**数据质量与数据质量八个维度指标**

全文摘自《企业经营数据分析》赵兴峰著，转载已获得作者认可。

数据的质量直接影响着数据的价值，并且直接影响着数据分析的结果以及我们以此做出的决策的质量。质量不高的数据不仅仅是数据本身的问题，还会影响着企业经营管理决策；错误的数据还不如没有数据，因为没有数据时，我们还会基于经验和基于常识的判断来做出不见得是错误的决策，而错误的数据会引导我们做出错误的决策。因此数据质量是企业经营管理数据治理的关键所在。

数据的质量可以从八个方面进行衡量，每个维度都从一个侧面来反映数据的品相。八个维度分别是：准确性、真实性、完整性、全面性、及时性、即时性、精确性和关联性。



我们在比较两个数据集的品相的时候往往采用这种图形表示。比如说，常规来讲内部数据采集的准确性、真实性、完整性高，而全面性、及时性、即时性、精确性和关联性方面取决于企业内部对数据的重视程度以及采用的技术手段的先进性有关；外部数据集，比如说微博数据、互联网媒体数据等，其全面性、及时性和即时性都可以通过技术手段，如网络爬虫等得到提高，但在准确性、真实性、精确性上难以保证，也难以控制，在关联性方面取决于数据采集和挖掘的相关技术。

我们也可以用这个模型来衡量公司内部各个职能部门数据的品相。下图是个示意，通过数据质量8大指标的评价，我们可以对企业内部数据治理有针对性地采取措施去提高企业的数据质量。

**数据的准确性**

数据的准确性（Accuracy）是指数据采集值或者观测值和真实值之间的接近程度，也叫做误差值，误差越大，准确度越低。数据的准确性由数据的采集方法决定的。

* **数据的精确性**

数据的精确性（Precision）是指对同一对象的观测数据在重复测量时所得到不同数据间的接近程度。精确性，也可以叫精准性。精确性与我们数据采集的精度有关系。精度高，要求数据采集的粒度越细，误差的容忍程度越低。

测量人的身高，我们可以精确到厘米，多次测量差异只会在厘米级别；测量北京到上海的距离，我们精确到公里，多次测量结果间的差异会在公里级别；采用游标卡尺测量一个零件的厚度，可以精确到1/50毫米，多次测量的结果间的误差也只会在1/50毫米间。采用的测量方法和手段直接影响着数据的精确性。

* **数据的真实性**

数据的真实性，也叫数据的正确性（Rightness）。数据的正确性取决于数据采集过程的可控程度，可控程度高，可追溯情况好，数据的真实性容易得到保障，而可控程度低或者无法追溯，数据造假后无法追溯，则真实性难以保证。

为了提高数据的真实性，采用无人进行过程干涉的智能终端直接采集数据，能够更好地保证所采集数据的真实性，减少人为干预，减少数据造假，从而让数据更加正确地反应客观事物。

* **数据的及时性**

数据的及时性（In-time）就是数据能否在需要的时候得到保证。我们月初会对上个月的经营和管理数据进行统计汇总，这些数据能否及时处理完成，财务能否在月度关账后及时核算。数据的及时性是我们数据分析和挖掘及时性的保障。如果公司的财务核算复杂，核算速度缓慢，上个月的数据在月中才能统计汇总完成，等需要调整财务策略的时候，已经到了月底了，一个月已经快过完了。特别是公司做大了之后，业务覆盖多个市场、多个国家，数据不能及时汇总，会影响到高层决策的及时程度。

数据的及时性与企业数据处理的速度和效率有直接的关系，为了提高数据的及时性，越来越多的公司采用管理信息系统，并在管理信息系统中附加各种自动数据处理功能，能够在数据上传系统之后自动完成绝大部分报表，从而保证数据处理的效率。计算机自动处理中间层数据是提高企业数据处理效率的有效手段。

除了保证数据采集的及时性和数据处理的效率问题外，还需要从制度和流程上保证数据传输的及时性。数据报表完成了，要及时或者在要求的时间范围内发送到指定的部门，或者上传到指定的存储空间。

