篇幅阐述证明，同意或不同意请自行脑补一下，有机会深入切磋。

1)所有的信息系统，不论国内国外的，不论操作系统、中间件系统、软件框架、平台、组件、应用，不论用哪种语言编写，不论刚面世亦或是诞辰百年，都有数据质量问题，只是数量和危害程度不同而已。

2)所有信息系统的数据质量，归根结底都是由人导致的，更进一步都是由人的能力和敬业精神导致的。

3)垃圾的数据是由具有bug的程序直接制造的。是人在设计、开发过程中的疏忽（姑且认为是疏忽）导致程序经年累月地制造垃圾数据。

- 这样的例子不胜枚举，我们最近就有一个，堪比神剧。
- 您是否听过流传在程序员之间的一个笑话，“你今天写了多少bug?”，“996是拼命写bug！”。您还别真的把这当笑话。

4)数据质量无法根治，只能容忍。

数据质量无法根治，只能容忍

数据质量无法根治的原因首先是没有机会修改程序的无奈，因为很多系统已投产，或许其程序写死在了ROM和芯片中，或许已经没有源程序和文档。

数据质量无法根治的原因其次是由于程序bug和数据质量不可穷尽，尤其是在投产之前的有限时间内。

数据质量无法根治的原因还可能是没有足够的时间和人员成本，或许不想修改引发更多新问题。

说数据质量无法根治，只能容忍绝不是悲观丧气，也不是放任程序员写bug。

说数据质量无法根治，只能容忍真正目的是写高质量程序。因为程序=数据+算法，由此可以通过改进算法从而提高程序的智能性、鲁棒性(鲁棒是Robust的音译，也就是健壮和强壮的意思。它是在异常和危险情况下系统生存的关键。比如说，计算机软件在输入错误、磁盘故障、网络过载或有意攻击情况下，能否不死机、不崩溃，就是该

软件的鲁棒性。)，即使是输入不完美的数据，甚至是非法数据，程序还能正常运行，给出正确的结果。

一个反例，波音737Max8型飞机的自动失速控制系统不容忍攻角传感器输入错误数据，跟驾驶员争夺飞机的控制权，频频发出与驾驶员相悖的控制指令，成功导致飞机高速撞地，惨不忍睹！

还有更多的反例，也是导致严重事故和后果的。

OLAP系统必须容忍输入数据的数据质量

OLAP系统也是信息系统，不论是叫BI系统、数据仓库系统、经营分析系统还是大数据系统，所以也必须容忍输入数据的数据质量。

进一步OLAP系统的数据全部从外部引入，抱怨内部生产系统存在大量bug既不能有效解决输入数据的数据质量问题，也不能成为搪塞领导和业务部门的借口。另外OLAP系统或多或少地需要爬取网络上的所需信息，抱怨数据质量就只是一种无能的表现罢了。

OLAP系统要容忍输入数据的数据质量，在自己可控的范围内，通过技术的手段，让数据质量问题不致命、不致错，力争基于当前数据质量状况给出真实的、客观的业务开展情况，存在的经营问题和改进机会。

OLAP系统要容忍输入数据的数据质量，也应“沧海横流，方显英雄本色。”这句话了！

如何容忍输入数据的质量不佳

OLAP系统要容忍输入数据的数据质量，有一个前提，其他任何系统也是一样，即预先需要知道会存在怎样的数据质量问题。之后才能再针对不同的数据质量问题制定合适的应对机制即容忍。没有针对性容忍的数据质量问题都是无法有效应对，结果也是不可控的。

知道存在多少及怎样的数据质量问题，尤其是影响比较严重的数据质量问题，变得非常关键，这需要靠多年行业经验积累，或者请这样的人帮助，没有捷径可走。

数据质量问题和对应的解决办法

下面举例说明一种数据质量问题和对应的解决办法。在2001年电信运营商建设经营分析系统，即数据仓库系统时，发现一种数据不一致的数据质量问题，常导致业务人员不相信报表数据及在线分析结果，负面的影响比较大。

在关系型数据库中，表的主键与外键关系就是用于保持数据的一致性，并且还有主外键的约束机制，关系型数据库也因这一显著特征而区别于其他类型数据库。但是，不是所有的数据都存储在关系型数据库中，有存文件的，有存储其他种类数据库中的。进一步，存储和使用关系型数据库的，未必所有关系都用主外键约束机制保持着，因此数据一致性的质量问题必然存在，而且存在这种问题最多的就是代码表，尤其是多个单位、多个生产系统数据整合到一起统计分析时。

数据库中的代码表在数据仓库系统中叫做维表，即Dimension Table，存放可枚举数据，用做统计分析的观察角度，比如城市维表，其中存放枚举出可能出现的所有城市标识和城市名称，形成了一个城市名单。

完成上述维表，需要经历如下过程：

- 数据仓库目标维表设计，即dim1表及其中成员的定义；
- 近源系统维表设计，多个系统逐一进行，而且必须建立在近源系统数据调查基础上。
- 上述2个步骤迭代完善至投产。

