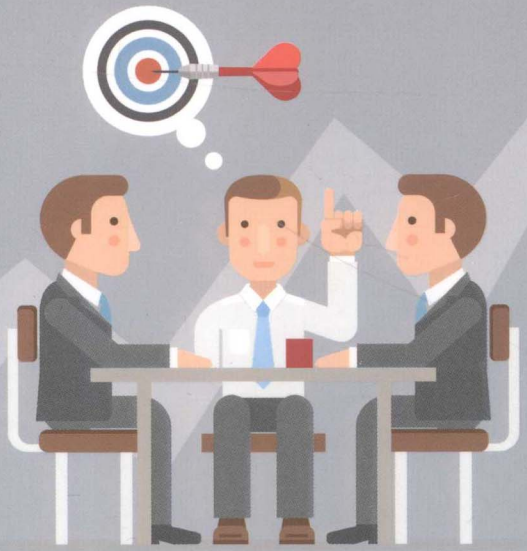
Broadview
www.broadview.com.cn

大数据丛书



用数据治理企业、改变企业！

企业数据化管理变革

数据治理与统筹方案

赵兴峰 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

第 1 篇

数据源头篇

1

你了解企业中的大数据吗

什么是企业大数据

企业大数据从哪里来

数据将成为像石油一样宝贵的资源

数据系统是企业的神经系统

企业大数据有什么用

大数据成了热门的词汇，从 2012 年开始到现在，其实这还仅仅是开始，未来 10 年绝对是大数据技术普及应用的最好时机。这与 10 年前比较火的“商业智能（BI）”不同，大数据的应用真的在发挥作用，影响我们的生活，甚至惊动了各国的国家领导人。奥巴马在 2012 年成立了大数据研究与发展局（Big Data Research and Development Institute）来研究大数据如何更好地推进政府治理工作；2015 年 9 月 6 日，李克强总理签发了《促进大数据发展行动纲要》；各国政府也在不断地构建数据开放平台。

而大数据是什么？与我们企业有什么关系？很多人还存有疑问，或者概念模糊。本章会探讨什么是大数据，什么是企业大数据，普通的企业与大数据有什么关系，以及为什么说企业大数据是企业宝贵的资源等。

1.1　什么是企业大数据

1.1.1　大数据的概念

首先我们先来了解一下大家常说的大数据是什么。

大家常提到的大数据，一般来讲是指企业外部的大数据。随着智能终端设备的普及、互联网技术的升级、移动互联网的快速应用以及数据存储、数据处理和数据分析与挖掘技术的革新，我们身边的各种数据都以“数字化”的形式被记录下来，从而产生了大量的数据记录，这个数据量级之大，超乎一般人的想象，因此就有了大数据这个说法。所以，一般意义上的大数据是指，数据量级非常大，以致我们常规的数据处理、数据存储以及数据分析能力无法满足要求，因而我们称其为大数据。

数据的处理能力是相对的，也是在不断发展和变化的。技术每天都在进步，经过一段时间之后回溯去看，我们会惊讶于其发展速度之快。随着技术的快速进步，包括

数据记录技术、数据存储技术、数据传输技术、数据分析技术以及数据挖掘技术等的发展，我们之前无法处理的数据量级，现在来看就会觉得非常小，甚至可以用微小来形容了。

20 年前，我们还在使用 286、386、486、586 的 PC 机器，100MB 对于我们来讲就是一个天文数字，而现在电脑的存储容量都是用 GB 来衡量的，一台普通的笔记本电脑都有 500GB 以上的存储容量，甚至有些智能手机都有超过 100GB 的存储容量。10 年前我们处理 1MB 的数据，计算机需要运行很长一段时间，而现在大型电商像淘宝、京东都已经进入了上百 PB 级别，百度的数据量更是接近 EB。大多数部署了管理信息系统的企业，数据量级都在 TB 以上级别。而亚马逊的 AWS 云服务器超过 300 万台，在全球共分布有几十个数据中心，这些在 20 年前都是无法想象的。所以说数据处理能力是一个相对的概念，其依然在高速发展，大数据的概念也会不断地演变，今天的“大数据”在不远的将来可能会被看作是“微数据”。

