第3章

大数据分析生命周期

【导读案例】

数据分析的五大思维方式

数据可视化的价值在于呈现数据背后的规律，从而帮助使用者提高决策效率与能力。对用户数据的分析是进行可视化系统建设必不可少的一个环节。首先，要知道什么叫数据分析。其实从数据到信息的这个过程就是数据分析。数据本身并没有什么价值，有价值的是从数据中提取出来的信息。其次，要搞清楚数据分析的目的是什么，目的是解决现实中的某个问题或者满足现实中的某个需求。

在这个从数据到信息的过程中，有一些固定的思路，或者称之为思维方式。

第一大思维：对照，俗称对比。单独看一个数据是不会有感觉的，必须与另一个数据做对比才能找到感觉(见图 3-1)。在图中单独看图 3-1(a)无感觉，而图 3-1(b)经过对比就会发现两天的销量实际上差了一大截。

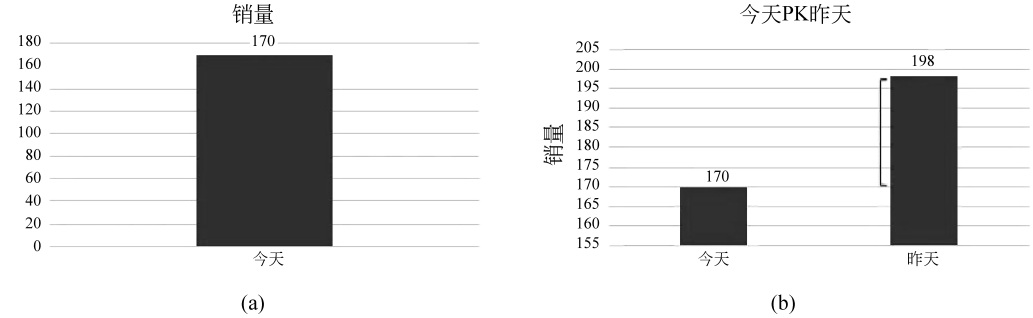

图 3-1　对比

对照是最基本也是最重要的思路，在现实中的应用非常广泛。例如，选款测算、监控店铺数据等，这些过程就是在做“对照”。分析人员拿到数据后，如果数据是独立的，无法进行对比的话，就无法判断，即无法从数据中读取有用的信息。

第二大思维：拆分。分析这个词的字面理解，就是拆解和分析，可见拆分在数据分析中的重要性。当某个维度可以对比的时候，我们选择对比。在对比后发现问题需要找出原因的时候，或者根本就无法对比的时候，就用到拆分了。

下面来看这样一个场景：运营小美经过对比店铺的数据，发现今天的销售额只有昨

天的50%，这个时候，再怎么对比销售额这个维度，已经没有意义了。这时需要对销售额这个维度做分解，拆分指标。

销售额＝成交用户数×客单价

其中，成交用户数又等于访客数×转化率。例如，图3-2(a)是一个指标公式的拆解，图3-2(b)是对流量的组成成分做的简单分解(还可以分得更细更全)。拆分后的结果相对于拆分前会清晰许多，便于分析查找细节。可见，拆分是分析人员必备的思维之一。

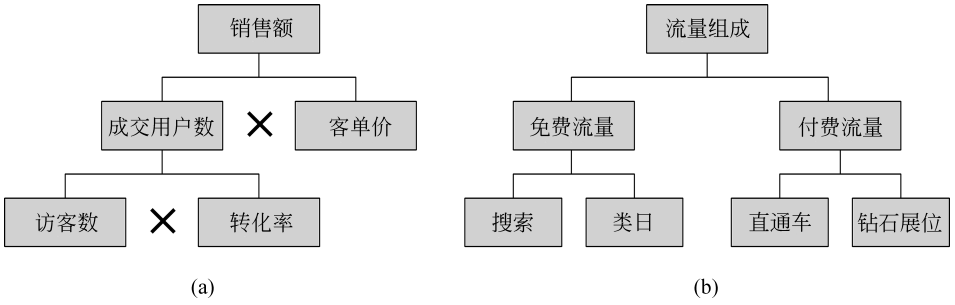

图3-2 拆分

第三大思维：降维。读者是否有面对一大堆维度的数据却束手无策的经历？当数据维度太多的时候，不可能每个维度都拿来分析，可以从一些有关联的指标中筛选出代表的维度(见表3-1)。

