脏数据潜在的隐患以及数据整合 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/475/2021_2022__E8_84_8F_E6_95_B0_E6_8D_AE_E6_c67_475816.htm 很少有什么IT项目比数据整合更令人头疼的了。如果我们换个方式思考，就会发现有一件事是比数据整合更可怕的，那就是数据整合出现了问题。 有时候，这是由于用户出错或者恶意用户的蓄意破坏，导致不良数据堆积引起的问题。有时候原始数据是完好无损的，但是从一个系统/数据库转移到另一个系统/数据库的过程中丢失、被删截或者被修改了，也会造成麻烦。数据会过时，也会在你企业内部的人事斗争过程中不幸被流弹击中，要知道每个人都是死抱着自己的一小片数据存储地盘，不愿与其他人分享。 有很多的方式会导致数据项目的流产，本文列举了其中五种最常见的情况，告诉你究竟是什么地方出错了，将会导致什么样的后果，以及可以采取什么措施避免同样的情况发生在自己身上。文中所涉及的公司名字一概隐去。希望不要让你自己的经历像本文所叙述的对象那样沦为他人口中的经验教训。 1.“ 亲爱的白痴 ”邮件事件 小心你的数据来源，它有可能会反过来摆你一道。这个事例源于一个大型金融服务机构的客户呼叫中心。就像几乎所有的客服柜台一样，这里的客户服务代表们要做的就是接听电话，并把客户信息输入到一个共享数据库里。 这个特殊的数据库里有一列是用来记录称谓的，并且是可编辑的。但是数据库管理员并没有对这一列的输入规则进行约束，例如只能输入某某先生，某某女士之类的称谓，反而可以接受客服代表输入的任何长达20或30字符的内容。在倾听一些客户愤怒的投诉时

，部分客服代表就会给每条记录添加一些他们自己想出来的不完全友善的注释，例如“这个客户真是个白痴”这类的注释。这种情况持续了很多年，因为机构里的其他系统都不会从这个称谓列中提取数据，所以没有人注意到这一情况。其后某天，市场部决定发起一次直接邮寄活动来推广一项新服务。他们想出了一个绝妙的点子：与其花钱购买一份名单，不如利用客服柜台的数据库。于是，以诸如“亲爱的白痴客户Linlin”这样的措词抬头的邮件开始源源不断的发到客户邮箱里。当然没有任何客户会签约使用这项新服务。该机构直到开始检查他们所发出的邮件时，才弄清楚前因后果。我们拥有的数据不是属于我们自己的。如今世界的联系日趋紧密，很可能会有人找到了你的数据，并把它利用在一个你完全想象不到的地方。如果你从别的地方获取数据，那么在你利用它们执行新任务时，必须要确保你的数据质量管理水平过关了。判断水平“过不过关”，取决于你要如何利用这些数据。正确性是判断数据质量的基本要素之一，对于直邮产业，数据的准确率达到70%至80%就可能就够了。而对于制药业，你就必须达到99%甚至更高。不过，没有什么公司想要或者需要完美的数据，更不用说为了得到完美数据而付出金钱，因为要数据保持完美的代价太昂贵了。问题是要怎样利用数据，以及数据的准确率达到什么程度才足够好。2.死去的人有没有选举权 相信大家对数据清洗（Data cleansing）这个术语并不陌生，它是数据整合过程中必须进行的一个复杂过程，通过检测和清除掉垃圾数据（包括不正确、过时、冗余以及不完整的数据），以保证数据的正确性、可靠性、完整性和一致性。从字面上，我们就可以看出数据清洗是一个“