大数据的应用越来越普及，我们常常听到的应用大数据的企业多是互联网企业、电信企业、电商以及金融服务企业，这些企业所在的行业本身就是“富数据”行业，企业自身的经营特征决定了能够存留大量的数据。比如，百度的主营业务就是通过后台的数据搜索服务器来收集互联网数据供用户查询；在 UGC（User Generated Content，用户产生内容）的时代，各种社交媒体，包括微博、微信、QQ 等，存留了大量的用户活动数据；电信运营商本身就在为客户提供各种数据传输服务，因而能够存留大量的客户沟通和传输的数据；亚马逊、淘宝、京东等电商企业，本身的客户数量庞大，加上所销售的产品种类繁多，也存留大量的交易活动记录数据；金融服务企业，例如银行，为巨量的用户提供资金的转移服务，拥有大量的交易记录信息数据。

这些都是我们常说的大数据，这些数据当中，有些是开放的，可以通过技术手段来获取和使用。比如，我们可以使用程序爬取微博数据，来分析微博用户的行为

和其对企业、品牌或者某些产品的看法；可以通过搜索引擎来提高企业品牌的曝光率或者被网络用户搜索到的概率；通过爬取电商平台上的信息来掌控产品的销量及价格的走势。有些富数据的企业，也在利用其所拥有的数据为自己和客户提供数据分析和挖掘的服务，甚至有的企业将数据作为自己的产品或者服务，销售给需要数据的企业。

现在所说的大数据几乎无处不在，可以是任何事情的记录，也包括任何智能数字化终端的数据记录。仅北京市，每天各种视频监控可以产生大概 0.6PB 的视频记录数据。而北京市的 2000 万市民拥有的近 1000 多万部智能手机的 GPS 产生的数据可达到上百 GB，如果包括智能手机中的微信、微博、QQ 等各种社交软件所产生的数据，则可以达到上百 TB。这些数据都是大数据的组成部分。现在，我们会利用清明上河图来了解那时的社会情景，而未来几百年之后，我们的后代在研究现在社会历史的时候，会利用更多的图文史料来研究这个时代。

1.1.2　企业大数据的概念

以上所提到的这些大数据，对产生数据的平台本身来讲是内部的大数据，但对多数企业来讲，这些是外部大数据。现在大多数人口中所说的“利用大数据来做某某事”，基本上指的是利用外部大数据。

利用外部大数据的案例有很多，一般都需要专业的数据人员，并需要投资足够的设备、网络带宽来实现对外部大数据的获取。对于大多数非“富数据”行业[1]的中小企业来讲，利用外部大数据还是比较难的。同时，探讨外部大数据应用的书已经非常丰富了，也有很有趣的案例可供大家参考，但本书将不再重复这些例子，本书

1　“富数据”行业是指行业公司本身的经营特征决定了该行业内的企业会有大量的数据，包括互联网行业中的搜索引擎、社交、媒体等，电商行业，电信行业、金融、零售、连锁服务、生产制造、贸易、物流等。这些行业因为自身经营模式或者业务特征，有大量的客户群、服务内容、巨量交易，就会沉淀大量的数据，从而称为“富”数据行业。

将从与每个企业都非常相关的内部大数据的视角来看企业的大数据治理和应用。

每个企业在日常经营和管理中都在产生数据。员工上下班打卡、销售人员销售产品、客户经理同客户通电话、生产线上在生产产品、公司财务在收款和付款、采购人员在同供应商询价及人力资源的员工在进行着招聘、面试、培训、考核、发工资等活动，这些都是企业经营管理的日常活动，只要企业还存续，这些活动就会持续不断地发生着，如果这些活动被记录下来，就形成了企业的内部数据。有些公司会比较重视数据的记录，有些公司并没有把这些活动记录下来留存成数据。大多数企业会对财务、人员工资、销售、采购等经济往来有相对明确的数据记录和管理。有些公司的数据量级非常小，有些公司则非常庞大，区别在于公司的规模、经营模式和业务内容。

