# 大数据架构和模式（一）: 大数据分类和架构简介

**1.本文对大数据做了哪些分类？**

**2.对数据进行分类后，如何将它与合适的大数据模式匹配?**

如何将大数据分为不同的类别

大数据问题的分析和解决通常很复杂。大数据的量、速度和种类使得提取信息和获得业务洞察变得很困难。以下操作是一个良好的开端：依据必须处理的数据的格式、要应用的分析类型、使用的处理技术，以及目标系统需要获取、加载、处理、分析和存储数据的数据源，对大数据问题进行分类。  
**概述**

大数据可通过许多方式来存储、获取、处理和分析。每个大数据来源都有不同的特征，包括数据的频率、量、速度、类型和真实性。处理并存储大数据时，会涉及到更多维度，比如治理、安全性和策略。选择一种架构并构建合适的大数据解决方案极具挑战，因为需要考虑非常多的因素。

这个 “大数据架构和模式” 系列提供了一种结构化和基于模式的方法来简化定义完整的大数据架构的任务。因为评估一个业务场景是否存在大数据问题很重要，所以我们包含了一些线索来帮助确定哪些业务问题适合采用大数据解决方案。  
**从分类大数据到选择大数据解决方案**

如果您花时间研究过大数据解决方案，那么您一定知道它不是一个简单的任务。本系列将介绍查找满足您需求的大数据解决方案所涉及的主要步骤。

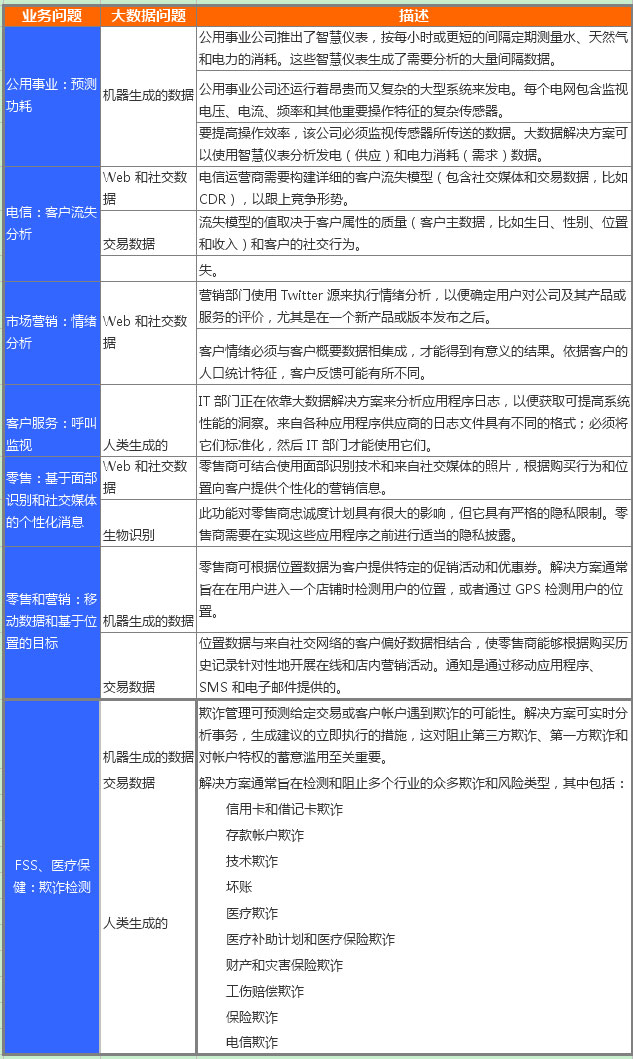
我们首先介绍术语 “大数据” 所描述的数据类型。为了简化各种大数据类型的复杂性，我们依据各种参数对大数据进行了分类，为任何大数据解决方案中涉及的各层和高级组件提供一个逻辑架构。接下来，我们通过定义原子和复合分类模式，提出一种结构来分类大数据业务问题。这些模式有助于确定要应用的合适的解决方案模式。我们提供了来自各行各业的示例业务问题。最后，对于每个组件和模式，我们给出了提供了相关功能的产品。

第 1 部分将介绍如何对大数据进行分类。本系列的后续文章将介绍以下主题：

* 定义大数据解决方案的各层和组件的逻辑架构
* 理解大数据解决方案的原子模式
* 理解用于大数据解决方案的复合（或混合）模式
* 为大数据解决方案选择一种解决方案模式
* 确定使用一个大数据解决方案解决一个业务问题的可行性
* 选择正确的产品来实现大数据解决方案

**依据大数据类型对业务问题进行分类**

业务问题可分类为不同的大数据问题类型。以后，我们将使用此类型确定合适的分类模式（原子或复合）和合适的大数据解决方案。但第一步是将业务问题映射到它的大数据类型。下表列出了常见的业务问题并为每个问题分配了一种大数据类型。  
按类型对大数据问题分类，更容易看到每种数据的特征。这些特征可帮助我们了解如何获取数据，如何将它处理为合适的格式，以及新数据出现的频率。来自不同来源的数据具有不同的特征；例如，社交媒体数据包含不断传入的视频、图像和非结构化文本（比如博客文章）。

表 1. 不同类型的大数据业务问题

我们依据这些常见特征来评估数据，下一节将详细介绍这些特征：

* 内容的格式
* 数据的类型（例如，交易数据、历史数据或主数据）
* 将提供该数据的频率
* 意图：数据需要如何处理（例如对数据的临时查询）
* 处理是否必须实时、近实时还是按批次执行。

**使用大数据类型对大数据特征进行分类**

按特定方向分析大数据的特征会有所帮助，例如以下特征：数据如何收集、分析和处理。对数据进行分类后，就可以将它与合适的大数据模式匹配：

1、分析类型 — 对数据执行实时分析还是批量分析。请仔细考虑分析类型的选择，因为这会影响一些有关产品、工具、硬件、数据源和预期的数据频率的其他决策。一些用例可能需要混合使用两种类型：

2、欺诈检测；分析必须实时或近实时地完成。

3、针对战略性业务决策的趋势分析；分析可采用批量模式。

4、处理方法 — 要应用来处理数据的技术类型（比如预测、分析、临时查询和报告）。业务需求确定了合适的处理方法。可结合使用各种技术。处理方法的选择，有助于识别要在您的大数据解决方案中使用的合适的工具和技术。

5、数据频率和大小 — 预计有多少数据和数据到达的频率多高。知道频率和大小，有助于确定存储机制、存储格式和所需的预处理工具。数据频率和大小依赖于数据源：

* 按需分析，与社交媒体数据一样
* 实时、持续提供（天气数据、交易数据）
* 时序（基于时间的数据）

6、数据类型 — 要处理数据类型 — 交易、历史、主数据等。知道数据类型，有助于将数据隔离在存储中。

7、内容格式（传入数据的格式）结构化（例如 RDMBS）、非结构化（例如音频、视频和图像）或半结构化。格式确定了需要如何处理传入的数据，这是选择工具、技术以及从业务角度定义解决方案的关键。

8、数据源 — 数据的来源（生成数据的地方），比如 Web 和社交媒体、机器生成、人类生成等。识别所有数据源有助于从业务角度识别数据范围。该图显示了使用最广泛的数据源。

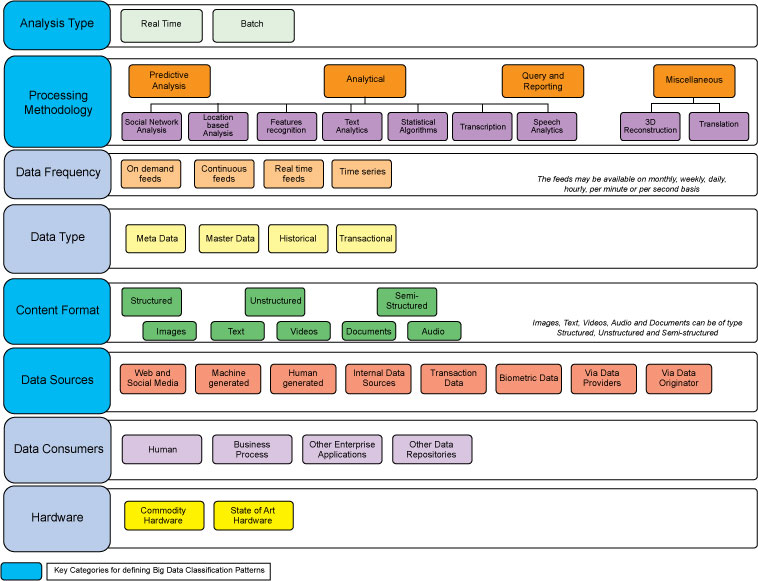
9、数据使用者 — 处理的数据的所有可能使用者的列表：

* 业务流程
* 业务用户
* 企业应用程序
* 各种业务角色中的各个人员
* 部分处理流程
* 其他数据存储库或企业应用程序

10、硬件 — 将在其上实现大数据解决方案的硬件类型，包括商用硬件或最先进的硬件。理解硬件的限制，有助于指导大数据解决方案的选择。

图 1 描绘用于分类大数据的各种类别。定义大数据模式的关键类别已识别并在蓝色方框中突出显示。大数据模式（将在下一篇文章中定义）来自这些类别的组合。

图 1. 大数据分类



**结束语和致谢**

在本系列剩余部分中，我们将介绍大数据解决方案的逻辑架构和各层，从访问到使用大数据。我们将提供数据源的完整列表，介绍专注于大数据解决方案的每个重要方面的原子模式。我们还将介绍复合模式，解释可如何结合使用原子模式来解决特定的大数据用例。本系列最后将提供一些解决方案模式，在广泛使用的用例与各个产品之间建立对应关系。