表3-1 关联指标的维度

日期	浏览量	访客数	访问深度	销售额	销售量	订单数	成交用户量	客单价	转化率
2015/2/1	2584	957	2.7	9045	96	80	67	135	7%
2015/2/2	2625	1450	2.5	9570	125	104	67	110	6%
2015/2/3	2572	1286	2.0	12 780	130	108	90	142	7%
2015/2/4	4125	1650	2.5	16 345	143	119	99	155	6%
2015/2/5	3699	1233	3.0	8362	107	89	74	113	6%
2015/2/6	4115	1286	3.2	14 040	130	108	90	166	7%

这么多的维度不必每个都分析。我们知道成交用户数÷访客数＝转化率，当存在这种维度可以通过其他两个维度经过计算转化出来的时候，就可以降维。

成交用户数、访客数和转化率，只要三选二即可。另外，成交用户数×客单价＝销售额，这三个也可以三择二。我们一般只关心对自己有用的数据，当有某些维度的数据与我们的分析无关时，就可以筛选掉，达到“降维”的目的。

第四大思维：增维。增维和降维是对应的，有降必有增。在当前的维度不能很好地解释我们的问题时，就需要对数据做一个运算，增加一个指标(见表3-2)。

我们发现一个搜索指数和一个宝贝数，这两个指标一个代表需求，一个代表竞争，有很多人把搜索指数÷宝贝数＝倍数，用倍数来代表一个词的竞争度，这种做法就是在增维。增加的维度也称为“辅助列”。

表 3-2　增加指标

序号	关键词	搜索人气	搜索指数	占比	点击指数	商城点击占比	点击率	当前宝贝数
1	毛呢外套	242 165	1 119 253	58.81%	512 673	30.76%	45.08%	2 448 482
2	毛呢外套(女)	33 285	144 688	7.29%	80 240	48.88%	54.79%	2 448 368
3	韩版毛呢外套	7460	29 714	1.45%	15 070	21.385%	50.04%	1 035 325
4	小香风毛呢外套	6400	22 543	1.09%	11.143	22.34%	48.72%	60.258
5	斗篷毛呢外套	5 463	23 443	1.14%	11.328	19.87%	19.87%	108.816

增维和降维是必须对数据的意义有充分的了解后，为了方便进行分析，有目的地对数据进行转换运算。

第五大思维：假说。当我们迷茫的时候，可以应用“假说”。假说是统计学中的专业名词，俗称假设。当我们不知道结果或者有几种选择的时候，那么就可以召唤“假说”，先假设有了结果，然后运用逆向思维。

从结果到原因，要有怎么样的因，才能产生这种结果。这有点儿寻根的味道。那么，我们可以知道，现在满足了多少因，还需要多少因。如果是在多选的情况下，就可以通过这种方法来找到最佳路径(决策)。

当然，“假说”的威力不仅如此。“假说”可是一匹天马(行空)，除了结果可以假设，过程也可以被假设。

资料来源：公众号零一. 数字冰雹大数据可视化. 2013-3-2.

阅读上文，请思考、分析并简单记录。

(1) 请回顾，文中介绍的数据分析的五大思维方式是指什么？

答：

(2) 试分析，这五大思维方式在运用时有顺序要求吗？为什么？

答：

(3) 请思考，列举并描述一个运用这五大思维方式(或者之一)来进行数据分析的例子。

答：

(4) 请简单描述你所知道的上一周发生的国际、国内或者身边的大事。

答：_____

_____

_____

_____

3.1 大数据分析生命周期概述

从组织上讲，采用大数据会改变商业分析的途径。大数据分析的生命周期从大数据项目商业案例的创立开始，到保证分析结果部署在组织中并最大化地创造价值时结束。在数据识别、获取、过滤、提取、清理和聚合过程中有许多步骤，这些都是在数据分析之前所必需的。

