第8章　外部数据/非结构化数据与数据仓库

大部分组织是以现有系统为来源的数据（即企业的内部数据）上建立其第一个数据仓库。在绝大部分情况下，从现有系统抽取的数据可称为内部结构化数据。数据来自于企业内部，并且数据已经被变换成一种规则的格式。

但是，企业合法使用的其他大量数据却并非产生于企业本身的系统。这类数据称作外部数据，通常这些数据是以非结构化的、不可预测的格式进入企业的。图8 -1表示了进入数据仓库的外部与非结构化数据。

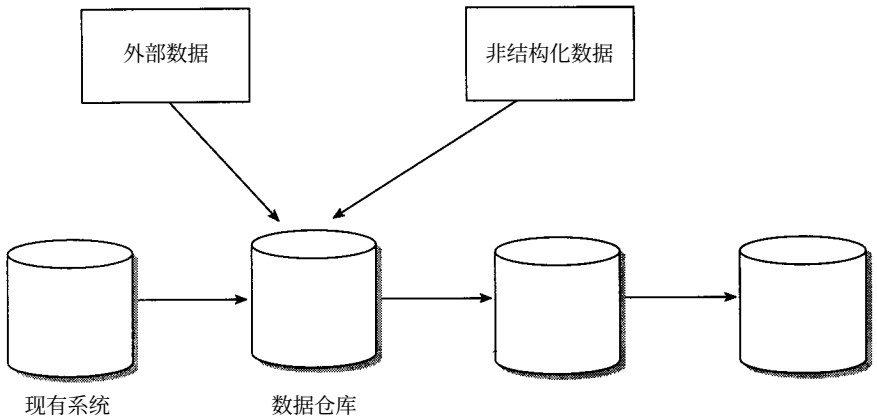

图8-1　外部数据与非结构化数据都归入数据仓库

数据仓库是存储外部与非结构化数据的理想场所。如果外部数据与非结构化数据没有存放在一个集中确定的位置，势必会产生一些问题。图8 -2表明当外部数据与非结构化数据以非规范的形式进入企业时，就失去了数据来源的标识，并且不管怎样有次序地使用数据都不存在数据间的协同。

典型地，当外部数据没有进入数据仓库时，这些数据就通过 PC进入企业。在PC级上，本质上进入的数据不存在任何错误。但是当数据在 PC级上进入时，几乎都是通过电子表格方式手工地进入数据仓库，并且绝对没有试图捕获有关附加在数据上的任何数据源的信息。例如，在图8-2中分析员得到了《华尔街日报》中一个报告。第二天，这个分析员采用《华尔街日报》中的数据作为某个报告的一部分，然而当此报告进入企业主数据流时，有关原始数据源的信息就丢失了。

获取外部数据的自由方式所导致的另一个困难是，在以后某个时刻很难回忆起这些数据的意义。数据进入企业系统，使用一次后便消失了。即使几星期后，也很难重新访问这些数据。这是很不幸的，因为许多来自外部源的数据在一定时间范围内都是非常有用的。