感谢 Rakesh R. Shinde 在定义本系列的整体结构上提供的指导，以及对本系列的审阅和提供的宝贵评论。

# 大数据架构和模式（二）如何知道一个大数据解决方案是否适合您的组织

**1.如何判断大数据问题是否需要大数据解决方案？**  
**2.如何评估大数据解决方案的可行性？**  
**3.可通过大数据技术获取何种洞察？**  
**4.是否所有大数据都存在大数据问题？**

**简介**

在确定投资大数据解决方案之前，评估可用于分析的数据；通过分析这些数据而获得的洞察；以及可用于定义、设计、创建和部署大数据平台的资源。询问正确的问题是一个不错的起点。使用本文中的问题将指导您完成调查。答案将揭示该数据和您尝试解决的问题的更多特征。

尽管组织一般情况对需要分析的数据类型有一些模糊的理解，但具体的细节很可能并不清晰。毕竟，数据可能具有之前未发现的模式的关键，一旦识别了一种模式，对额外分析的需求就会变得很明显。要帮助揭示这些未知的未知信息，首先需要实现一些基本用例，在此过程中，可以收集以前不可用的数据。构建数据存储库并收集更多数据后，数据科学家就能够更好地确定关键的数据，更好地构建将生成更多洞察的预测和统计模型。

组织可能也已知道它有哪些信息是不知道的。要解决这些已知的未知，组织首先必须与数据科学家合作，识别外部或第三方数据源，实现一些依赖于此外部数据的用例。

本文首先尝试回答大多数 CIO 在实施大数据举措之前通常会提出的问题，然后，本文将重点介绍一种将帮助评估大数据解决方案对组织的可行性的基于维度的方法。

**我的大数据问题是否需要大数据解决方案？**

**大数据，曾几何时似乎很少出现**

组织多半会选择以增量方式实现大数据解决方案。不是每个分析和报告需求都需要大数据解决方案。如果对于对大型数据集或来自多个数据源的临时报告执行并行处理的项目，那么可能没有必要使用大数据解决方案。

随着大数据技术的到来，组织会问自己：“大数据是否是我的业务问题的正确解决方案，或者它是否为我提供了业务机会？”大数据中是否隐藏着业务机会？以下是我从 CIO 那里听到的一些典型问题：

* 如果我使用大数据技术，可能会获得何种洞察和 业务价值？
* 它是否可以扩充我 现有的数据仓库？
* 我如何评估 扩展当前环境 或采用新解决方案的成本？
* 对我现有的 IT 治理 有何影响？
* 我能否 以增量方式实现 大数据解决方案？
* 我需要掌握哪些 具体的技能 来理解和分析构建和维护大数据解决方案的需求？
* 我的 现有企业数据 能否用于提供业务洞察？
* 来自各种来源的 数据的复杂性 在不断增长。大数据解决方案对我有帮助吗？

**维度可帮助评估大数据解决方案的可行性**

为了回答这些问题，本文提出了一种依据下图中所示的维度来评估大数据解决方案的可行性的结构化方法。



* 来自可通过分析数据获得的洞察的业务价值
* 针对新数据来源和数据使用方式的治理考虑因素
* 拥有相关技能和赞助商的承诺的人员
* 捕获的数据量
* 各种各样的数据源、数据类型和数据格式
* 生成数据的速度，需要对它执行操作的速度，或者它更改的速度
* 数据的真实性，或者数据的不确定性和可信赖性

对于每个维度，我们都给出了一些关键问题。依据业务上下文，为每个维度分配一个权重和优先级。评估会因业务案例和组织的不同而有所不同。您可以考虑在与相关的业务和 IT 利益相关者召开的一系列研讨会中探讨这些问题。

**业务价值：可通过大数据技术获取何种洞察？**

许多组织想知道，他们在寻找的业务洞察能否通过大数据解决方案解决。没有权威的指南能够用来定义可从大数据获取的洞察。具体场景需要由组织识别，而且这些场景在不断演变。在确定和识别在实现后会给企业带来重大价值的业务用例和场景的过程中，数据科学家起着至关重要的作用。

数据科学家必须能够理解关键绩效指标，对数据应用统计算法和复杂算法来获得一个用例列表。用例因行业和业务不同而有所不同。研究市场竞争对手的行动、发挥作用的市场力量，以及客户在寻找什么，会很有帮助。下表给出了来自各行各业的用例示例。

表 1. 来自各行各业的示例用例



潜在的客户正在社交网络和评论站点上生成大量新数据。在企业内，随着客户切换到在线渠道来执行业务和与公司交互，交易数据和 Web 日志与日俱增。

**确定数据的优先级**

首先为企业内存在的数据创建一个清单。识别内部系统和应用程序中存在的数据以及从第三方传入的数据。如果业务问题可使用现有数据解决，那么有可能不需要使用来自外部来源的数据。

请考虑构建一个大数据解决方案的成本，并权衡它与带给业务部门的新洞察的价值。

在有关现有客户的归档数据的上下文中分析此新数据时，业务人员将获得对新业务机会的洞察。

主要满足以下条件，大数据可提供可行的解决方案：

* 从数据中开发的洞察所生成的价值，值得在大数据解决方案中投入的资本成本
* 面向客户的场景可证明来自洞察的潜在价值

评估通过大数据解决方案获取的业务价值时，请考虑您当前的环境是否可扩展并权衡此投资的成本。

**我当前的环境能否扩展？**

询问以下问题，确定您能否扩充现有的数据仓库平台？

1. 当前的数据集是否非常大，是否达到了 TB 或 PB 数量级？
2. 现有的仓库环境是否包含生成或获取的所有 数据的存储库？
3. 是否有大量冷数据或人们很少接触的数据未分析，可以通过分析这些数据获得业务洞察？
4. 您是否需要丢弃数据，因为无法存储或处理它？
5. 您是否希望能够在复杂且大量的数据上执行数据探索？
6. 您是否希望能够对非操作数据执行分析？
7. 您是否有兴趣使用数据执行传统和新类型的分析？
8. 您是否试图延迟对现有数据仓库的升级？
9. 您是否在寻求途径降低执行分析的总体成本？

如果任何这些问题的答案是 “是”，那么您就可以探索扩充现有数据仓库环境的方式。

**扩展我当前的环境的成本是多少？**

扩展现有数据仓库平台或 IT 环境与实现大数据解决方案的成本和可行性取决于：

* 现有工具和技术
* 现有系统的可伸缩性
* 现有环境的处理能力
* 现有平台的存储能力
* 执行的治理和策略
* 现有 IT 应用程序的异构性
* 组织中存在的技术和业务技能。

它还依赖于将从新数据来源收集的数据量、业务用例的复杂性、处理的分析复杂性，以及获取数据和拥有恰当技能集的人员的成本。现有的资源池能否开发新的大数据技能，或者是否可从外部雇佣拥有稀缺技能的人员？

请注意，大数据举措会对其他正在实施的项目产生影响。从新的来源获取数据具有很高的成本。您首先应当识别系统和应用程序内部存在的数据，以及目前收到的第三方数据，这一点很重要。如果业务问题可以使用现有数据解决，那么有可能不需要使用来自外部来源的数据。

在生成新工具和应用程序之前，请评估组织的应用程序组合。例如，一个普通的 Hadoop 平台可能无法满足您的需求，您可能必须购买专业的工具。或者相对而言，Hadoop 的商业版本对当前用例而言可能很昂贵，但可能需要用作长期投资来支持一个战略性的大数据平台。考虑大数据工具和技术需要的基础架构、硬件、软件和维护的成本。

**对数据的治理和控制：对现有的 IT 治理有何影响？**

在决定是否实现一个大数据平台时，组织可能会查看新数据源和新的数据元素类型，而这些信息当前的所有权尚未明确定义。一些行业制度会约束组织获取和使用的数据。例如，在医疗行业，通过访问患者数据来从中获取洞察是否合法？类似的规则约束着所有行业。除了 IT 治理问题之外，组织的业务流程可能也需要重新定义和修改，让组织能够获取、存储和访问外部数据。

请在您的情况的上下文中考虑以下治理相关问题：

**安全性和隐私**— 为了与当地法规一致，解决方案可以访问哪些数据？可以存储哪些数据？哪些数据应在移动过程中加密？静止数据呢？谁可以查看原始数据和洞察？

**数据的标准化**— 是否有标准约束数据？数据是否具有专用的格式？是否有部分数据为非标准格式？

**数据可用的时段**— 数据在一个允许及时采取操作的时段是否可用？

**数据的所有权**— 谁拥有该数据？解决方案是否拥有适当的访问权和权限来使用数据？

**允许的用法**：允许如何使用该数据？

**我能否增量地实现大数据解决方案？**

大数据解决方案可以采用增量方式实现。明确地定义业务问题的范围，并以可度量的方式设置预期的业务收入提升，这样做会很有帮助。

对于基础业务案例，请仔细列出问题的范围和解决方案带来的预期收益。如果该范围太小，业务收益将无法实现，如果范围太大，获得资金和在恰当的期限内完成项目就会很有挑战性。在项目的第一次迭代中定义核心功能，以便能够轻松地赢得利益相关者的信任。

**人员：是否已有恰当的技能并调整了合适的人员？**

需要特定的技能来理解和分析需求，并维护大数据解决方案。这些技能包括行业知识、领域专长，以及有关大数据工具和技术的技术知识。拥有建模、统计、分析和数学方面的专业经验的数据科学家，是任何大数据举措成功的关键。

在实施一个新的大数据项目之前，确保已安排了合适的人员：

* 您是否获得利益相关者和其他愿意投资该项目的业务赞助者的支持？
* 是否拥有熟悉该领域、能分析大量数据、而且能识别从数据生成有意义且有用的洞察的途径的数据科学家？