我们把以上这些数据叫作企业大数据，给一个明确的定义就是：**企业大数据是指全面记录企业经营和管理活动的数据**。这个定义是从企业实践应用的角度出发的，不过多地强调数据量级的大小，即使是一个非常微小的数据，也是企业大数据中的一部分。该定义更强调数据涉及范围的全面性。在企业经营和管理过程中，单独的数据或者孤立的数据价值会大打折扣。只有全面记录数据和信息并实现相互间的关联，才能够使其更好地发挥作用。

如果充分且有效地记录公司人、财、物各种资源以及资源的活动，形成数据库，并长期坚持采集记录，那么这个数据的量级对中等规模以上的企业都不是小数据。几百人的企业，规模虽然小，每个员工的活动和每个客户交易的活动、每次市场调研、每次产品推广等，如果能够详细记录，形成完整的数据库，经过几年，即使不计算图片、音频、视频等多媒体数据，这个数据量级也可以达到 TB 级别。

1.1.3 数据的价值密度概念

数据的价值在于挖掘，但数据本身对于不同的对象，也有不同的价值。作为外部大数据的微博数据信息量非常大，因为微博来自于千千万万兴趣不同的用户，记

载着不同的内容，表达着对各种事物的看法和想法，这些内容因为不够聚焦，所以对单个企业来讲，其价值含量就非常低。但因为数据量级的巨大，可以通过在上亿条记录中找出部分与企业业务相关的信息，就能够帮助企业了解客户需求、了解客户对产品或者竞争对手产品的评价，从而帮助企业随时了解外部动向，这是有意义的，只是数据的价值密度低而已。

而企业大数据则不同，每一条信息记录都是与企业相关的，每一条信息都可能蕴含着巨大的信息量。所以说，其数据的价值密度就很高。一个公司月度销售额数据一年 12 个月的数据才 12 条，可这 12 条数据能够反映企业每个月的销售额变化以及企业环比增长情况；加上每个月的销售目标情况才 24 条数据，但能够反映出这个企业每个月完成销售目标的情况；如果把几年的月度数据叠加对比，就会反映出这个企业所在行业的季节性变化情况。所以，微量的数据可能蕴含着大量的信息。这些高价值密度的内部数据，需要企业更加重视起来。

1.1.4　开始积累企业大数据

很多企业在谈大数据时，艳羡外部大数据的量级，以及部分企业从对大数据应用中所获得的利益与价值，却并未重视内部经营和管理活动的数据采集。很多有价值的数据并未在历史的过程中记录下来，甚至有些上规模的企业仍然舍不得在管理信息系统上进行投资，主要的原因还是没有充分认识到这些数据的价值，也不知道这些数据有什么用。

受实用主义理念的影响，当企业的管理者看不到数据的价值的时候，就不会注重对数据的收集和管理，因而很多企业在发展过程中，并没有将上面谈到的各种数据记录在一起，这就有了部分企业觉得自己的企业中没有数据这样的想法。其实企业不是没有数据，而是没有记录、整理，或者说没有对数据进行管理。

我们不可能分析和挖掘没有的数据。如果我们现在不记录下企业经营管理活动

所产生的数据，以后肯定无法再找到这样的数据，靠回忆是无法将数据记录得完整、全面和准确的。没有数据就无从分析，也就无从挖掘数据的价值，而挖掘不到数据价值的时候，就更不会去注重数据的收集和管理，这就演变为一个“先有鸡还是先有蛋”的问题争论。

未来的市场竞争环境和过去已经完全不同，依靠经验做出的判断往往是有非常高的风险的，没有数据的企业就像没有昨天、没有历史一样，无法“以史为鉴”，曾经缴纳的“学费”还要继续去缴，甚至还会犯同样的错误，走同样的弯路。现在的市场竞争环境越来越复杂，瞬息万变，企业如果没有历史数据，就无法做到心中有数，而“心中有数”这句古语本身就在强调数据的重要性。