来自外部数据源的数据类型是多种多样的。一些典型的主要数据来源如下：

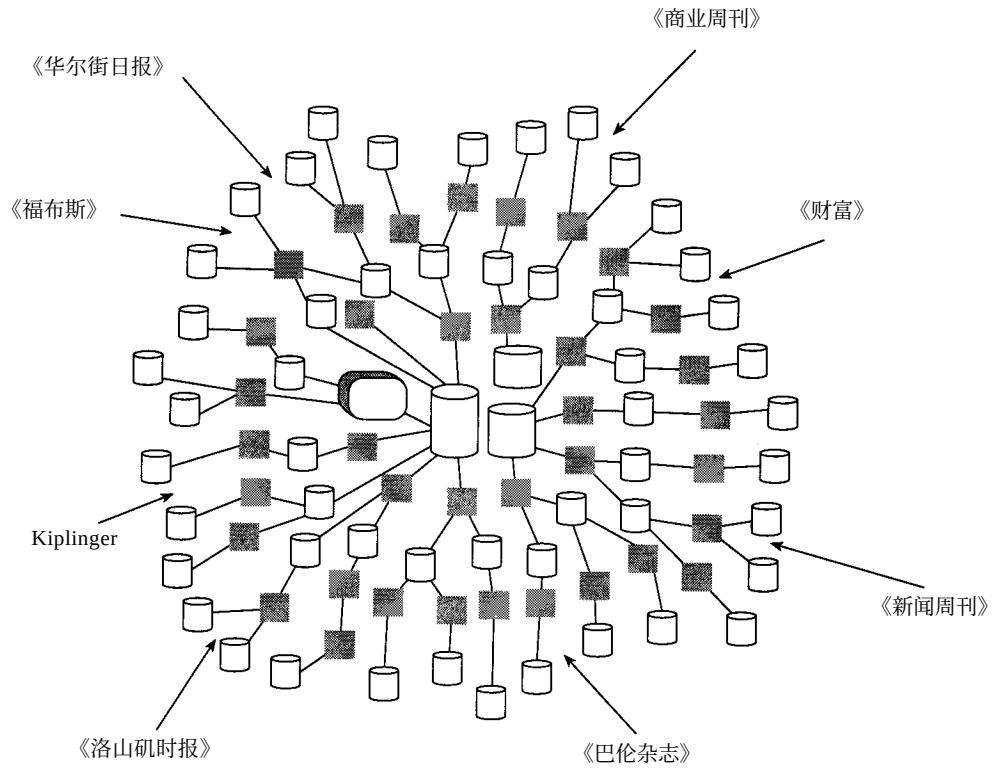

图8-2 外部数据与非结构化数据所带来的问题：

- 当数据进入企业时，其来源被删掉了
- 一个分析员没有意识到另一个分析员已将类似的信息输入

■ 《华尔街日报》。
■ 《商业周刊》。
■ 《福布斯》。
■ 《财富》。
■ 工业通信。
■ 技术报告。
■ 《Dun and Bradstreet》。
■ 咨询员专门为企业研究的报告。
■ 《Equifax reports》。
■ 竞争分析报告。
■ 市场比较与分析报告。
■ 销售分析与比较报告。
■ 新产品通告。

另外，还有一些企业内部的报告也同样值得注意：

■ 审计季报。
■ 年度报告。

8.1 数据仓库中的外部数据/非结构化数据

在数据仓库中，存在一些与外部数据 /非结构化数据的使用和存储相关的问题。在图8 -3中所给出的非结构化数据所存在的一个问题是可用频率。与内部出现的数据不同，外部数据没有呈现的真正模式。当为了确保捕获正确的数据，而必须建立永久的监控方式时，呈现的频率就成为一个问题。

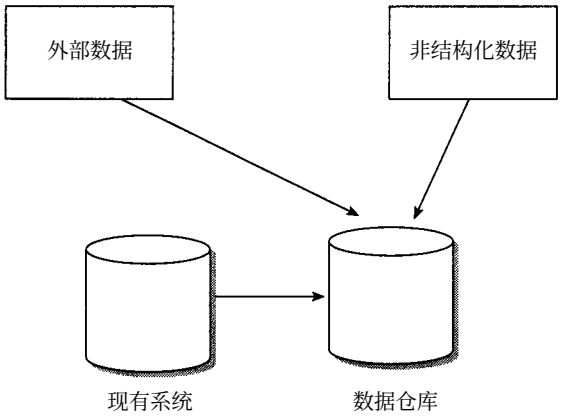

图8-3 与外部数据有关的问题：

- 访问的频率
- 数据的形式
- 不可预测性

外部数据的第二个问题是数据的形式。外部数据的形式是完全没有规则的。为了使之有用并能放入数据仓库，就必须对外部数据进行一定的重新格式化，将其转化成为内部可接受的与有用的形式。

导致外部数据难以获得的第三个因素是其不可预测性。外部数据可能在几乎任何时侯来源于任何实际的数据源。外部数据本身可用性的不可预测性使得很难一致地与完全地获得所需要的外部数据。