**是否拥有可用于获取洞察的现有数据？**

所有组织都拥有大量未用于获取业务洞察的数据。这些数据包括日志文件、错误文件和来自应用程序的操作数据。不要忽略此数据，它是宝贵信息的潜在来源。

**数据复杂性是否在增长？**

查找数据复杂性增长的线索，尤其是在数据量、种类、速度和真实性方面。

**数据量是否已增长？**

如果满足以下条件，您可能希望考虑大数据解决方案：

* 数据大小达到 PB 和 EB 级，而且在不久的将来，它们可能增长到 ZB 级别。
* 这一数据量给使用传统方法（比如关系数据库引擎）存储、搜索、共享、分析和可视化数据带来的技术和经济挑战。
* 数据处理目前可使用可用硬件上的大规模并行处理能力。

**数据种类是否已增多？**

如果满足以下条件，各种各样的数据可能都需要大数据解决方案：

* 数据内容和结构无法预期或预测。
* 数据格式各不相同，包括结构化、半结构化和非结构化数据。
* 用户和机器能够以任何格式生成数据，例如：Microsoft® Word 文件、Microsoft Excel® 电子表格、Microsoft PowerPoint 演示文稿、PDF 文件、社交媒体、Web 和软件日志、电子邮件、来自相机的照片和视频、信息感知的移动设备、空中感知技术、基因组和医疗记录。
* 以前没有为了获得洞察而被挖掘的数据来源不断地在产生新的数据类型。
* 领域实体在不同的上下文中具有不同的含义。

**数据的速度是否已增长或改变？**

考虑您的数据是否：

* 在快速更改，必须立即响应
* 拥有过多的传统技术和方法，它们不再足以实时处理传入的数据

**您的数据是否值得信赖？**

如果满足以下条件，那么请考虑使用大数据解决方案：

* 数据的真实性或准确性未知。
* 数据包含模糊不清的信息。
* 不清楚数据是否完整。

如果数据的量、种类、速度或真实性具有合理的复杂性，那么有可能会适合地采用大数据解决方案。对于更复杂的数据，需要评估与实现大数据解决方案关联的任何风险。对于不太复杂的数据，则应该评估传统的解决方案。

**是否所有大数据都存在大数据问题？**

不是所有大数据情形都需要大数据解决方案。请在市场中寻找线索。竞争对手在做什么？哪些市场力量在发挥作用？客户想要什么？

使用本文中的问题，帮助确定大数据解决方案是否适合于您的业务情形和您需要的业务洞察。如果认为是时候实施大数据项目了，请阅读下一篇文章，其中会介绍如何定义一个逻辑架构，而且将会确定您的大数据解决方案需要的关键组件。

作者: Divakar等  来源: DeveloperWorks

End.

# 大数据架构和模式（三）理解大数据解决方案的架构层

**问题导读**

**1.大数据解决方案通常哪些逻辑层组成？**

**2.本文讲了大数据来源有哪些？**

**3.大数据治理包含哪些因素？**

**概述**

这个 “大数据架构和模式” 系列的 第 2 部分 介绍了一种评估大数据解决方案可行性的基于维度的方法。如果您已经使用上一篇文章中的问题和提示分析了自己的情况，并且已经决定开始构建新的（或更新现有的）大数据解决方案，那么下一步就是识别定义项目的大数据解决方案所需的组件。

**大数据解决方案的逻辑层**

逻辑层提供了一种组织您的组件的方式。这些层提供了一种方法来组织执行特定功能的组件。这些层只是逻辑层；这并不意味着支持每层的功能在独立的机器或独立的进程上运行。大数据解决方案通常由以下逻辑层组成：

1、大数据来源

2、数据改动 (massaging) 和存储层

3、分析层

4、使用层

**大数据来源：**考虑来自所有渠道的，所有可用于分析的数据。要求组织中的数据科学家阐明执行您需要的分析类型所需的数据。数据的格式和起源各不相同：

**格式— 结构化、半结构化或非结构化。**

**速度和数据量**— 数据到达的速度和传送它的速率因数据源不同而不同。

**收集点**— 收集数据的位置，直接或通过数据提供程序，实时或以批量模式收集数据。数据可能来自某个主要来源，比如天气条件，也有可能来自一个辅助来源，比如媒体赞助的天气频道。

**数据源的位置**— 数据源可能位于企业内或外部。识别您具有有限访问权的数据，因为对数据的访问会影响可用于分析的数据范围。

**数据改动和存储层**：此层负责从数据源获取数据，并在必要时，将它转换为适合数据分析方式的格式。例如，可能需要转换一幅图，才能将它存储在 Hadoop Distributed File System (HDFS) 存储或关系数据库管理系统 (RDBMS) 仓库中，以供进一步处理。合规性制度和治理策略要求为不同的数据类型提供合适的存储。

**分析层**：分析层读取数据改动和存储层整理 (digest) 的数据。在某些情况下，分析层直接从数据源访问数据。设计分析层需要认真地进行事先筹划和规划。必须制定如何管理以下任务的决策：

* 生成想要的分析
* 从数据中获取洞察
* 找到所需的实体
* 定位可提供这些实体的数据的数据源
* 理解执行分析需要哪些算法和工具。

**使用层**：

此层使用了分析层所提供的输出。使用者可以是可视化应用程序、人类、业务流程或服务。可视化分析层的结果可能具有挑战。有时，看看类似市场中的竞争对手是如何做的会有所帮助。

每一层包含多种组件类型，下面将会介绍这些类型。

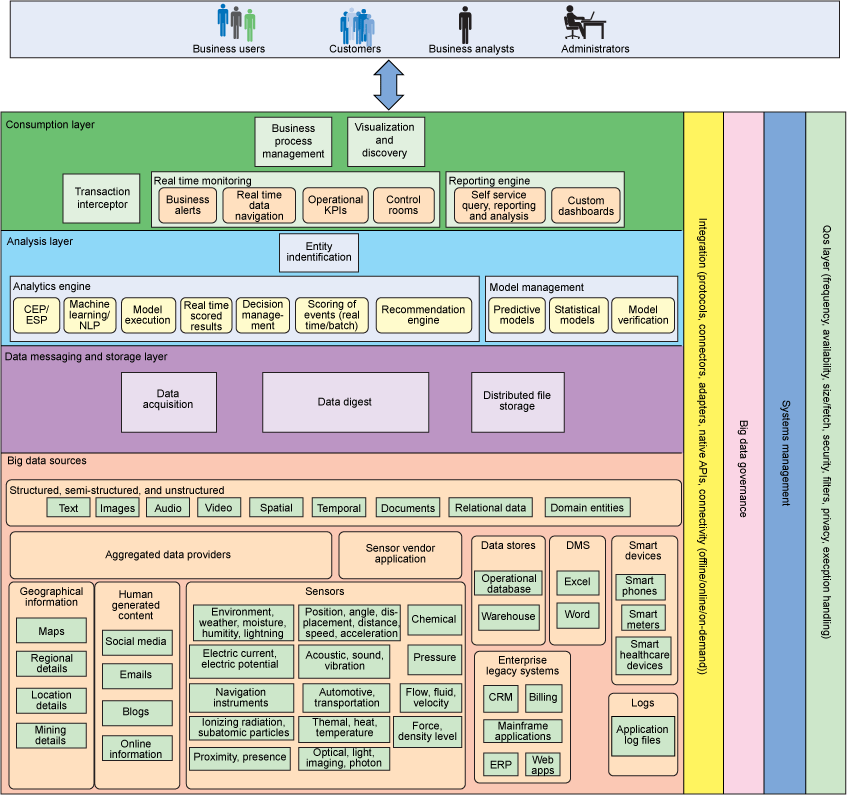


图 1. 逻辑和垂直层的组件

该图显示了逻辑和垂直层的组件

**大数据来源**

此层包含所有必要的数据源，提供了解决业务问题所需的洞察。数据是结构化、半结构化和非结构化的数据，而且来自许多来源：

**1、企业遗留系统—** 这些系统是企业应用程序，执行业务需要的分析并获取需要的洞察：

* 客户关系管理系统
* 结算操作
* 大型机应用程序
* 企业资源规划
* Web 应用程序开发

Web 应用程序和其他数据来源扩充了企业拥有的数据。这些应用程序可使用自定义的协议和机制来公开数据。

**2、数据管理系统 (DMS)**— 数据管理系统存储逻辑数据、流程、策略和各种其他类型的文档：

* Microsoft® Excel® 电子表格
* Microsoft Word 文档

这些文档可以转换为可用于分析的结构化数据。文档数据可公开为领域实体，或者数据改动和存储层可将它转换为领域实体。

**3、数据存储**— 数据存储包含企业数据仓库、操作数据库和事务数据库。此数据通常是结构化数据，可直接使用或轻松地转换来满足需求。这些数据不一定存储在分布式文件系统中，具体依赖于所处的上下文。

**4、智慧设备**— 智慧设备能够捕获、处理和传输使用最广泛的协议和格式的信息。这方面的示例包括智能电话、仪表和医疗设备。这些设备可用于执行各种类型的分析。绝大多数智慧设备都会执行实时分析，但从智慧设备传来的信息也可批量分析。

**5、聚合的数据提供程序**— 这些提供程序拥有或获取数据，并以复杂的格式和所需的频率通过特定的过滤器公开它。每天都会产生海量的数据，它们具有不同的格式，以不同的速度生成，而且通过各种数据提供程序、传感器和现有企业提供。

其他数据源— 有许多数据来自自动化的来源：

**地理信息：**

* 地图
* 地区详细信息
* 位置详细信息
* 矿井详细信息

**人类生成的内容：**

* 社交媒体
* 电子邮件
* 博客
* 在线信息

**传感器数据：**

* 环境：天气、降雨量、湿度、光线
* 电气：电流、能源潜力等
* 导航装置
* 电离辐射、亚原子粒子等
* 靠近、存在等
* 位置、角度、位移、距离、速度、加速度
* 声音、声震动等
* 汽车、运输等
* 热量、热度、温度
* 光学、光、成像、见光度
* 化学
* 压力
* 流动、流体、速度
* 力、密度级别等
* 来自传感器供应商的其他数据