由于被处理数据的容量、速率和多样性的特点，大数据分析不同于传统的数据分析。为了处理大数据分析需求的多样性，需要一步步地使用采集、处理、分析和重用数据等方法。大数据分析生命周期可以组织和管理与大数据分析相关的任务和活动。从大数据的采用和规划的角度来看，除了生命周期以外，还必须考虑数据分析团队的培训、教育、工具和人员配备的问题。生命周期的执行需要组织重视培养或者雇佣新的具有相关能力的人。

大数据分析的生命周期可以分为 9 个阶段(见图 3-3)。

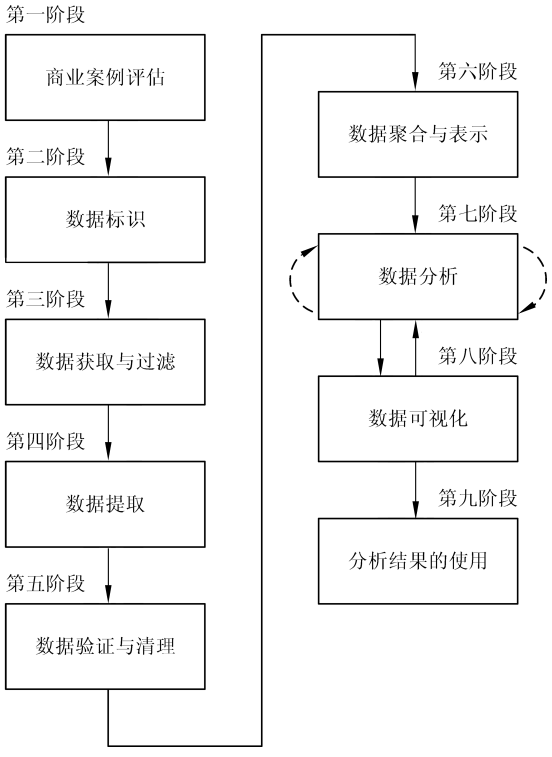

图 3-3 大数据分析生命周期的 9 个阶段

3.2 商业案例评估

在分析阶段中，每一个大数据分析生命周期都必须起始于一个被很好定义的商业案例，它有着清晰的执行分析的理由、动机和目标，并且应该在着手分析之前就被创建、评估和改进。

商业分析案例的评估能够帮助决策者了解需要使用哪些商业资源，需要面临哪些挑战。另外，在这个环节中详细区分关键绩效指标，能够更好地明确分析结果的评估标准和评估路线。如果关键绩效指标不容易获取，则需要努力使这个分析项目变得 SMART，即 Specific（具体的）、Measurable（可衡量的）、Attainable（可实现的）、Relevant（相关的）和 Timely（及时的）。

基于商业案例中记录的商业需求，可以确定所定位的商业问题是否是真正的大数据问题。为此，这个商务问题必须直接与一个或多个大数据的特点相关。

同样还要注意的是，本阶段的另一个结果是确定执行这个分析项目的基本预算。任何如工具、硬件、培训等需要购买的东西都要提前确定，以保证可以对预期投入和最终实现目标所产生的收益进行衡量。比起能够反复使用前期投入的后期迭代，大数据分析生命周期的初始迭代需要在大数据技术、产品和训练上有更多的前期投入。

3.3 数据标识

数据标识阶段主要用来标识分析项目所需要的数据集和所需的资源。标识种类众多的数据资源可能会提高找到隐藏模式和相互关系的可能性。例如，为了提供洞察能力，尽可能多地标识出各种类型的相关数据资源非常有用，尤其是当我们探索的目标并不是那么明确的时候。

根据分析项目的业务范围和业务问题的性质，我们需要的数据集和它的数据源可能是企业内部和/或企业外部的。在内部数据集的情况下，如数据集市和操作系统等一系列可供使用的内部资源数据集，往往靠预定义的数据集规范来进行收集和匹配。在外部数据集的情况下，如数据市场和公开可用的数据集这样的一系列可能的第三方数据集会被收集。一些外部数据的形式则会内嵌到博客和一些基于内容的网站中，这些数据需要通过自动化工具来获取。