生死攸关”的问题。下面讲述的也是“生死攸关”的事例。2006年美国国会选举期间，某政府工作志愿者在通过电话让已登记的选民来投票的过程中发现，每十个选民中有三个是已经死去的人，因此没有资格投票。现代社会里死者数据不全所引发的问题很常见，确实也给生者带来了很大的困扰。 对于诸如保险公司、投资公司、基金公司、通讯公司等拥有大量客户的服务类企业而言，客户数据是其重要的财富来源。然而，客户数据质量问题却一直是困扰企业开发新服务项目的绊脚石。在一项关于客户数据质量的调查研究中发现，平均而言，8-15%的客户数据记录存在各种问题，例如各种证件号码输入错误、联系方式过期等等。其中有五分之一的数据问题是由于客户的死亡造成的，其中一部分客户死亡时间超过十年却仍保留着股东的身份。 这并不是客户的疏忽，只是自然发生的问题。私营企业上市、被并购或者拆分，而他们的股东数据却一直被保留着，甚至长达数十年之久。不过这些垃圾数据所引起的问题可能比起在不必要的邮寄费用上浪费一点钱更为严重。最令人担心的问题莫过于欺诈和盗窃ID，如果这些情况发生在颇具影响力的机构组织里，必会导致更为严重的现实问题，例如已故股东的红利被陌生人兑现，继承人的继承权被剥夺，公司机密泄漏等等。 那么要怎么解决这个问题呢？利用商业评测软件可以识别不同系统的异常数据并做好标记方便检查。即便如此，所有的企业都应当加强重视，做好内部监控，严格执行例行的基本检查。事实上，每一个企业都或多或少存在垃圾数据方面的问题。从风险管理的观点来看，最好的解决方案就是持之以恒地检查。如果你从上文的内容能认识到这个自然发生的现象可能会

对你产生什么影响的话，已经有了一个好的开始。 3. 数据重复的代价 用户出错会引发麻烦事，用户自作聪明造成的问题可能更严重。某保险公司从上世纪70年代开始就将大部分客户资料保存在一个主应用软件中，并规定数据录入操作员录入新数据前先要搜索数据库中是否已经有该客户的记录，但是搜索功能执行起来非常慢而且不够准确，所以大多数操作员不再执行这一步骤，而从头开始输入新记录，这样做确实简单轻松多了。然而，结果是很多客户公司的记录在数据库里重复达几百次，使系统运行地更慢，数据搜索结果更加不准确，形成了恶性循环。 不幸的是，这个应用软件已经根深蒂固的嵌入到该公司的其他系统了，管理部门不愿意花钱把它替换掉。最后，该公司的IT部门发现如果公司再也无法查找用户资料了，将会造成的每天75万美元的损失。直到这时候，公司才如梦初醒，使用识别系统来清洗数据，最终清除了近四万条重复记录。 重复数据的问题一直都让IT管理员头痛不已。数据库越庞大，这个问题越严重。但是，很少有人真正认识到问题的严重性。如果有人告诉你他的客户数据库里有2.7%的重复数据，很可能低估了。不过，我们也没有什么灵丹妙药彻底解决这个问题，即使我们能够利用数据匹配技术来沙里淘金，跨越多个数据库找出唯一有用的信息，最难的一关可能是让企业里的不同利益团体就什么数据可以大家共享以及如何构建匹配达成一致。同一个机构里的两个不同的部门可能对匹配和重复项有完全不同的定义。类似的数据整合工作会因为相关人员不能对“ 谁才是数据的所有者 ”以及“ 什么数据可以拿来与别人交换 ”的意见不和而土崩瓦解。 4. 小心老化的数据 相信很多人对魔域大冒险（Zork）这

款最经典的文字冒险游戏还记忆犹新，通过问答形式由游戏设置提供情景描述，而玩家输入选择关键词判断来推动游戏发展，是现代RPG游戏的鼻祖。现在，还有不少人仍在开发这类古老的游戏，这也没什么，问题是他们数据库里保存的用户资料也同样的古老。 某老款游戏开发商利用MailChimp的网络营销服务来联系以前的一万名客户，就是为了提醒他们游戏的第二版终于完成了。他们所用的大部分电子邮件地址至少是十年前的，其中有一部分是Hotmail帐户，很久之前就被遗弃不用了，以致微软已经把这些邮件地址当成垃圾邮件陷阱了。于是，一天之内，所有的MailChimp邮件都被Hotmail的垃圾邮件过滤器列入了黑名单。 幸好游戏开发商以前保留了原始记录，包括每位客户下载其游戏时的IP地址，这成了MailChimp的救命稻草。MailChimp给Hotmail的客服发了紧急申明，证明这些邮箱帐户是合法客户，只是年代比较久远。第二天，hotmail就把MailChimp从黑名单中解救出来了。 所有的数据都会快速老化，就像放射性物质发生衰变一样，而联络数据比其他数据老化得更快。数据库管理人员必须定期更新每一个系统的数据。 美国工商资料库是个巨额产业，而联络资料是所有资料中最受销售人员青睐的，但也是最难维护的。2004年成立于美国的Jigsaw.com是一个在线商务联络资料数据库，面向销售专业人员，采用Wiki式数据清洗方式来维护。该网站的三十多万名用户通过上传新名片资料或纠正错误的名片资料来换取点数，上传的每条记录必须完整，如果上传不正确或是资料太老旧，就会扣除相应的点数。而用户能得到的利益就是用获得的点数购买自己所需要的名片资料。 Jigsaw的首席执行官Jim Fowler称一家科技公司想