**数据改动和存储层**

因为传入的数据可能具有不同的特征，所以数据改动和存储层中的组件必须能够以各种频率、格式、大小和在各种通信渠道上读取数据：

数据获取— 从各种数据源获取数据，并将其发送到数据整理组件或存储在指定的位置中。此组件必须足够智能，能够选择是否和在何处存储传入的数据。它必须能够确定数据在存储前是否应改动，或者数据是否可直接发送到业务分析层。

数据整理— 负责将数据修改为需要的格式，以实现分析用途。此组件可拥有简单的转换逻辑或复杂的统计算法来转换源数据。分析引擎将会确定所需的特定的数据格式。主要的挑战是容纳非结构化数据格式，比如图像、音频、视频和其他二进制格式。

分布式数据存储— 负责存储来自数据源的数据。通常，这一层中提供了多个数据存储选项，比如分布式文件存储 (DFS)、云、结构化数据源、NoSQL 等。

**分析层**

这是从数据中提取业务洞察的层：

分析层实体识别— 负责识别和填充上下文实体。这是一个复杂的任务，需要高效的高性能流程。数据整理组件应为这个实体识别组件提供补充，将数据修改为需要的格式。分析引擎将需要上下文实体来执行分析。

分析引擎— 使用其他组件（具体来讲，包括实体鉴别、模型管理和分析算法）来处理和执行分析。分析引擎可具有支持并行处理的各种不同的工作流、算法和工具。

模型管理— 负责维护各种统计模型，验证和检验这些模型，通过持续培训模型来提高准确性。然后，模型管理组件会推广这些模型，它们可供实体识别或分析引擎组件使用。

**使用层**

这一层使用了从分析应用程序获取的业务洞察。分析的结果由组织内的各个用户和组织外部的实体（比如客户、供应商、合作伙伴和提供商）使用。此洞察可用于针对客户提供产品营销信息。例如，借助从分析中获取的洞察，公司可以使用客户偏好数据和位置感知，在客户经过通道或店铺时向他们提供个性化的营销信息。

该洞察可用于检测欺诈，实时拦截交易，并将它们与使用已存储在企业中的数据构建的视图进行关联。在欺诈性交易发生时，可以告知客户可能存在欺诈，以便及时采取更正操作。

此外，可以根据在数据改动层完成的分析来触发业务流程。可以启动自动化的步骤 — 例如，如果客户接受了一条可自动触发的营销信息，则需要创建一个新订单，如果客户报告了欺诈，那么可以触发对信用卡使用的阻止。

分析的输出也可由推荐引擎使用，该引擎可将客户与他们喜欢的产品相匹配。推荐引擎分析可用的信息，并提供个性化且实时的推荐。

使用层还为内部用户提供了理解、找到和导航企业内外的链锁信息的能力。对于内部使用者，为业务用户构建报告和仪表板的能力使得利益相关者能够制定精明的决策并设计恰当的战略。为了提高操作有效性，可以从数据中生成实时业务警告，而且可以监视操作性的关键绩效指标：

**交易拦截器**— 此组件可实时拦截高容量交易，将它们转换为一种容易被分析层理解的实时格式，以便在传入数据上执行实时分析。事务拦截器应能够集成并处理来自各种来源的数据，比如传感器、智能仪表、麦克风、摄像头、GPS 设备、ATM 和图像扫描仪。可以使用各种类型的适配器和 API 来连接到数据源。也可以使用各种加速器来简化开发，比如实时优化和流分析，视频分析，银行、保险、零售、电信和公共运输领域的加速器，社交媒体分析，以及情绪分析。

**业务流程管理流程**— 来自分析层的洞察可供业务流程执行语言 (BPEL) 流程、API 或其他业务流程使用，通过自动化上游和下游 IT 应用程序、人员和流程的功能，进一步获取业务价值。

**实时监视**— 可以使用从分析中得出的数据来生成实时警告。可以将警告发送给感兴趣的使用者和设备，比如智能电话和平板电脑。可以使用从分析组件生成的数据洞察，定义并监视关键绩效指标，以便确定操作有效性。实时数据可从各种来源以仪表板的形式向业务用户公开，以便监视系统的健康或度量营销活动的有效性。

**报告引擎**— 生成与传统商业智能报告类似的报告的能力至关重要。用户可基于从分析层中得到的洞察，创建临时报告、计划的报告或自助查询和分析。

**推荐引擎**— 基于来自分析层的分析结果，推荐引擎可向购物者提供实时的、相关的和个性化的推荐，提高电子商务交易中的转换率和每个订单的平均价值。该引擎实时处理可用信息并动态地响应每个用户，响应基于用户的实时活动、存储在 CRM 系统中的注册客户信息，以及非注册客户的社交概况。

**可视化和发现**— 数据可跨企业内外的各种联邦的数据源进行导航。数据可能具有不同的内容和格式，所有数据（结构化、半结构化和非结构化）可组合来进行可视化并提供给用户。此能力使得组织能够将其传统的企业内容（包含在企业内容管理系统和数据仓库中）与新的社交内容（例如 tweet 和博客文章）组合到单个用户界面中。

**垂直层**

影响逻辑层（大数据来源、数据改动和存储、分析和使用层）的所有组件的各方面都包含在垂直层中：

* 信息集成
* 大数据治理
* 系统管理
* 服务质量

**信息集成**

大数据应用程序从各种数据起源、提供程序和数据源获取数据，并存储在 HDFS、NoSQL 和 MongoDB 等数据存储系统中。这个垂直层可供各种组件使用（例如数据获取、数据整理、模型管理和交易拦截器），负责连接到各种数据源。集成将具有不同特征（例如协议和连接性）的数据源的信息，需要高质量的连接器和适配器。可以使用加速器连接到大多数已知和广泛使用的来源。这些加速器包括社交媒体适配器和天气数据适配器。各种组件还可以使用这一层在大数据存储中存储信息，从大数据存储中检索信息，以便处理这些信息。大多数大数据存储都提供了服务和 API 来存储和检索该信息。

**大数据治理**

数据治理涉及到定义指南来帮助企业制定有关数据的正确决策。大数据治理有助于处理企业内或从外部来源传入的数据的复杂性、量和种类。在将数据传入企业进行处理、存储、分析和清除或归档时，需要强有力的指南和流程来监视、构建、存储和保护数据。

除了正常的数据治理考虑因素之外，大数据治理还包含其他因素：

1、管理各种格式的大量数据。

2、持续培训和管理必要的统计模型，以便对非结构化数据和分析进行预处理。请记住，设置处理非结构化数据时的重要一步。

3、为外部数据设置有关其保留和使用的策略和合规性制度。

4、定义数据归档和清除策略。

5、创建如何跨各种系统复制数据的策略。

6、设置数据加密策略。

**服务质量层**

此层复杂定义数据质量、围绕隐私和安全性的策略、数据频率、每次抓取的数据大小和数据过滤器：

数据质量

1、完整地识别所有必要的数据元素

2、以可接受的新鲜度提供数据的时间轴

3、依照数据准确性规则来验证数据的准确性

4、采用一种通用语言（数据元组满足使用简单业务语言所表达的需求）

5、依据数据一致性规则验证来自多个系统的数据一致性

6、在满足数据规范和信息架构指南基础上的技术符合性

**围绕隐私和安全的策略**

需要策略来保护敏感数据。从外部机构和提供程序获取的数据可能包含敏感数据（比如 Facebook 用户的联系信息或产品定价信息）。数据可以来源于不同的地区和国家，但必须进行相应的处理。必须制定有关数据屏蔽和这类数据的存储的决策。考虑以下数据访问策略：

A、数据可用性

B、数据关键性

C、数据真实性

D、数据共享和发布

E、数据存储和保留，包括能否存储外部数据等问题。如果能够存储数据，数据可存储多长时间？可存储何种类型的数据？

F、数据提供程序约束（政策、技术和地区）

G、社交媒体使用条款

**数据频率**

提供新鲜数据的频率是多少？它是按需、连续还是离线的？

**抓取的数据大小**

此属性有助于定义可抓取的数据以及每次抓取后可使用的数据大小。

**过滤器**

标准过滤器会删除不想要的数据和数据中的干扰数据，仅留下分析所需的数据。

**系统管理**

系统管理对大数据至关重要，因为它涉及到跨企业集群和边界的许多系统。对整个大数据生态系统的健康的监视包括：

A、管理系统日志、虚拟机、应用程序和其他设备

B、关联各种日志，帮助调查和监视具体情形

C、监视实时警告和通知

D、使用显示各种参数的实时仪表板

E、引用有关系统的报告和详细分析

F、设定和遵守服务水平协议

G、管理存储和容量

G、归档和管理归档检索

I、执行系统恢复、集群管理和网络管理

J、策略管理

**结束语**

对开发人员而言，层提供了一种对大数据解决方案必须执行的功能进行分类的途径，为组织建议必需执行这些功能所需的代码。但是，对于想要从大数据获取洞察的业务用户，考虑大数据需求和范围通常会有所帮助。原子模式解决了访问、处理、存储和使用大数据的机制，为业务用户提供了一种解决需求和范围的途径。下一篇文章将介绍用于此用途的原子模式。

作者: Divakar等  来源: DeveloperWorks

End.