企业最大的经营风险来自于外部和内部环境的不确定性，越是在复杂多变的市场环境下，企业要想持续经营就越加需要注重确定性，而提高企业经营和管理确定性的基础就是数据。大多数企业的灭亡都是因为管理决策失误造成的，而管理决策的准确性依靠对内外部环境准确地判断，如果我们能够有明确的数据，判断的准确性就能得到大幅度的提高，决策失误概率就会大大降低，企业持续时间就会更长久。由于数据化管理或者是说数据思维在发达国家企业的普遍性，其企业平均持续时间就会更长久些，根据财富杂志的研究表明，美国企业的平均寿命在 7 年左右，而中国企业的平均寿命不到 3 年。

一件事情做不成有两个原因，一个是“不会”，另一个是“不为”，“不会”可以通过学习来解决，而“不为”则需要转换理念，改变习惯。企业的数据化管理也需要从“不会”和“不为”两个方面去诊断。

企业大数据的概念还非常新，相关的知识也比较匮乏，市面上能买到的书也比较少，管理学院的课程也待开发，“不会”的问题肯定是存在的，而制约企业数据化管理方式推进的更大阻力则来自于“不为”。“你不可能叫醒一个装睡的人”，数据化管理方面也一样，你不可能教会一家不愿意推进数据化管理的企业；将视角放到

企业内部也一样，企业的大数据积累和沉淀都需要企业全员的数据思维和数据意识，如果中层管理者和基层员工都没有数据意识和数据思维，企业高层也无法推动。

1.2　企业大数据从哪里来

随着大数据概念的火爆和普及，每个人都逐步意识到大数据的重要作用，都开始思考企业大数据的问题。有些公司的高层开始请外部的大数据专家来讲课，希望内部的员工能够开始使用大数据。而中层的管理者总是一头雾水：“大数据在哪儿呢？”

1.2.1　企业大数据来自我们的日常工作活动

其实每一位管理者仔细思考一下自己日常的工作，就会发觉自己日常接触到的内部数据其实有很多。在这里简单罗列一下，以下这份清单几乎是所有的企业都应该有的，即使不保存在公司电脑里或者是说存储在企业管理信息系统里，各个岗位的管理者也应有一份自己的数据清单，以方便自己的工作。

部　门	数 据 表
人力资源管理	• 员工花名册 • 员工基本信息表 • 员工工资表 • 员工考勤表 • 员工绩效指标数据表（KPI） • 员工绩效考核表 • 员工岗位说明书 • 内部员工通讯录 • 组织架构图与岗位人员花名册

续表

部　门	数 据 表
财务管理	• 收款记录流水单 • 付款记录流水单 • 银行对账单 • 报销记录流水单 • 固定资产清单 • 固定资产信息表
销售管理	• 客户名录 • 销售订单记录表 • 产品和服务清单、价格表 • 客户服务记录表 • 竞争产品名录 • 竞争对手活动记录 • 潜在客户名单

以上只是从三个部门的角度出发列出了一些基本的数据表，这些基本数据表的完整程度、管理的规范程度直接反映了企业基础数据管理的完善程度和规范程度，这些数据表中的数据质量也会直接体现出这个企业所拥有的内部数据的质量。因此，在判断企业目前数据化管理程度时，笔者一般会直接让企业相关部门提供以上清单中的几个数据表，就能快速做出相对准确的判断。

企业的每个岗位、每个人员都在进行着与企业相关的经营和管理活动，都在掌握着企业相关资源，拥有这些资源的信息和记录，这些资源与资源转换活动就是企业大数据的发源地。只要每个岗位的员工都能参与到数据采集和数据记录的过程中，或者配合着相关的设备完成对数据的采集工作，企业积累自己的大数据就是一件非常容易的事情。