3.4 数据获取与过滤

在数据获取和过滤阶段，前一阶段标识的数据已经从所有的数据资源中获取，这些数据接下来会被归类并进行自动过滤，以去掉被污染的数据和对分析对象毫无价值的数据。

根据数据集的类型，数据可能会是档案文件，如购入的第三方数据；可能需要 API 集成，如微博、微信上的数据。在许多情况下，我们得到的数据常常是并不相关的数据，特别是外部的非结构化数据，这些数据会在过滤程序中被丢弃。

被定义为“坏”数据的，是其包括遗失或毫无意义的值或是无效的数据类型。但是，被一种分析过程过滤掉的数据集还有可能对于另一种不同类型的分析过程具有价值。因此，在执行过滤前存储一份原文件备份是个不错的选择。为了节省存储空间，可以对原文件备份进行压缩。

内部数据或外部数据在生成或进入企业边界后都需要继续保存。为了满足批处理分析的要求，数据必须在分析之前存储在磁盘中，而在实时分析之后，数据需要再存储到磁盘中。

元数据会通过自动操作添加到内部和外部的数据资源中来改善分类和查询(见图 3-4)。扩充的元数据例子主要包括数据集的大小和结构、资源信息、日期、创建或收集的时间、特定语言的信息等。确保元数据能够被机器读取并传送到数据分析的下一个阶段是至关重要的，它能够帮助我们在大数据分析的生命周期中保留数据的起源信息，保证数据的精确性和高质量。

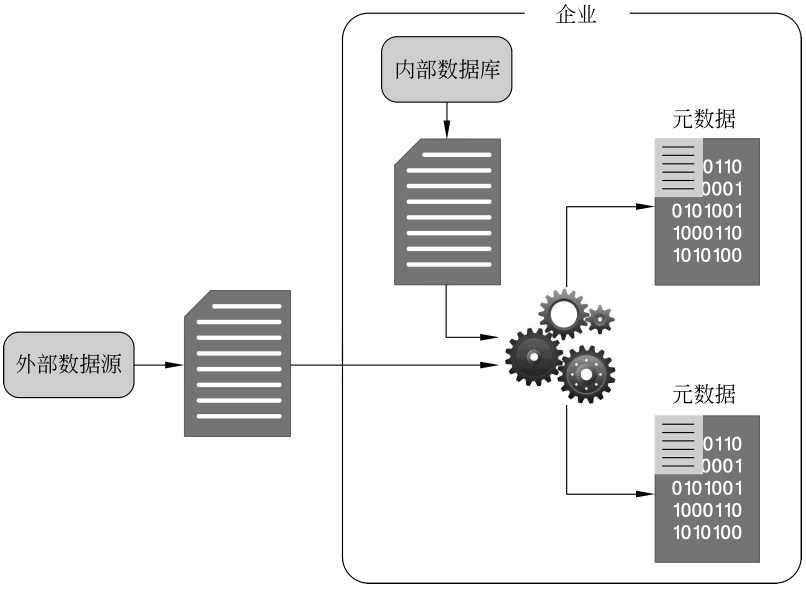

图 3-4 元数据从内部资源和外部资源中添加到数据中

3.5 数据提取

为分析而输入的一些数据可能会与大数据解决方案产生格式上的不兼容，这样的数据往往来自于外部资源。数据提取阶段主要是要提取不同的数据，并将其转换为大数据解决方案中可用于数据分析的格式。

需要提取和转换的程度取决于分析的类型和大数据解决方案的能力。例如，如果相关的大数据解决方案已经能够直接加工文件，那么从有限的文本数据(如网络服务器日志文件)中提取需要的域，可能就不必要了。类似地，如果大数据解决方案可以直接以本地格式读取文稿的话，对于需要总览整个文稿的文本分析而言，文本的提取过程就会简化