# 大数据架构和模式（四）大数据解决方案的原子模式和复合模式

**1.原子模式大数据解决方案有什么特点？  
2.原子模式下的样例模式有哪些？  
3.复合模式有什么特点？**

**简介**

本系列的 第 3 部分 介绍了大数据解决方案的逻辑层。这些层定义了各种组件，并对它们进行分类，这些组件必须处理某个给定业务用例的功能性和非功能性需求。本文基于层和组件的概念，介绍了解决方案中所用的典型原子模式和复合模式。通过将所提出的解决方案映射到此处提供的模式，让用户了解需要如何设计组件，以及从功能角度考虑，应该将它们放置在何处。模式有助于定义大数据解决方案的架构。利用原子模式和复合模式可以帮助进一步完善大数据解决方案的每个组件的角色和责任。

本文介绍原子模式和复合模式。本系列的最后一篇文章将介绍解决方案模式。

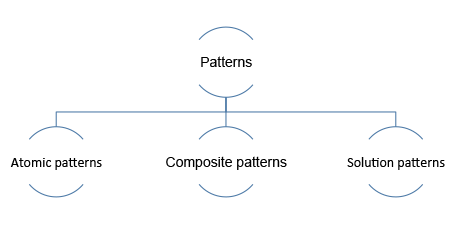


图 1. 模式的类别

该图显示了一下模式：原子模式、复合模式和解决方案模式

**原子模式**

对于大数据上下文中经常出现的问题，原子模式 有助于识别数据如何是被使用、处理、存储和访问的。它们还有助于识别所需的组件。访问、存储和处理来自不同数据源的多种数据需要不同的方法。每种模式都用于满足特定的需求：例如，可视化、历史数据分析、社交媒体数据和非结构化数据的存储。可以将多种原子模式结合使用，组成一个复合模式。这些原子模式没有进行分层或排序。例如，可视化模式可以与社交媒体的数据访问模式直接交互，可视化模式还可以与高级分析处理模式进行交互。

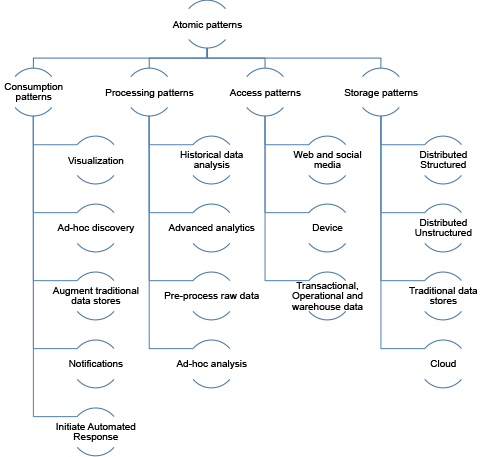


图 2. 用于数据使用、处理、数据访问和存储的原子模式的示例

该图显示了每种原子模式下的样例模式的列表

**数据使用模式**

这种类型的模式处理使用数据分析结果的各种方式。本节包括的数据使用模式可以满足几个需求。

**可视化模式**

可视化数据的传统方式以图表、仪表板和摘要报告为基础。这些传统的方法并不总是用来可视化数据的最佳方式。

大数据可视化的典型需求（包括新出现的需求）如下所示：

1、执行流数据的实时分析和显示

2、基于上下文，以交互方式挖掘数据

3、执行高级搜索，并获得建议

4、并行可视化信息

5、获得先进的硬件，支持未来的可视化需求

正在进行研究，以确定人类和机器如何使用大数据洞察。这些挑战包括所涉及的数据量，并且需要将数据与上下文相关联。必须在适当的上下文中显示洞察。

可视化数据的目的是为了更容易、更直观地使用数据，因此报告和仪表板可能提供全高清的观看效果和 3-D 互动视频，并且可以为用户提供使用应用程序控制业务活动和结果的能力。

**即席发现模式**

创建满足所有业务需求的标准报告往往是不可行的，因为企业的业务数据查询会有不同的需求。用户在查找特定信息时，可能需要获得根据问题的上下文执行即席查询的能力。

即席分析可以帮助数据科学家和关键业务用户了解业务数据的行为。即席处理中涉及的复杂性来自多种因素：

1、多个数据源可用于相同的域。

2、单一的查询可以有多个结果。

3、输出可以是静态的，并具有多种格式（视频、音频、图形和文本）。

4、输出可以是动态和交互式的。

**加强传统的数据存储**

在大数据的初步探索中，许多企业选择使用现有的分析平台来降低成本，并依赖于现有的技能。加强现有的数据存储有助于拓宽可用于现有分析的数据的范围，包括驻留在组织边界内外的数据，比如社交媒体数据，它可以丰富主数据。通过拓宽数据范围，使之包含现有存储中的新事实表、维度和主数据，并从社交媒体获取客户数据，组织可以获得更深入的客户洞察。

但要牢记的是，新的数据集通常比较大，而现有的提取、转换和加载工具可能不足以处理它。您可能需要使用具有大规模并行处理能力的高级工具来解决数据的数量、多样性、真实性和速度特征。

**通知模式**

大数据洞察使人类、企业和机器可以通过使用事件通知而立即采取行动。通知平台必须能够处理及时发送出去的预计数量的通知。这些通知与大量邮件或群发短信不同，因为内容一般是特定于使用者的。例如，推荐引擎可以提供有关世界各地的庞大客户群的洞察，而且可以将通知发送给这样的客户。

**启动一个自动响应模式**

从大数据获得的业务洞察，可用于触发或启动其他业务流程或事务。

**处理模式**

无论数据是处于静止状态还是在运动中，都可以处理大数据。具体情况取决于分析的复杂性，有可能不需要对数据进行实时处理。这种模式解决了对大数据进行实时、近实时或批量处理的方式。

以下高级的大数据处理类别适用于大多数分析。这些类别通常也适用于基于 RDBMS 的传统系统。惟一的区别是庞大规模的数据、多样性和速度。在处理大数据时，要使用机器学习、复杂事件处理、事件流处理、决策管理和统计模型管理等技术。

**历史数据分析模式**

传统的历史数据分析仅限于预定义的数据时间段，这通常取决于数据保留策略。由于处理和存储的限制，超出此时间段的数据通常会被归档或清除。基于 Hadoop 的系统和其他等效的系统可以克服这些限制，因为它们具有丰富的存储以及分布式大规模并行处理能力。运营、业务和数据仓库的数据被移动到大数据存储，您通过使用大数据平台功能对它们进行处理。

历史分析包括分析给定时间段、季节组合和产品的历史趋势，并与最新的可用数据进行比较。为了能够存储和处理如此庞大的数据，您可以使用 HDFS、NoSQL、SPSS® 和 InfoSphere® BigInsights™。

**高级分析模式**

大数据提供了很多实现创意洞察的机会。不同的数据集可以在多种上下文中存在关联。发现这些关系需要创新的复杂算法和技术。

高级分析包括预测、决策、推理过程、模拟、上下文信息标识和实体解析。高级分析的应用包括生物统计数据分析（例如，DNA 分析）、空间分析、基于位置的分析、科学分析、研究，等等。高级分析要求大量的计算来管理大量的数据。

数据科学家可以指导您识别合适的技术、算法和数据集，以及在给定上下文中解决问题所需的数据源。比如 SPSS、InfoSphere Streams 和 InfoSphere BigInsights 等工具提供了这类功能。这些工具访问存储在大数据存储系统（比如 BigTable、HBase，等等）中的非结构化数据和结构化数据（例如，JSON 数据）。

**预处理原始数据模式**

大数据解决方案主要由基于 MapReduce 的 Hadoop 系统和技术组成，MapReduce 是开箱即用的分布式存储和处理解决方案。然而，从非结构化数据提取数据（例如，图像、音频、视频、二进制提要，甚至是文本）是一项复杂的任务，需要具有机器学习能力并掌握自然语言处理等技术。另一个主要挑战是如何验证这些技术和算法的输出的准确度和正确性。

要对任何数据执行分析，数据都必须是某种结构化格式。从多个数据源访问的非结构化数据可以按原样存储，然后被转化成结构化数据（例如 JSON），并被再次存储到大数据存储系统中。非结构化文本可以转换成半结构化或结构化数据。同样，图像、音频和视频数据需要转换成可用于分析的格式。此外，使用预测和统计算法的高级分析的准确性和正确性取决于用来训练其模型的数据和算法的数量。

下面的列表显示了将非结构化数据转换成结构化数据所需的算法和活动：

1、文档和文本分类

2、特征提取

3、图像和文本分割

4、关联特征、变量和时间，然后提取包含时间的值

5、输出的准确度检查使用了混淆矩阵（confusion matrix）等技术和其他手动活动

数据科学家可以帮助用户选择合适的技术和算法。

**即席分析模式**

处理大数据的即席查询所带来的挑战不同于对结构化数据执行即席查询时所面临的挑战，由于数据源和数据格式不是固定的，所以需要使用不同的机制来检索和处理数据。

虽然大数据供应商可以处理简单的即席查询，但在大多数情况下，查询是复杂的，因为必须在运行时动态地发现数据、算法、格式和实体解析。所以需要利用数据科学家和业务用户的专业知识来定义下列任务所需的分析：