要把他们公司的数据库和Jigsaw的数据库进行比较，以便清除不良数据。该科技公司拥有四万条记录，其中只有65%是当前可用的，而且全部数据都不完整。Jigsaw发现他们大部分合作客户都拥有很多毫无价值的数据，根本就没办法去匹配纠正。公司花费了数百万美元在客户关系管理软件上，可见这些数据有多糟糕。有时候公司的真正价值不在拥有的数据本身，而在于有没有能力与时俱进地跟上数据变化的速度。Jigsaw的能力正是在于完善数据并进行自我清洗，如果没有自我修正的机制，Jigsaw也只不过是一家毫无价值的数据公司而已。 5. 小错误与大麻烦 好数据和不良数据之间的差别很可能就体现在一个小点上。某专案优化解决方案供应商的高级顾问告诉我们，他曾为一个大型数据整合项目做顾问，这个项目看起来一切都运行正常，但六个月后，某人打开一个数据表，只看到了一排排符号，什么数据都没有。 这其实只是一个字符代码错误：本来在一些域里应该用省略号（三个点）的，但有人只输入了两个点，导致了整个数据线的崩溃。该公司不得不费尽力气从备份中重新创建整个数据库：查找省略号，然后用正确数据替换。 很多时候，问题不仅仅是简单的数据录入错误或者是“ 脏数据进脏数据出 ”的问题而已。很多企业在进行不同操作系统之间的数据移植或从老的SQL版本中升级数据等操作时并没有做好充分计划。他们总是希望利用手头上任何可利用资源火速进行，而把数据清洗任务冀望于以后完成。更甚者，他们的测试环境和操作环境可能并不一致，或者他们只用少量数据子集来测试，没有测试过的数据很可能会在后面的操作引发大麻烦。 企业经历着深刻的技术革命，却没有在数据整合和维护的管理上花费

足够的时间和精力，最终只会成为不良数据的牺牲品。在数据迁移的过程中，有无数的机会让它们成为不良数据。 不要指望IT部门来验证你的数据。让与这些数据密切相关的有能力的用户来帮助你做好数据整合计划和测试。在你决定进行整合之前，先查看一下所有数据，确定用于从中提取数据的应用软件。如果可以，最好测试所有的数据而不是其中某个子集，要知道正如上面的例子所示，就算是一个小的不能再小的错误都会把你和你的数据拉进痛苦的深渊。 我们最后再用一个实例来说明小错误和大麻烦之间的关系。 某商业风险管理解决方案供应商的某位客户创建了一个SQL服务器数据库，用来确定是否有错误的CAD文件在其网络内部流窜。原本的设想是，如果错误的数据包超过某设定阈值，公司管理员就会知道并进行数据挖掘和清洗工作。问题是他们不小心颠倒了数据库的规则设置（把两个阈值放反了），导致错误数据包越多，提交公司的报告里显示的网络运行情况就越好。最后该公司网络被某种蠕虫病毒入侵，破坏了他们的工程CAD档案。他们不得不重头开始花费大量的金钱来重建大部分的文档。这一切都是因为一个非常简单数据提取设置错误造成的。 希望本文讲述的内容能够让大家对数据整合有个正确的认识，数据整合不可规避，并且要谨慎行事。 100Test下载频道开通，各类考试题目直接下载。详细请访问 www.100test.com