许多。

图 3-5 显示了从没有更多转换需求的 XML 文档中对注释和内嵌用户 ID 的提取。

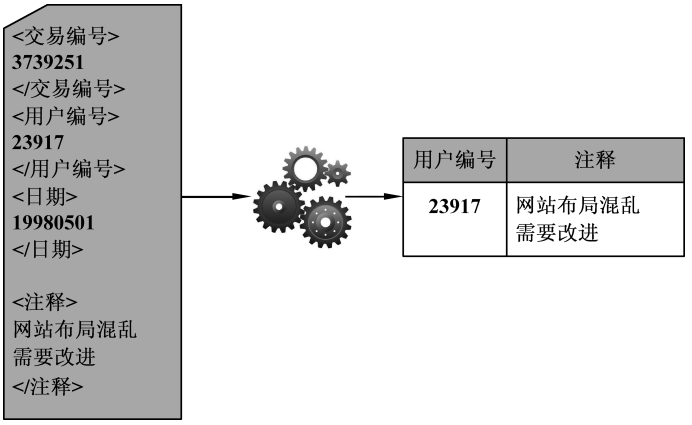

图 3-5　从 XML 文档中提取注释和用户编号

图 3-6 显示了从单个 JSON 字段中提取用户的经纬度坐标。为了满足大数据解决方案的需求，将数据分为两个不同的域，这就需要做进一步的数据转换。

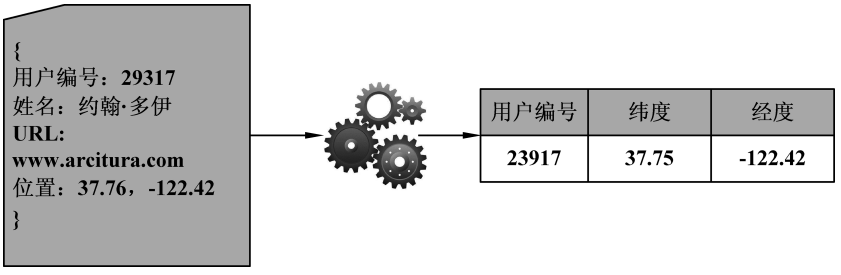

图 3-6　从单个 JSON 文件中提取用户编号和相关信息

3.6　数据验证与清理

无效数据会歪曲和伪造分析的结果。和传统的企业数据那种数据结构被提前定义好、数据也被提前校验的方式不同，大数据分析的数据输入往往没有任何的参考和验证来进行结构化操作，其复杂性会进一步使数据集的验证约束变得困难。

数据验证和清理阶段是为了整合验证规则并移除已知的无效数据。大数据经常会从不同的数据集中接收到冗余的数据，这些冗余数据往往会为了整合验证字段、填充无效数据而被用来探索有联系的数据集。数据验证会检验具有内在联系的数据集，填充遗失的有效数据。

对于批处理分析，数据验证与抽取可以通过离线 ETL（抽取/转换/加载）来执行。对于实时分析，则需要一个更加复杂的在内存中的系统来对从资源中得到的数据进行处理，在确认问题数据的准确性和质量时，来源信息往往扮演着十分重要的角色。有的时候，看

起来无效的数据(见图 3-7)可能在其他隐藏模式和趋势中具有价值,在新的模式中可能有意义。

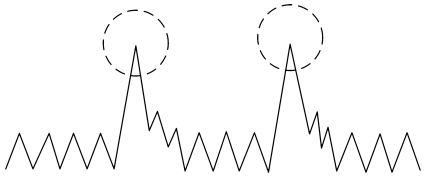

图 3-7 无效数据的存在造成了一个峰值

3.7 数据聚合与表示

数据可以在多个数据集中传播,这要求这些数据集通过相同的域被连接在一起,就像日期和 ID。在其他情况下,相同的数据域可能会出现在不同的数据集中,如出生日期。无论哪种方式都需要对数据进行核对的方法或者需要确定表示正确值的数据集。