除了来自于期刊上的文章和咨询报告之外，外部数据的另一种来源是现在能够自动操作的整个数据类，即非结构化数据，如图8 -4所示。

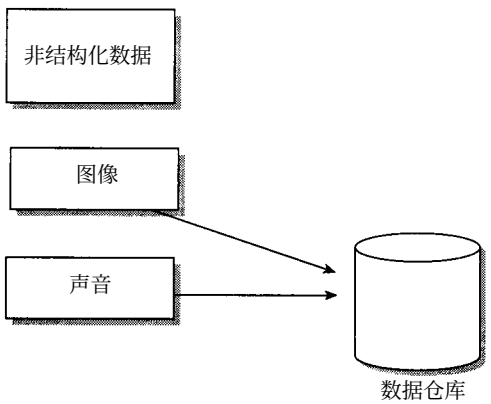

图8-4 能够存储在数据仓库中的非结构化数据的一些形式

非结构化数据的两种最常见的类型是图像和声音。图像数据是以图片形式存储的数据，声音数据是数字存储的数据并能转换为一种声音格式。图像数据和声音数据的问题主要是技术上的，获取并处理图像和声音数据的技术并不象传统技术那么成熟。另外，即使能够获取图像和声音数据，存储它们也需要大量的DASD设备，并且它们的查找、显示或回放都可能是不方便的和缓慢的。

无论如何，非结构化信息的获取及其在数据仓库中的存储都还是有许多可能性的。

8.2 元数据和外部数据

在任何方案中，元数据都是数据仓库的一种重要组成部分。但是，当面对存储和管理外部数据与非结构化数据时，元数据的作用呈现出完全不同的一面。图8 -5表示了元数据的作用。

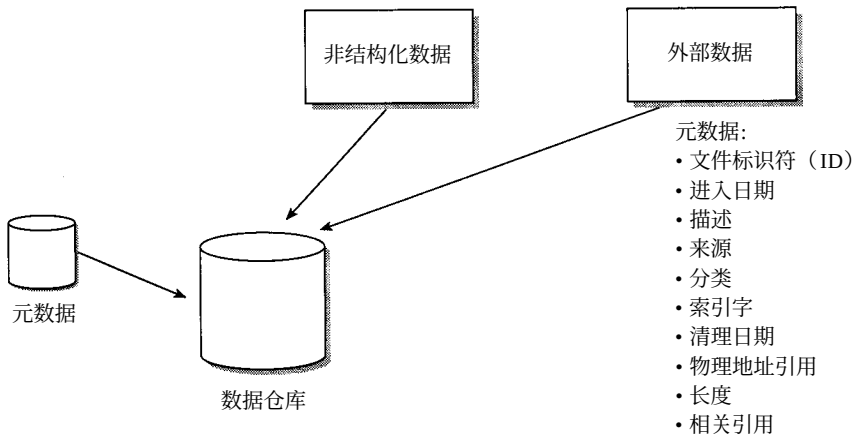

图8-5 对外部数据/非结构化数据，元数据起着新的作用

元数据是很重要的，因为在数据仓库环境中正是通过元数据来对外部数据进行注册、访问与控制的。在数据仓库中对于外部数据来说，元数据的典型内容就是元数据重要性的最好解释，例如：

- 文件标识符（ID）。
- 进入数据仓库的日期。
- 文件描述。
- 文件来源。
- 文件来源的日期。
- 文件的分类。
- 索引字。
- 清理日期。
- 物理地址引用。
- 文件长度。
- 相关参考。

正是通过元数据，管理者来判断许多有关外部数据的信息。在许多情况下，管理者甚至不看

源文件，只看元数据。在清除不相关的或过时的文件中，浏览元数据可为管理者减少大量的工作。

就外部数据而言，适当地建立和维护元数据对于数据仓库的操作是完全必要的。

与元数据有关的另一种数据类型是通知数据。图8 -6所示的通知数据仅仅是一个为系统的用户创建的文件，它表明与用户有关的数据分类。当数据进入数据仓库和元数据时，要进行检查来发现谁与该数据有关。一旦发现获得了与某人有关的数据，就发出通知，以便让他或她将来知道已经获得了有关的外部数据。