* **数据的即时性**

数据的即时性是指数据采集时间节点和数据传输的时间节点，一个数据在数据源头采集后立即存储，并立即加工呈现，就是即时数据，而经过一段时间之后再传输到信息系统中，则数据即时性就稍差。

微博的数据采集，当用户发布了微博，数据立即能够被抓取和加工，会生成即时微博数据报告，并随着时间推移，数据不断变化，我们可以称作是即时采集和处理的。一个生产设备的仪表即时反应着设备的温度、电压、电流、气压等数据，这些数据生成数据流，随时监控设备的运行状况，这个数据可以看作是即时数据。而当设备的即时运行数据存储下来，用来分析设备运行状况与设备寿命的关系，这些数据就成为历史数据。

* **数据的完整性**

数据的完整性是从数据采集到的程度来衡量的，是应采集和实际采集到数据之间的比例。一条信息采集12个数据点，如我们采集员工信息数据的时候，要求填写姓名、出生日期、性别、民族、籍贯、身高、血型、婚姻状况、最高学历、最高学历专业、最高学历毕业院校、最高学历毕业时间等12项信息，而某一员工仅仅填写了部分信息，如只填写了其中的5项，则该员工所填写数据的完整性只有一半。

一个公司数据的完整性体现着这个公司对数据的重视程度。要求采集数据而实际上并未完整采集，只采集了一部分，这就是不完整的，往往是公司对数据采集质量要求不到位导致的。公司要求每个人都填写完整的个人信息表，而有部分员工拒绝填写，公司2000员工，只有1200人填写了完整的个人信息表，则这个数据集就是不完整的。

另外，对于动态数据，我们可以从时间轴上去衡量数据采集的完整性。比如，我们要求每小时采集一次数据，每天会形成24个数据点，记录为24条数据，但是员工渎职，只记录了20次，那么这个数据集也是不完整的。

* **数据的全面性**

数据的全面性和完整性不同，完整性衡量的是应采集和实际采集的差异。而全面性指的是数据采集点的遗漏情况。比如说，我们要采集员工行为数据，我们只采集了员工上班打卡和下班打卡的数据，上班时间的员工行为数据并未采集，或者没有找到合适的方法来采集。那么，这个数据集就是不全面的。

我们描述一个产品的包装，仅仅描述了产品包装的正面和背面，没有记录产品包装的侧面，则就是不全面的。我们记录一个客户的交易数据，我们只采集了客户订单中的产品、订单中产品的价格和数量，而没有采集客户送货地址、采购时间，这个数据采集就是不全面的。

腾讯QQ和微信的用户数据记录了客户交流沟通的数据；阿里和京东的用户数据记录了用户的购买交易数据；百度地图记录了用户出行的数据；大众点评和美团记录了客户餐饮娱乐的数据。对于全面描述一个人的生活的衣食住行各方面，这些公司的数据都是不全面的，而如果把他们的数据整合起来，则会形成更加全面的数据。所以说，数据的全面性说一个相对的概念。过度追求数据的全面性说不现实的。

* **数据的关联性**

数据的关联性是指各个数据集之间的关联关系。比如员工工资数据和员工绩效考核数据是通过员工这个资源关联在一起来的，而且绩效数据直接关系到工资的多少。采购订单数据与生产订单数据之间通过物料的追溯机制进行关联，而生产订单又是由员工完成的，即通过员工作业数据与员工信息数据关联起来。

其实，我们本书探讨的企业大数据，每个数据集都是相关关联的，有些是直接关联的，比如员工工资数据和员工绩效数据，有些是间接关联的，比如说物料采购订单数据与员工工资数据。这些数据的关联关系是由公司的资源，包括人、财、物和信息等，连接起来的。如果有任何的数据集不能连接到其他的数据集，就会存在数据割裂或者数据孤岛。数据割裂和数据孤岛是企业数据关联性不足导致的。而数据的关联性直接影响到企业数据集的价值。