1.2.2　企业数据源头管理需要系统化

从前面这份数据表清单示例中可以看到：有的数据是基本的信息表，有的数据

是活动的记录表，会形成一个流水清单；有的数据是主动记录下来的信息，有些数据是机器自动采集完成的；有的不是公司内部的资源，但是需主动采集的信息。可以说，数据源头是各种各样的，有的信息比较容易管理，比如说公司安装了门禁和指纹考勤机，要求每个员工上下班打卡，就能够自动记录考勤情况。

而有的信息，比如竞争产品信息数据、竞争对手活动数据、潜在客户名单等相关的数据表，就需要销售部门的人员主动去外部采集，数据的质量和数量都与销售人员的积极主动性直接相关。员工自己比较主动、勤快，或者说有数据意识，就会去收集整理这些数据，如果公司不要求，基本很少有人去做，即使要求了，应付差事的情况也很多见。

企业大数据管理不能依赖于个人的积极性和主动性，因为不同的员工会带来不同的结果。要想构建比较完善的企业大数据，就需要系统化地管理。为保障源头数据的质量，企业需要明确什么源头需要什么样的记录，在数据信息字段的采集、数据的格式、数据记录的载体、数据的存储和传输形式等方面形成规范性的要求，并对相关源头数据的负责人提供足够的培训，在过程中进行监督检查。

比如，最基本的《员工个人基本信息登记表》是基础数据表，人力资源部对该表所采集数据的质量，包括数据的全面性、准确性、及时性和完整性负有管理责任。人力资源部在入职管理或招聘岗位相关人员时，需在人员招聘面试、入职等时间节点上对该数据进行采集，让每个新员工填写完整的《员工个人基本信息登记表》，并在日常工作中，随着员工个人情况的异动，定期进行更新。比如说，每个季度需要员工填写个人信息异动表；在某些管理工作节点发生异动后，及时更新信息库，如员工请婚假，需要及时更新员工的婚姻状况、家庭成员状况的信息；员工请产假，需要及时更新员工的子女状况信息；为员工开具个人收入证明，其买房时，需要更新员工个人资产、个人居住地址等相关的信息。一方面，需要数据负责人对自己所负责的数据有质量意识；另一方面，在内部管理上，需要建立并不断完善这种活动

与数据更新的联动机制。这需要在内部管理制度、岗位说明、任务说明、流程要求等方面做出数据管理的规范性要求。

系统化的数据管理制度与流程能够保障企业大数据的质量：在数据采集的全面性上、在数据的完善程度上、在数据的准确性上、在数据采集的及时性上、在数据积累的持续性都要有保障；同时，系统化的管理能够将以上各种数据关联到一起，形成高度关联的大数据集合。

1.2.3 企业大数据的分类

本章探讨的企业大数据会将重点放到企业内部大数据上，这里的“内部”更多的是从数据拥有方式上定义的，指企业所能够自主拥有的大数据，具有“自主产权”的数据，包括企业主动采集或者采购的外部数据。

从数据所描述的“主体”上，我们把企业大数据分成两个大类，一类是资源信息数据，另一类是资源活动记录数据。

第一类，资源信息数据。资源信息数据是静态数据，记录企业相关内外部资源主体的相关信息，企业的资源包括人、财、物和信息四大类资源，其中的信息资源包括企业的无形资产、技术专利、经营诀窍、客户关系以及内部的数据等资源。