1、识别并发现计算和算法

2、识别并发现数据源

3、定义所需的可以由计算使用的格式

4、对数据执行并行计算

**访问模式**

在大数据解决方案中，有许多数据源，还有很多访问数据的方式，本节将介绍最常见的几种。

**Web 和社交媒体访问模式**

Internet 是提供许多目前可以获得的洞察的数据源。在几乎所有分析中，都会用到 Web 和社交媒体，但获得这种数据需要不同的访问机制。

在所有数据源中，因为 Web 和社交媒体的多样性、速度和数量，所以 Web 和社交媒体是最为复杂的。网站大约有 40-50 个类别，每一个类别都需要使用不同的方式来访问数据。本节将列出这些类别，并介绍一些访问机制。从大数据的角度讲，高级的类别是商业站点、社交媒体站点，以及具有特定和通用组件的站点。有关的访问机制见图 3。如果需要的话，在完成预处理后，可将所访问的数据存储在数据存储中。

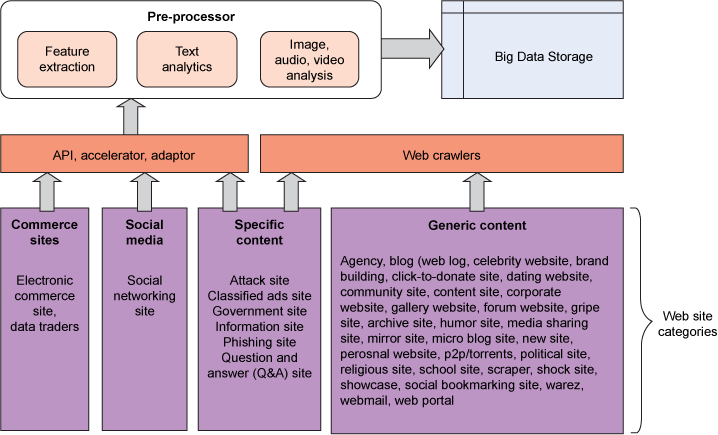


图 3. Web 和社交媒体访问

该图显示了 Web 和社交媒体数据访问

需要执行以下步骤来访问 Web 媒体信息。

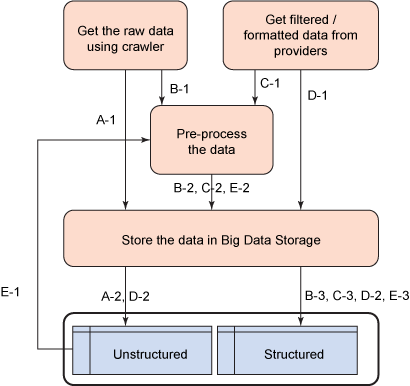


图 4. 大数据访问步骤

**非结构化数据存储中的 Web 媒体访问**

步骤 A-1. 爬网程序读取原始数据。

步骤 A-2. 数据被存储在非结构化存储中。

**Web 媒体访问为结构化存储预处理数据**

步骤 B-1. 爬网程序读取原始数据。

步骤 B-2. 对数据进行预处理。

步骤 B-3. 数据被存储在结构化存储中。

**Web 媒体访问预处理非结构化数据**

步骤 C-1. 在极少数情况下，来自供应商的数据可以是非结构化数据。

步骤 C-2. 对数据进行预处理。

步骤 C-3. 数据被存储在结构化存储中。

**非结构化或结构化数据的 Web 媒体访问**

步骤 D-1. 数据供应商提供结构化或非结构化数据。

步骤 D-2. 数据被存储在结构化或非结构化存储中。

**Web 媒体访问预处理非结构化数据**

步骤 E-1. 不能使用在存储时未经过预处理的非结构化数据，除非它是结构化格式的数据。

步骤 E-2. 对数据进行预处理。

步骤 E-3. 经过预处理的结构化数据被存储在结构化存储中。

如图所示，数据可以直接存储在存储器中，或者可以对它们进行预处理，并将它们转换成一个中间格式或标准格式，然后再存储它们。

在可以分析数据之前，数据格式必须可用于实体解析或用于查询所需数据。这种经过预处理的数据可以存储在一个存储系统中。

虽然预处理通常被认为是微不足道的，但这项处理可能非常复杂和耗时。

**设备生成的数据模式**

设备生成的内容包括来自传感器的数据数据是从天气信息、电气仪表和污染数据等数据来源检测到的，并且由传感器捕获。这些数据可以是照片、视频、文本和其他二进制格式。

下图说明了处理机器生成的数据的典型过程。

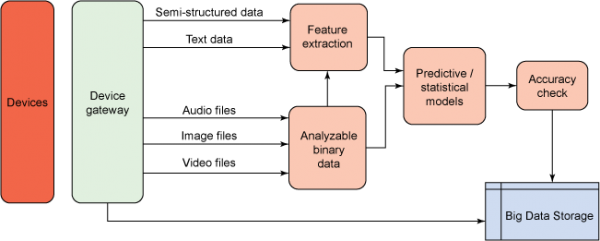


图 5. 设备生成的数据访问

该图显示了设备生产的数据已被处理

图 5 说明了访问来自传感器的数据的过程。由传感器捕获的数据可以发送到设备网关，设备网关会对数据执行一些初始预处理，并缓冲高速数据。机器生成的数据大多为二进制格式（音频、视频和传感器读数）或文本格式。这样的数据最初可以存储在存储系统中，也可以对它们进行预处理，然后再存储它们。对于分析来说，要求执行预处理。

**事务、运营和仓库数据模式**

可以存储现有的事务、运营和仓库数据，避免清除或归档数据（因为存储和处理的限制），或减少在数据被其他使用者访问时对传统存储的负载。

对于大多数企业而言，事务、运营、主数据和仓库信息都是所有分析的核心。如果用在 Internet 上，或者通过传感器和智能设备提供的非结构化数据以及外部数据来增强此数据，那么可以帮助组织获得准确的洞察，并执行高级分析。

使用由多个数据库厂商提供的标准连接器，事务和仓库数据可以被推入存储。预处理事务性数据要容易得多，因为数据大多是结构化的。可以使用简单的提取、转换和加载流程将事务数据移动到存储中。事务数据可以很容易地转换成 JSON 和 CSV 等格式。使用 Sqoop 等工具可以更容易将事务数据推入存储系统，如 HBase 和 HDFS。

**数据访问模式的特殊变化：生物特征数据访问**

此信息的数据访问与对机器生成的数据的访问非常相似。生物特征数据被归类为生理和行为数据，可以通过许多方式对大量数据进行分析。

有些数据可通过传感器来获取，有些数据则需要身体样品（血液、尿液等）。处理生物特征数据（如 DNA 数据）需要更长的时间。

生理数据包括指纹、掌纹、气味和香味的信息，以及面部、声音、视网膜和虹膜特征。行为数据包括打字模式、打字节奏、说话、走路、签名匹配和步态。

**存储模式**

存储模式有助于确定适当的存储各种数据的类型和格式。数据可以按原样存储，根据键值对存储，或者以预定义的格式存储。

分布式文件系统（如 GFS 和 HDFS）都能够存储任何类型的数据。但是，高效地检索或查询数据的能力会影响性能。技术的选择很重要。

**分布式非结构化数据的存储模式**

大部分大数据是非结构化数据，而且可以通过不同的方式针对不同的上下文提取它所拥有的信息。大多数时候，非结构化数据必须按原样并以其原始格式进行存储。

这样的数据可以存储在分布式文件系统（如 HDFS）和 NoSQL 文档存储（如 MongoDB）中。这些系统提供了检索非结构化数据的有效方法。

**分布式结构化数据的存储模式**

结构化数据包括从数据源到达的已经是结构化格式的数据，以及经过预处理，被转换为 JSON 数据等格式的非结构化数据。必须存储已经过转换的数据，避免从原始数据到结构化数据的频繁数据转换。

可以使用 Google 的 BigTable 等技术来存储结构化数据。BigTable 是一个大规模容错式自我管理系统，包括 TB 级的内存和 PB 级的存储。

Hadoop 中的 HBase 可媲美 BigTable。它使用了 HDFS 作为底层存储。

**传统数据存储的存储模式**

对于存储大数据而言，传统的数据存储并不是最佳选择，但在企业执行初步数据探索的情况下，企业可能会选择使用现有的数据仓库、RDBMS 系统和其他内容存储。这些现有的存​​储系统可用来存储使用大数据平台消化和过滤的数据。不要认为传统的数据存储系统适用于大数据。

**云存储的存储模式**

许多云计算基础架构供应商都有分布式结构化、非结构化的存储能力。从传统的配置、维护、系统管理、编程和建模角度讲，大数据技术有点不同。此外，实现大数据解决方案所需的技能既罕见又昂贵。探索大数据技术的企业可以使用云解决方案来提供大数据的存储、维护和系统管理。

要存储的数据往往是敏感数据，这些数据包括医疗记录和生物特征数据。您需要考虑数据安全性、数据共享、数据治理，以及有关数据的其他政策，在考虑将云作为大数据存储库的时候尤其如此。传输大量数据的能力也是云存储的另一个重要考虑因素。

**复合模式**

原子模式 侧重于提供执行各项功能所需的能力。但是，复合模式 是基于端到端的解决方案进行分类的。每个复合模式都要考虑一个或多个维度。在将复合模式应用到每个模式时，会有许多变化。可以将复合模式映射到一个或多个原子模式，以解决某个给定的业务问题。本文所述的复合模式列表是基于经常发生的典型业务问题，但这不是复合模式的完整列表。

**存储和探索模式**

如果业务问题需要存储大量新数据和现有数据，而且先前由于缺乏足够的存储和分析能力而一直未使用这些数据，那么这种模式就非常有用。该模式旨在缓解对现有数据存储的负载。所存储的数据可用于初始勘探和即席发现。用户可以推演报告，通过进一步的处理来分析数据的质量和价值。您可以使用 ETL 工具来预处理和净化原始数据，然后再进行任何类型的分析。