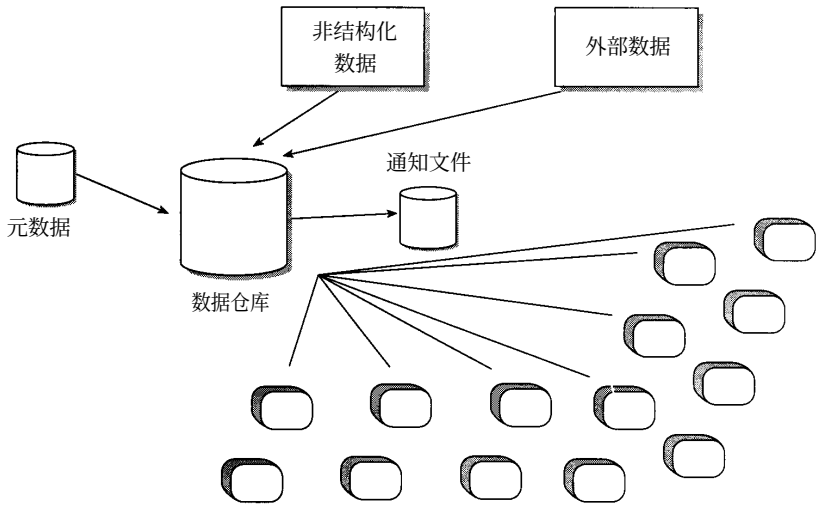

图8-6 外部数据和元数据的另一个优点是能够创建专门通知文件

8.3 存储外部数据/非结构化数据

如果方便且费用允许的话，外部数据 /非结构化数据实际上可以存储在数据仓库中。但在

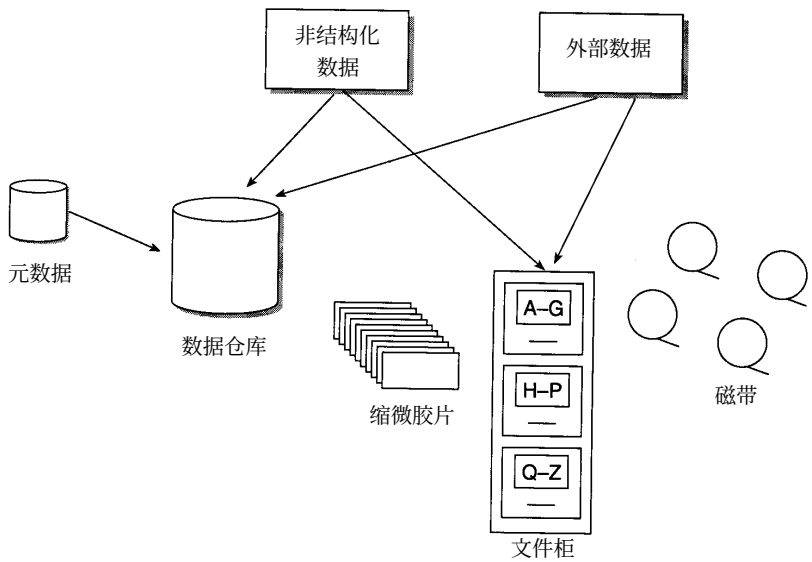

图8-7 在任何情况下，外部数据/非结构化数据总是与元数据一起进行登记的，但实际数据依据其大小和存取概率来决定是否存储在数据仓库中

许多情况下，将所有的外部数据存储在数据仓库中是不可能的或者是不经济的。于是，在数据仓库的元数据中，对外部数据/非结构化数据进行登录，创建一个条目以说明什么地方能找到外部数据本身，外部数据存储在任何一个方便的地方，如图8-7所示。外部数据可能存储在文件柜中，缩微胶片、磁带上，等等。当然，如果需要的话，外部数据可以存储在数据仓库中。

8.4 外部数据/非结构化数据的不同组成部分

外部数据/非结构化数据的重要设计问题之一是其经常包括许多不同的组成部分，其中一些组成部分比另外一些更有用。作为一个例子，考虑一个产品的完整生产历史记录。生产过程的某些方面是很重要的，如从开始到最后装配的时间。另一个重要的生产度量是所有非装配的原材料的总成本。但还有许多其他不重要的信息同样也与生产信息相关，例如生产的实际日期、装运说明书、生产时的温度。

为了管理这些数据，有经验的DSS分析员或工程师需要决定什么是最重要的数据部分。然后将最重要的数据存储在一个联机的、容易访问的位置。这是一个存储和访问效率的问题。其余不重要的细节不能丢弃，而是将其放在大容量的存储位置。这样，大量的非结构化数据就能够有效地存储和管理。

8.5 建模与外部数据/非结构化数据

数据模型和外部数据的作用是什么？图8-8反映了这个问题。数据模型通常的作用是根据设计塑造环境。但外部数据与非结构化数据是根本不可塑的。看起来好象数据模型和外部数据之间的关系很小。能做的事最多不过是在涉及的关键词和关键字解释范围内，记录数据模型和外部数据之间的区别。任何将数据模型用于数据的重大改造的企图都将是一个错误。