看起来维表的设计和使用还比较简单，但还有问题：

1)业务系统不仅仅是新增数据，业务及业务系统也在不断的调整，即缓慢变化。相应的dim1、dim1A、dim1B、……是否也在维护？维护及时么？

2)业务系统中出现设计之初不存在的代码，在数据仓库中如何区分及回溯？

解决办法1：自动维表维护

自动维表维护，就是用程序、自动地优先于其他数据，同步源系统的代码表，按照维表定义机制，自动补充之前不曾定义的新代码。

解决办法2：使用自描述维度

自描述维度数据就是不转码存储的数据，举例如：

明细数据直接保存“全球通”、“神州行”、“动感地带”、“男”、“女”、“天蝎座”等内容，而不是其编码，也不再编码，就存储数据的本身，而数据本身更有自描述的好处，从而生命周期更长，也更便于交换，最最重要的是保障了数据的质量。

不会因为原始数据中出现“GoTone”、“GO-Tone”、“Go-Tone”、“神洲行”而导致筛选结果集数量错误，品牌构成的错误！

数据质量保障的新方法

从ODS层出数据实现最严苛的生产报表给了我们足够的启示，数据质量还可以更轻松、更低成本、从根本上解决，思路如下图所示。

- **自然维**

上例存储“全球通”、“男”、“天蝎座”数据的几个列，不重复值的数量很少，是一个维度，只是未抽取成独立的维表并存储维度id而已，这样的列都是一个自然维（欢迎提供更贴切命名的建议）。

之前的文章中说过，像自然维字段，存储数据本身并不比存储维表id或维度名称浪费空间，也不会耗费更多的计算时间。相反，如果处理的好，则空间开销更小、处理速度更快。

- **去预处理和聚合**

数据以自然维方式存储，不用做维度转换，进一步像日期、地址、身份证号码、手机号、多级维度等字段也都不用做解析，即数据不做预处理，只简单、直接地存储。省去很多程序开发工作，又可以更快交付，更低开销运行，更快速运行，同时程序和数据质量更可靠，何乐而不为呢？

预先聚合曾经是非常好的性能优化方法，但如果聚合结果集较大、聚合结果复用度低、存在多种聚合形式，则通过预聚合进行性能改进就被其本身存在的缺点抵消殆尽了。因为预先聚合也要编程，运行也花时间，结果集存储也占空间，多结果间还可能带来一致性问题。内存计算是比预聚合更好的性能优化技术，简单、快速、可靠！

某企业数据质量管控及效果

某企业在建设商业智能系统（BI系统）时，对数据质量提出了很高的要求，同时也投入了大量的人员、时间和资源来改善、提升数据质量。不同分子公司之间经常性探讨、交流成功经验，逐步形成了一套有效的管控方法，还编制了数据质量管控规范贯彻执行，为数据应用保驾护航。

某企业BI系统数据质量管控有很多可取之处，这里只能粗略介绍一二。总体上BI系统是通过人工和系统相结合，分阶段侧重改善数据质量的。

商业智能系统（BI系统）中的基本质量管控程序包括但不限于：

1）**自动维表维护**，自动并优先于其他数据，同步源系统的代码表

2）**数据质量核查**，一个很大、很复杂的系统，让人工配置数据质量核查点、核查方法，系统自动采集数据、比较、告警、报告

3）**元数据管理系统**，在数据质量保障方面主要应用在定位出问题的程序

估计是企业出于对服务SLA的习惯，也出于对数据质量的深刻认识，在数据质量保障上不遗余力，普遍采用上述质量改善及保障流程，**但是每引入新数据时，每上线新应**

用时，都要部分地从第一阶段开始重复数据质量管控过程，存在投入大、周期长的缺点。

作者简介

冯海文，具有20年的数据仓库、BI系统规划、建设经验。设计并主持研发分布式的分析型内存数据库产品和即席查询分析工具。擅长高吞吐率和低延时的高性能计算应用设计及开发。曾经设计、搭建中国移动第一个数据仓库系统，为北京、上海、浙江、山东、四川、辽宁、吉林等16家移动公司的经营分析系统设计数据存储模型、搭建数据仓库，实现ETL，查证解决并保证数据质量，开发可视化的分析应用，建成数据仓库系统。丰富数据模型设计及评估经验。

我们的使命：发展数据治理行业、普及数据治理知识、改变企业数据管理现状、提高企业数据质量、推动企业走进大数据时代。
我们的愿景：打造数据治理专家、数据治理平台、数据治理生态圈。
我们的价值观：凝聚行业力量、打造数据治理全链条平台、改变数据治理生态圈。

了解更多精彩内容

长按，识别二维码，关注我们吧！

数据工匠俱乐部

微信号：zgsjgjjlb

专注数据治理，推动大数据发展。