数据聚合和表示阶段是专门为了将多个数据集进行聚合,从而获得一个统一的视图。在这个阶段会因为以下情况变得复杂。

(1) 数据结构——数据格式相同时,数据模型可能不同。

(2) 语义——在两个不同的数据集中,具有不同标记的值可能表示同样的内容,如“姓”和“姓氏”。

通过大数据解决方案处理的大量数据能够使数据聚合变成一个时间和劳动密集型的操作。调和这些差异需要可以自动执行的无须人工干预的复杂逻辑。

在此阶段,需要考虑未来的数据分析需求,以帮助数据的可重用性。是否需要对数据进行聚合,了解同样的数据能以不同形式来存储十分重要。一种形式可能比另一种更适合特定的分析类型。例如,如果需要访问个别数据字段,以 BLOB(Binary Large Object,二进制大对象)存储的数据就会变得没有多大的用处。

BLOB 是一个可以存储二进制文件的容器。在计算机中,BLOB 常常是数据库中用来存储二进制文件的字段类型。BLOB 是一个大文件,典型的 BLOB 是一张图片或一个声音文件,由于它们的尺寸,必须使用特殊的方式来处理(例如上传、下载或者存放到一个数据库)。在 MySQL 中,BLOB 是一个类型系列,例如 TinyBlob 等。

由大数据解决方案进行标准化的数据结构可以作为一个标准的共同特征被用于一系列的分析技术和项目。这可能需要建立一个像非结构化数据库一样的中央标准分析仓库(见图 3-8)。

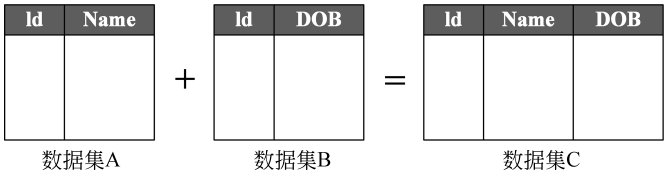

图 3-8 使用 ID 域聚集两个数据域的简单例子

图 3-9 展示了存储在两种不同格式中的相同数据块。数据集 A 包含所需的数据块，但是由于它是 BLOB 的一部分而不容易访问。数据集 B 包含相同的以列为基础来存储的数据块，使得每个字段都被单独查询到。

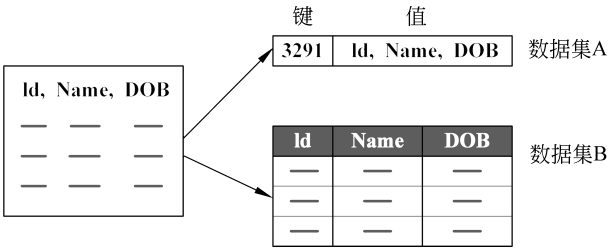

图 3-9　数据集 A 和 B 能通过大数据解决方案结合起来创建一个标准化的数据结构

3.8　数据分析

数据分析阶段致力于执行实际的分析任务，通常会涉及一种或多种类型的数据分析。在这个阶段，数据可以自然迭代，尤其在数据分析是探索性分析的情况下，分析过程会一直重复，直到发现适当的模式或者相关性。

根据所需的分析结果的类型，这个阶段可以被尽可能地简化为查询数据集以实现用于比较的聚合。另一方面，它可以像结合数据挖掘和复杂统计分析技术来发现各种模式和异常，或是生成一个统计或是数学模型来描述变量关系一样具有挑战性。

数据分析可以分为验证分析和探索分析两类，后者常常与数据挖掘相联系。

验证性数据分析是一种演绎方法，即先提出被调查现象的原因，被提出的原因或者假说称为一个假设。接下来使用数据分析以验证和反驳这个假设，并为这些具体的问题提供明确的答案。我们常常会使用数据采样技术，意料之外的发现或异常经常会被忽略，因为预定的原因是一个假设。

探索性数据分析是一种与数据挖掘紧密结合的归纳法。在这个过程中没有假想的或是预定的假设产生。相反，数据会通过分析探索来发展一种对于现象起因的理解。尽管它可能无法提供明确的答案，但这种方法会提供一个大致的方向以便发现模式或异常。