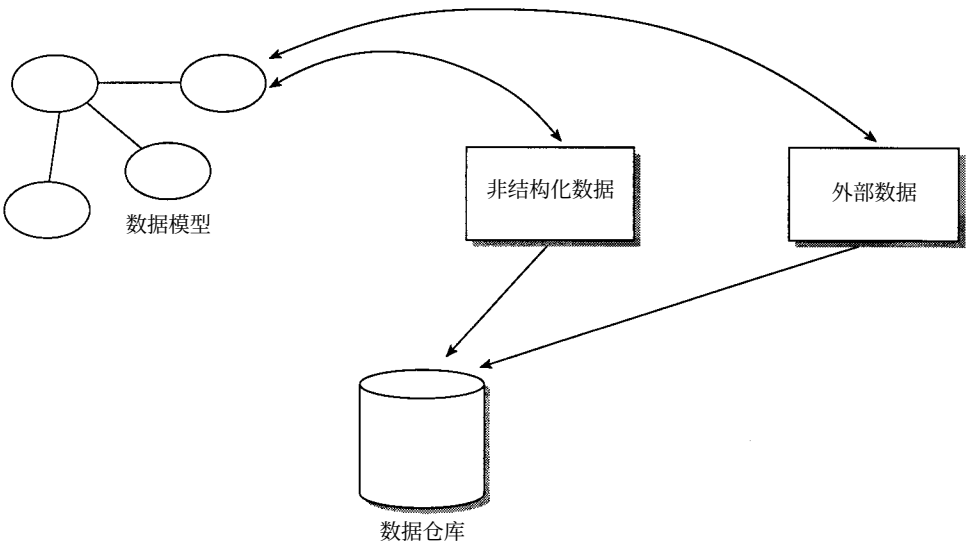

图8-8 外部数据/非结构化数据与数据模型通常只有极少的相似，而且数据模型对外部数据和非结构化数据的改造无能为力

8.6 间接报告

不仅原始数据能放入数据仓库，而且当数据循环重复时，间接报告可以按时间根据细节数据来产生。例如图8-9所示的月底道·琼斯平均指数报告。

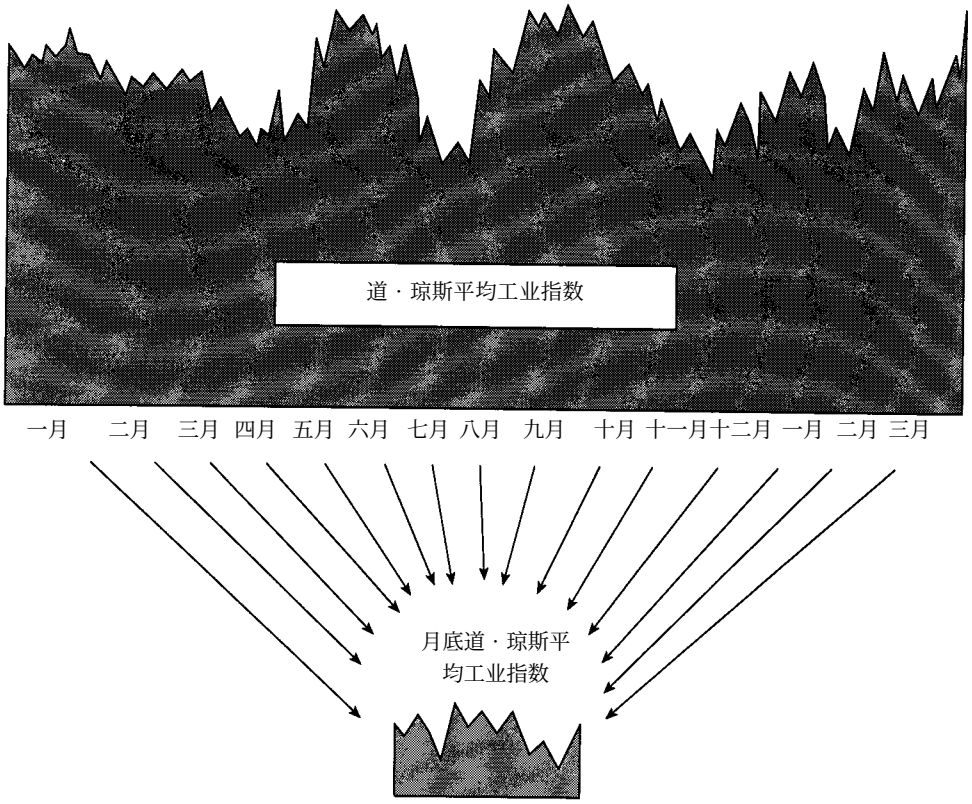

图8-9 根据每日或每月的信息创建一个总结报告

在图8-9中，道·琼斯平均指数信息每天输入数据仓库环境。每天的信息是有用的，但更感兴趣的是由此产生的长期趋势信息。月底，有关道·琼斯平均指数的信息记入一个“间接”报告中，于是间接报告就成为数据仓库中外部数据存储的一部分。

8.7 外部数据归档

每一条信息（外部的或其他的）都有个有用的生命周期。一旦超出了这个生命周期，保存这些信息就不经济了。管理外部数据的一个基本部分就是决定数据的使用生命周期。即使确定了使用生命周期，仍然还有一个数据是否丢弃或归档的问题。通常，外部数据可能从数据仓库移出并放在较便宜的存储设备。元数据对外部数据的引用应该更新以反映新的存储位置，并且新的存储位置仍然保留在元数据存储单元中。访问元数据存储单元的费用是很低的，因此一旦放在那里，最好留在那里。

8.8 内部数据与外部数据的比较

可以对外部数据做的最有用的事情是过一段时间将其与内部数据进行比较。这种比较可

以使管理工作“以树看森林”。对非常即时性的个体的活动和趋势与非常普遍的活动和趋势进行比较，能使高级管理人员获得在其他地方不能得到的见解。图8 -10给出这样的比较。