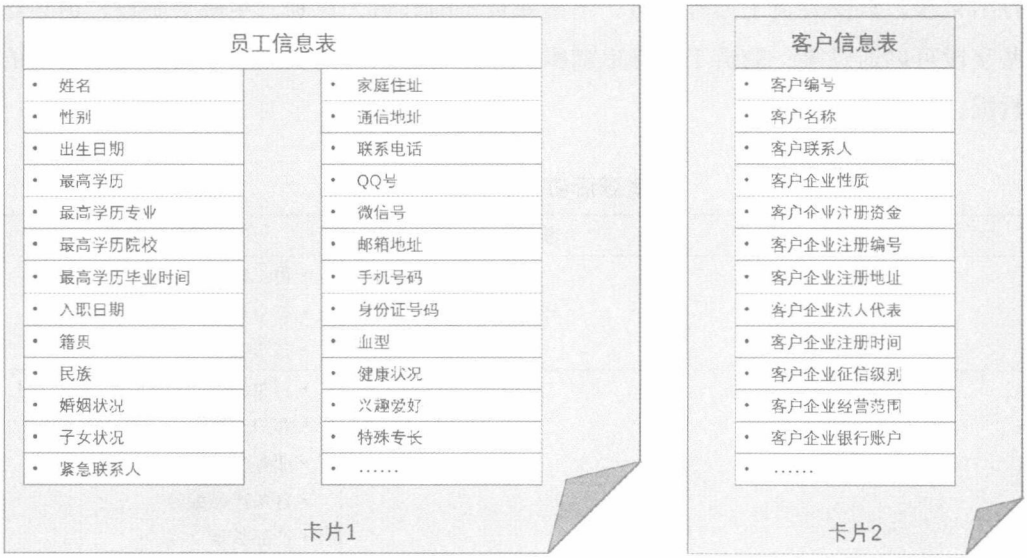
比如，人这个资源，指所有与企业经营活动相关的人，包括公司的领导者、管理者、员工，还包括与公司经营有利益关系的人，例如客户、供应商、竞争对手、政府、社区、协会等。

资源信息类的数据相对于资源活动记录数据来讲，具有相对的稳定性，对即时性要求相对较低。比如，对人这个资源的描述信息相对是固定或者稳定的，但内部员工会随着岗位变迁、人员流失、招聘等活动而发生变化，但人的基本信息变动频

率不像资源活动记录那样有着非常高的时间节点性，对记录的即时性要求不高，即使事后补充记录，对数据质量的影响也不会太大。

对资源信息的记录，比较强调信息记录的全面性。但受限于法律规定、信息获取手段等，数据的完整性不见得都能得到保证。比如，收集内部员工的个人信息受隐私法保护的限制，有些信息比较敏感，可能无法强制获取；对客户信息的收集，受客户提供信息的意愿和采集数据的手段限制，对客户信息的采集往往难以保证完整性。这里就需要把握一个度，通过长期的坚持和积累，实现数据的不断丰富。

对资源信息类数据源进行系统性梳理时，常常会采用一些卡片工具进行采集或者诊断现有数据信息的完整性，如下图所示。



信息字段定义卡片工具

第二类，资源活动记录数据是指公司经营和管理活动所必然牵动的数据。比如，员工的考勤数据，跟客户进行的买卖交易活动，这些都是资源的活动，具有非常敏感的时效性，所以可以称之为“动态数据”。根据笔者在实践中的观察，企业对活动

的记录往往是比较缺乏的，容易发生“事情做了，但没记录下来”这样的情况。为了更好地保留企业内部各种经营管理活动所带来的资源活动数据，需要建立严格的管理流程和制度，并配以足够的技术手段，实现活动记录的即时记录。在宝洁公司，为了追求数据的即时记录，内部流行一句话：“没有记录下来的事情都没有发生过。”就是说，如果不记录下来形成数据，你的工作相当于没有做。

在对动态数据进行梳理的时候，笔者经常采用的是表格工具，如下表所示。表格的左边是梳理企业所有相关资源的企业资源列表，右边是资源对应的活动，这样就将活动对应应该记录的内容进行了明确化。因为不同的公司有不同的业务特征，信息记录字段的要求也不同，此处仅仅作为示例。从数据结构的角度讲，注意不要有太多重复的记录，这样会加大以后进行数据校验时的工作量。比如，员工上下班打卡记录，只要有员工编号即可，不需要员工的姓名、性别、年龄等字段，因为这些字段可以通过唯一的员工编号追溯得到，这个对应的是员工基本信息数据表中的数据。

资源活动记录字段定义工具表

资源名称	资源活动	记录信息字段
人-员工-销售员工	上下班打卡	• 员工编号 • 打卡日期 • 打卡时间
	客户订单	• 订单编号 • 客户编号 • 业务员编号 • 订单产品编号 • 产品数量 • 产品价格 • 交货时间 • 交货地点 • 交货方式 • 验货方式 • 付款方式