3.9　数据可视化

如果只有分析师才能解释数据分析结果，那么分析海量数据并发现有用的见解的能力就没有什么价值了。数据可视化阶段致力于使用数据可视化技术和工具，并通过图形表示有效的分析结果(见图 3-10)。为了从分析中获取价值并在随后拥有向下一阶段提供反馈的能力，商务用户必须充分理解数据分析的结果。

完成数据可视化阶段得到的结果能够为用户提供执行可视化分析的能力，这能够让用户去发现一些未曾预估到的问题的答案。相同的结果可能会以许多不同的方式来呈

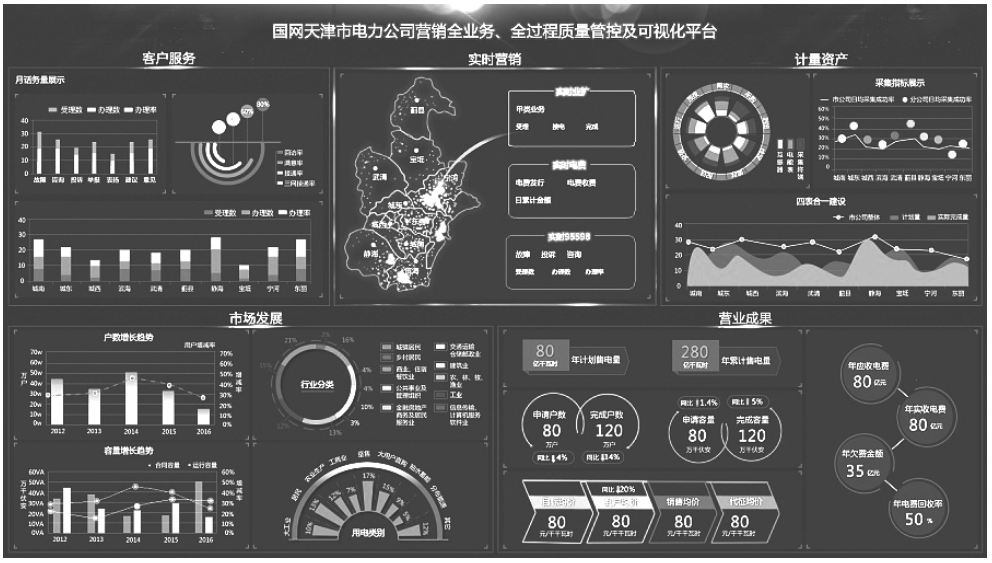

图 3-10 数据分析仪表盘

现，这会影响最终结果的解释。因此，重要的是保证商务域在相应环境中使用最合适的可视化技术。

另一个必须要记住的方面是：为了让用户了解最终的积累或者汇总结果是如何产生的，提供一种相对简单的统计方法也是至关重要的。

3.10 分析结果的使用

大数据分析结果可以用来为商业使用者提供商业决策支持，例如使用图表之类的工具，可以为使用者提供更多使用这些分析结果的机会。在分析结果的使用阶段，致力于确定如何以及在哪里处理分析数据能保证产出更大的价值。

基于要解决的分析问题本身的性质，分析结果很可能会产生对被分析的数据内部一些模式和关系有着新的看法的“模型”。这个模型可能看起来会比较像一些数据公式和规则的集合，它们可以用来改进商业进程的逻辑和应用系统的逻辑，也可以作为新的系统或者软件的基础。

在这个阶段常常会被探索的领域主要有以下几种。

（1）企业系统的输入。数据分析的结果可以自动或者手动输入到企业系统中，用来改进系统的行为模式。例如，在线商店可以通过处理用户关系分析结果来改进产品推荐方式。新的模型可以在现有的企业系统或是在新系统的基础上改善操作逻辑。