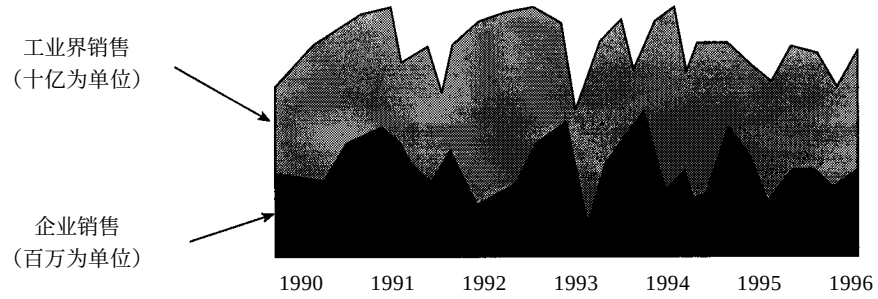

图8-10 外部数据与内部数据比较可以是很明晰的

当进行外部数据和内部数据的比较时，假设比较是在一个公共键码上进行的。任何其他的假设都会使外部数据和内部数据的比较丢失其有用性。不幸的是，在外部数据和内部数据之间实际得到一个公共的键码基础是不容易的。

为了理解这种难度，我们来看两个例子。一个例子中所卖的商品是大的、昂贵的物品，如汽车或电视机。为了进行有意义的比较，由实际销路卖出的商品需要进行度量。零售商的实际售出是比较的基础。不幸的是，数据的外部数据源使用的键码结构与内部系统使用的键码结构并不相同。要将外部数据源转换成内部数据源的键码结构，反之亦然。这种转换不是件小事情。

现在来考虑体积大、成本低的商品的销售度量，例如可乐。公司的内部销售曲线反映了可乐的销售，但外部销售数据将可乐的销售与其他饮料（如啤酒）的销售混在一起。将这两种类型销售数据进行比较将导致错误的结论。为了进行有意义的比较，需要有一个外部销售数据的“清理过程”以便只包含可乐。事实上，只要可能，将包括各种各样生产与销售的瓶装可乐 。不仅要将啤酒从外部销售数据中剔除出去，也要将非竞争的可乐类型剔除出去。

8.9 小结

数据仓库不仅仅能够拥有内部的、结构化的数据。还有许多与企业运营有关的来自企业以外数据源的信息。

外部数据是捕获的，而元数据是存储在数据仓库中的数据。元数据为高级管理人员检索信息服务。另外，只要有新的项目进入数据仓库，经常会提供一种“通知”服务。

外部数据/非结构化数据实际并不一定存储在数据仓库中。