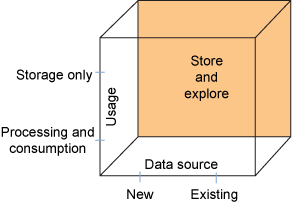


图 6. 存储和探索复合模式

该图显示了模式的维度包括仅存储、处理和使用

图 6 说明了这种模式的多个维度。数据的使用目的可能只是存储它，或处理和使用它。

仅存储的示例是，数据的获取和存储只是为了将来能够满足合规性或法律的要求。在处理和使用的情况下，分析的结果可以被处理和使用。可以从最近发现的来源或从现有的数据存储访问数据。

**专用和可预测的分析复合模式**

使用此模式的情况是，使用多种处理技术执行分析，因此，可以用新洞察丰富现有数据，或创建可由各种用户使用的输出。该分析可以在事件发生的同时实时发生，或使用批量模式，根据收集到的数据获得洞察。作为可以分析的静态数据的示例，某电信公司可能构建客户流失模型，包括分析呼叫数据记录、社交数据和事务数据。作为分析运动数据的示例，预测某个给定事务正在经历欺诈的需求必须实时或近实时地发生。

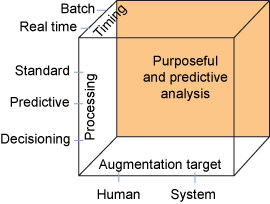


图 7. 专用和预测分析复合模式

该图显示了专用和预测分析复合模式

图 7 说明了这种模式的多个维度。所执行的处理可以是标准的或预测性的，并且可以包括决策。

此外，可以将通知发送给与特定任务或消息有关的系统或用户。该通知可以使用可视化功能。该处理可实时发生或以批量模式发生。

**可操作的分析模式**

大数据解决方案的最高级形式是，对数据集执行分析，并且基于可重复的过去的行动或行动矩阵来暗示行动。该操作可以是手动、半自动或全自动的。基础分析需要高度准确。行动是预定义的，分析的结果被映射到行动。可操作分析中所涉及的典型步骤是：

1、分析数据以获得洞察。

2、制定决策。

3、激活相应的渠道，对正确的使用者采取行动。

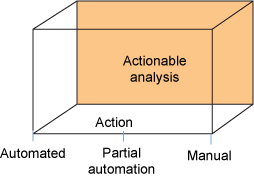


图 8. 可操作的分析复合模式

该图显示了可操作的分析复合模式

图 8 说明该分析可以是手动、半自动或全自动的。如图中的说明所示，它使用了原子模式。

手动操作 意味着系统基于分析的结果来提供建议操作，并由人类决定和执行操作。半自动 意味着，分析建议操作，但不需要通过人类干预来启动操作，或从一组建议的操作中进行选择。全自动 表示在决策之后，系统立即执行操作。例如，在设备被预测会发生故障之后，系统可以自动创建一个工作订单。

下面的矩阵显示了如何将原子模式映射到复合模式，复合模式是原子模式的组合。每个复合模式都被设计为针对具有一组特定特征的数据在特定情况下使用。矩阵显示了模式的典型组合。必须对模式进行调整，以满足特定的情况和需求。在矩阵中，按照从最简单到最复杂的顺序列出了复合模式。“store and explore（存储和探索）”模式是最简单的。

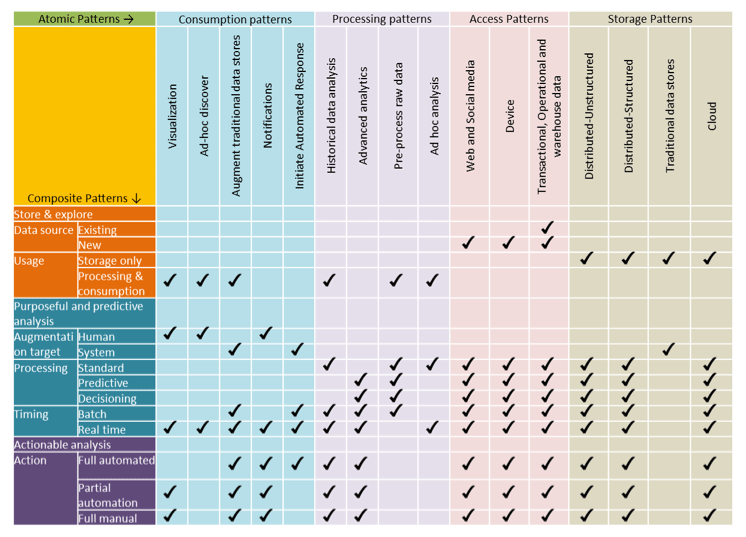


图 9. 复合模式对原子模式的映射

该图显示了复合模式对原子模式的映射

**结束语**

采用基于模式的方法可以帮助业务团队和技术团队在解决方案的首要目标上达成一致意见。技术团队可以使用模式定义架构性原则，并制定一些关键架构决策。技术团队可以将这些模式应用到架构层，并导出实现解决方案所需的组件集。通常情况下，解决方案从有限的范围开始，然后企业会变得越来越有信心，相信解决方案会带来价值。随着演变的发生，与解决方案一致的复合模式与原子模式将会得到细化。在初始阶段，可以使用模式来定义一个基于模式的架构，并映射出在该架构中如何逐步设计组件。

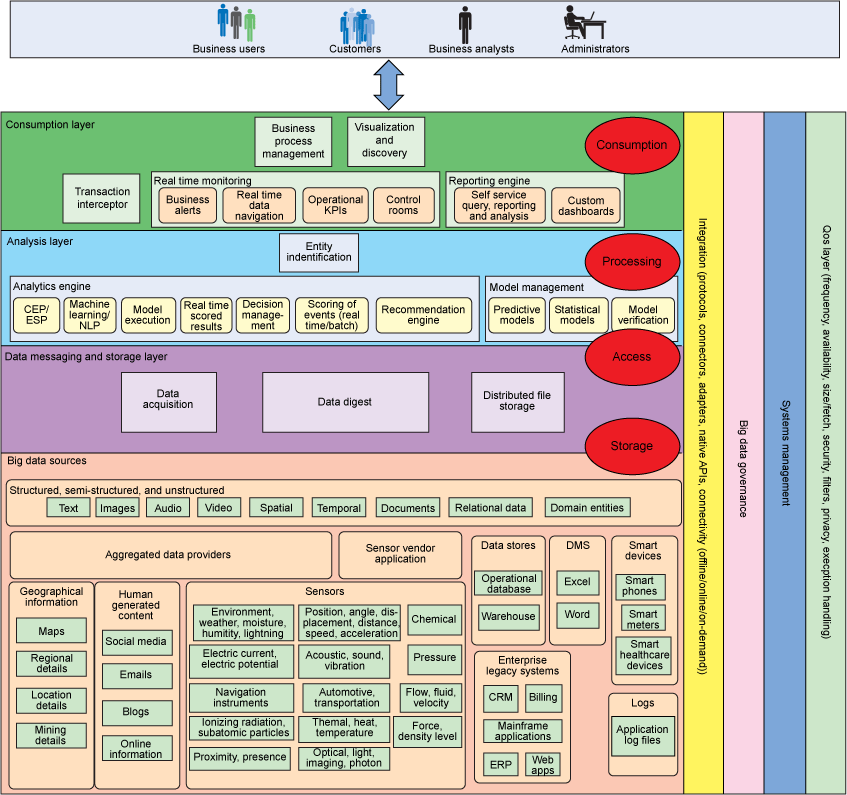


图 10. 将原子模式映射到架构层

该图显示了原子模式对架构层的映射

在本系列的 第 2 部分 中，我们描述了与大数据有关的复杂性，以及如何确定是否应实现或更新您的大数据解决方案。在本文中，我们讨论了原子模式和复合模式，并解释了一个解决方案可以由多种模式组成。给定一个特定的上下文，您可能会发现某些模式是比其他模式更合适。我们建议您采用端到端的解决方案视图，并考虑所涉及的模式，然后定义大数据解决方案的架构。

对于架构师和设计师，映射到模式可以支持对架构中各组件的责任进一步细化。对于业务用户而言，它通常有助于更好地理解大数据问题的业务范围，从而获得有价值的洞察，让解决方案满足所期望的结果。

此外，解决方案模式有助于定义最优的组件集，根据业务问题是否需要使用数据发现和探索功能、专用和可预测的分析或者可操作的分析。请记住，在实现一个解决方案时，并没有建议的原子、复合或解决方案模式的顺序或次序。在本系列的下一篇文章中，将针对此用途介绍解决方案模式。

作者: Divakar等  来源: DeveloperWorks

End.