（2）商务进程优化。在数据分析过程中识别出的模式、关系和异常能够用来改善商务进程。例如，作为供应链的一部分整合运输线路。模型也有机会能够改善商务流程逻辑。

（3）警报。数据分析的结果可以作为现有警报的输入或者是新警报的基础。例如，

可以创建通过电子邮件或者短信的警报来提醒用户采取纠正措施。

作　　业

1. 大数据分析的生命周期中，在数据(　　)过程中有许多步骤，这些都是在数据分析之前所必需的。

A. 识别、获取、过滤、提取、清理和聚合

B. 打印、计算、过滤、提取、清理和聚合

C. 统计、计算、过滤、存储、清理和聚合

D. 存储、提取、统计、计算、分析和打印

2. 由于被处理数据的容量、速率和多样性的特点，大数据分析不同于传统的数据分析。数据分析生命周期可以(　　)与大数据分析相关的任务和活动。

A. 收集和整理　　B. 组织和管理

C. 分析和处理　　D. 打印和存储

3. 每一个大数据分析生命周期都必须起始于一个被很好定义的(　　)，它应该在着手分析任务之前被创建、评估和改进，并且有着清晰的执行分析的理由、动机和目标。

A. 商业计划　　B. 社会目标

C. 营利方针　　D. 商业案例

4. 在大数据分析商业案例的评估中，如果关键绩效指标不容易获取，则需要努力使这个分析项目变得 SMART，即(　　)。

A. 实际的、大胆的、有价值的、可分析的

B. 有风险的、有机会的、能实现的、有价值的

C. 具体的、可衡量的、可实现的、相关的、及时的

D. 有理想的、有价值的、有前途的、能实现的

5. 大数据分析的生命周期可以分为 9 个阶段，但以下(　　)不是其中的阶段之一。

A. 商业案例评估　　B. 数值计算

C. 数据获取与过滤　　D. 数据提取

6. 大数据分析的生命周期可以分为 9 个阶段，但以下(　　)不是其中的阶段之一。

A. 数据删减　　B. 数据聚合与表示

C. 数据分析　　D. 数据可视化

7. 大数据分析的生命周期可以分为 9 个阶段，但以下(　　)不是其中的阶段之一。

A. 数据标识　　B. 数据验证与清理

C. 分析结果的使用　　D. 数据打印

8. 数据标识阶段主要是用来标识分析项目所需要的数据集和所需的资源。标识种类众多的数据资源可能会提高找到(　　)的可能性。

A. 数据获取和数据打印　　B. 算法分析和打印模式

C. 隐藏模式和相互关系　　D. 隐藏价值和潜在商机

9. 在数据获取和过滤阶段，从所有的数据资源中获取到的所需要的数据接下来会

被(　　)并进行自动过滤,以去除掉所有被污染的数据和对于分析对象毫无价值的数据。

A. 整理　　B. 归类　　C. 打印　　D. 处理

10. 数据提取阶段主要是要提取不同的数据,并将其转换为大数据解决方案中可用于(　　)的格式。需要提取和转换的程度取决于分析的类型和大数据解决方案的能力。

A. 数据分析　　B. 打印输出　　C. 数据存储　　D. 数据整合

11. 大数据分析的数据输入中,数据验证和清理阶段是为了(　　)并移除任何已知的无效数据。

A. 完善数据结构　　B. 建立存储结构

C. 整合验证规则　　D. 充实合理数据

12. 数据聚合和表示阶段是专门为了将(　　)进行聚合,从而获得一个统一的视图。

A. 关键数据集　　B. 离散数据

C. 单个数据集　　D. 多个数据集

13. 数据分析阶段致力于执行实际的分析任务,通常会涉及一种或多种类型的数据分析。在这个阶段,尤其是在探索性分析的情况下,分析过程会(　　)。

A. 重复进行,直到数据被清零

B. 循环进行,直到人为终止

C. 自然迭代,直到适当的模式或者相关性被发现

D. 一次完成,分析结果被打印和存储

14. 数据可视化阶段致力于由使用者使用(　　)技术和工具,并通过图形表示有效的分析结果。

A. 图形设计　　B. 数据可视化　　C. Photoshop　　D. 数字媒体

15. 大数据分析结果可以用来为商业使用者提供商业决策支持,为使用者提供更多使用这些分析结果的机会。分析结果的使用阶段致力于确定(　　)分析数据能保证产出更大的价值。

A. 如何以及在哪里处理　　B. 怎样以及什么时候

C. 是否以及怎样　　D. 如何打印以及存储