续表

资源名称	资源活动	记录信息字段
	客户订单	• 付款日期 • 品质规格要求（编码） • 服务条款 • …

之所以要把数据分成静态数据和动态数据，主要是为了企业能够系统化地梳理数据源头，解决“数据从哪里来”和数据记录全面性的问题。即在对企业大数据进行系统性地梳理时，首先要梳理企业所有的相关资源，然后再对资源的活动进行梳理，这样就能够全面地、系统地梳理企业所有的大数据，然后再根据技术条件（可获取性）、经济条件（成本投入高低）和数据本身价值进行分类，将最紧迫、最重要、高价值密度的数据优先获得，并逐步纳入数据库中，从而构成企业的大数据源头。

1.2.4　企业大数据的六大主要来源

为了更加全面地梳理或者评测企业大数据的源头，需要从企业经营活动主体边界角度再进一步看企业大数据的来源，从而为企业构筑更加完整和全面的数据源头提供思路。

从数据描述对象与企业的关系角度以及动态和静态信息来分类，企业大数据的来源主要有六大类（如下图）：

（1）企业资源的信息数据（静态数据）；

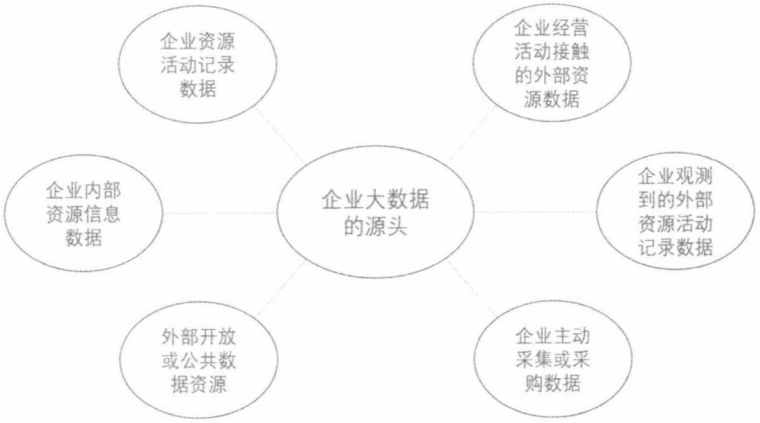
（2）企业资源活动的记录数据（动态数据）；

（3）企业经营活动所接触外部资源的信息数据（静态数据）；

（4）企业观测到相关资源活动的记录数据（动态数据）；

（5）企业主动采集或者采购的外部数据（静态 + 动态数据）；

（6）外部开放数据和公共数据资源（静态 + 动态数据）。



企业大数据的六个主要来源

以上分类中，通过 1.2.3 节中梳理数据源头的方法基本可以梳理清楚，这里重点介绍一下第 5 类和第 6 类。

企业主动采集或者采购的外部数据是企业根据经营决策需要，采用数据采集的手段和方法，成立数据采集项目，完成数据采集的工作。比如，公司为了了解市场中消费者的分类，为公司选择目标客户群体，并定位关键细分客户群体重点研发新产品时，可以发起消费者研究活动，通过市场研究项目，定性或者定量研究消费者的需求，然后形成数据分析报告。这样采集的数据就是企业跨出自己的经营边界所能够接触到的资源，属于主动采集数据。

如果企业能够坚持每年做一次市场调查，经过几年的跟踪、监控，就能掌握消费者对产品需求的变化线路，从而敏锐地感知到消费者需求的变迁，及时根据消费者需求的变化调整自己的产品线和品牌路线，让产品能够更好地满足消费者的需求，保证最佳的客户体验，公司就能在市场上一直保持较好的竞争优势。国内的企业能够坚持这样做的不多，大多数是跟随企业领导做出产品线的调整，或者看到市场上哪一类产品开始受欢迎就跟进模仿，而不是自己花费人力和物力去研究、创新。宝