# 大数据架构和模式（五）：利用大数据识别保险行业中的欺诈业务案例

**1.你认为该如何利用大数据识别保险行业中的欺诈业务？**  
**2.传统解决欺诈的方案是什么？**  
**3.本文解决方案模式：入门及获得高级业务洞察分别是什么？**

**简介**

本系列的 第 3 部分 描述了针对最常见的、经常发生的大数据问题及其解决方案的原子模式和复合模式。本文将推荐可以用于架构大数据解决方案的三个解决方案模式。每个解决方案模式都使用了一个复合模式，该模式由逻辑组件构成（参见第 3 部分的介绍）。在本文末尾处，列出了产品和工具清单，它们可映射到每一个解决方案模式的组件。

**解决方案模式**

以下各节将介绍可以用于架构大数据解决方案的三个解决方案模式。为了说明这些模式，我们将它们应用到特定的用例（例如，如何检测医疗保险欺诈），但这些模式可以用于解决其他许多业务场景。每个解决方案模式都利用了一个复合模式的优势。在下表中，列出了本文介绍的解决方案模式，以及作为其基础的复合模式。

**用例描述：保险欺诈**

财务欺诈对金融业的所有领域都带来了巨大的风险。在美国，保险公司每年要损失数十亿美元。在印度，仅仅是 2011 年的亏损总额就达到 3000 亿印度卢比。除了经济损失，保险公司还会失去一些业务，因为客户感到不满意。虽然许多保险监管机构已经定义了框架和流程来控制欺诈行为，但他们往往只是对欺诈做出反应，而不是采取主动措施来预防它们。传统的方法（如循环列入黑名单的客户、保险代理人和员工）并不能解决欺诈问题。

本文为大数据解决方案提出了一种解决方案模式，以本系列的第 3 部分中介绍的逻辑架构以及 第 4 部分 中介绍的复合模式为基础。

保险欺诈是为了让做出欺诈的当事人或其他关联方获得不正当或非法的好处的行为或疏忽。欺诈行为的种类包括：

**保单持有人欺诈和索赔欺诈**— 在购买和执行保险产品时对保险公司的欺诈包括提出保险索赔时的欺诈。

**中介欺诈**— 保险代理人、企业代理、中介或第三方代理人对保险公司或者保单持有人所做的欺诈。

**内部欺诈**— 保险公司的董事、经理，或其他任何人员或办公室成员对保险公司所做的欺诈。

**目前的欺诈检测流程**

保险监管委员会已经建立了反欺诈政策，其中包括明确定义的欺诈行为监控流程、搜索潜在的欺诈指标（并发布列表）的流程，以及与执法部门协调的流程。保险公司配置了专门分析欺诈索赔的工作人员。

**目前的欺诈检测流程的问题**

保险监管机构已明确定义了欺诈检测和缓解的流程。传统的解决方案使用的模型基于历史欺诈数据、被列入黑名单的客户和保险代理人，以及有关特定于领域的欺诈的数据。可用于检测欺诈的数据被局制于给定保险公司的 IT 系统和一些外部源。

目前的欺诈检测流程大多是手工的，只能处理有限的数据集。保险公司可能无法调查所有指标。通常很迟才检测到欺诈，而且保险公司很难对每个欺诈案例都进行适当的跟进。

目前的欺诈检测依赖于对现有欺诈案件的已知情况，所以每一种新型诈骗发生时，保险公司都不得不承担第一次的损失。最传统的方法在一个特定的数据源内工作，无法容纳不断增长的各种不同来源的数据。大数据解决方案可以帮助解决这些挑战，并在保险公司的欺诈检测中发挥重要作用。

**解决方案模式：入门**

该解决方案模式基于存储和探索复合模式。它集中处理数据的获得并存储来自企业内部或外部的不同来源的相关数据。在图 1 所示的数据源只是一个示例；领域专家可以识别适当的数据源。

因为必须收集、存储和处理来自多个来源的大量不同数据，此业务挑战是大数据解决方案的良好候选场景。

下图显示了解决方案模式，它映射到了在 第 3 部分 中介绍过的逻辑架构。

图 1. 入门解决方案模式

该图显示了逻辑层图，此图显示了入门解决方案模式

图 1 使用以下数据提供者：

1、外部数据源

2、结构化数据存储

3、经过转换的结构化数据

4、实体解析

5、大数据浏览器组件

医疗保健欺诈检测所需的数据可以从不同的数据源和系统中获得，比如银行、医疗机构、社交媒体和 Internet 机构。这些数据包括来自博客、社交媒体、新闻机构、各机构的报告，以及 X 光报告等来源的非结构化数据。更多示例请参见图 1 中的数据源层。利用大数据分析，这些不同来源的信息可相互关联和组合，并且被分析（在已定义规则的帮助下），以确定欺诈的可能性。

在这种模式中，所需的外部数据是从数据供应者那里获得的，他们贡献经过预处理的非结构化数据，这些数据已被转化为结构化或半结构化数据。在经过初始预处理后，这些数据被存储在大数据存储中。下一步是找出可能的实体，并从数据生成即席报告。

实体识别是在数据中识别命名元素的任务。识别分析所需的所有实体必须都被识别出来，包括那些与其他实体没有关系的松散实体。实体识别主要由数据科学家和业务分析师执行。实体解析可以像根据数据关系和上下文识别单一实体或复杂实体一样简单。此模式使用了简单形式的实体解析组件。

您可以简单地将结构化数据转换成最适合于分析的格式，并直接存储在大数据结构化存储中。

可以在此数据上执行即席查询，以获得相关信息：

* 给定的客户、区域、保险产品、代理或审批人员在给定期间内的整体欺诈风险
* 某些代理或审批人员，或跨保险公司的客户过往的索赔检查

顾名思义，组织一般采用这种模式开始使用大数据。组织采用探索式方法，根据可用的数据评估可以生成什么样的洞察。在这个阶段，组织一般不会对高级分析技术进行投资，比如机器学习、特征提取和文本分析。

**解决方案模式：获得高级业务洞察**

这种模式比入门模式更高级。它在索赔处理的三个阶段预测欺诈：

* 索赔已经解决。
* 索赔处理正在进行中。
* 刚刚收到索赔请求。

对于前两种情况，可以批量处理索赔，而且可以启动欺诈检测流程，它可能是常规报告流程的一部分，也可以由业务请求启动。第 3 种情况可以近实时地进行处理。索赔请求拦截器会拦截索赔请求，启动欺诈检测流程（如果指示器报告这可能是一个欺诈案件），然后通知在系统中识别的利益相关者。越早检测到欺诈，风险或损失的严重性就会越低。

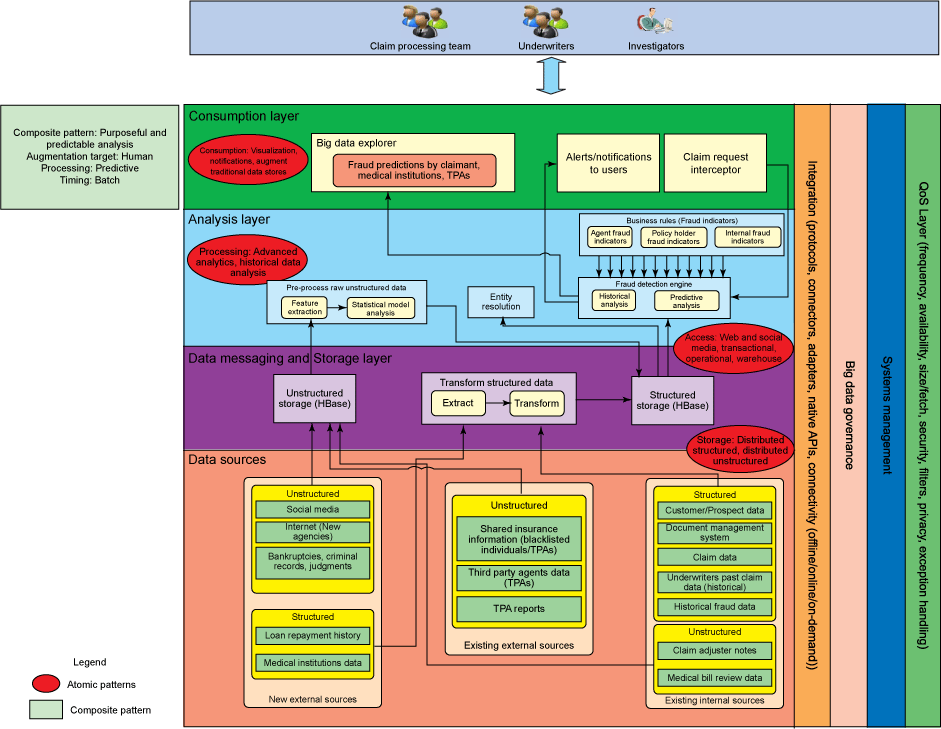


图 2. 获得高级业务洞察的解决方案模式

逻辑层图显示了用于获得高级业务洞察的解决方案模式

图 2 使用：

1、非结构化数据存储

2、结构化数据存储

3、经过转换的结构化数据

4、经过预处理的非结构化数据

5、实体解析

6、欺诈检测引擎

7、业务规则

8、大数据浏览器

9、对用户的警告和通知

10、索赔请求拦截器

在这个模式中，组织可以在分析之前选择对非结构化数据进行预处理。

数据被获取并按原样存储在非结构化数据存储中。然后，它被预处理成可以被分析层使用的格式。有时，预处理可能会非常复杂和费时。您可以使用机器学习技术进行文本分析，Hadoop Image Processing Framework 对于处理图像很有用。最广​​泛使用的技术是 JSON。经过预处理的数据随后被存储在结构化数据存储中，如 HBase。

此模式的核心组件是欺诈检测引擎，由高级分析功能构成，可以帮助预测欺诈。被明确定义并经常更新的欺诈指标有助于识别欺诈行为。下面欺诈指标可以帮助检测欺诈，并且可以使用技术来实现打击欺诈行为的系统。下面是常见欺诈指标的列表：

1、保单生效后不久进行索赔。

2、在处理索赔时发生严重核保失误。

3、被保险人在追求快速解决时具有明显的攻击性。

4、被保险人愿意接受小额赔偿，而不是记录所有损失。

5、文件的真实性值得怀疑。

6、被保险人在背后偿还贷款。

7、所产生的伤害是不可见的。

8、不存在已知伤亡的高价值索赔。

9、个人集群之间存在关系，包括保单持有人、医疗机构、联营公司、供应商与合作伙伴。

10、持牌及非持牌医疗服务提供者之间存在联系。

仅使用传统方法不足以预测诈骗。用户还需要使用社交网络分析来检测持牌及非持牌医疗服务提供者之间的联系，并检测保单持有人、医疗机构、联营公司、供应商与合作伙伴之间的关系。验证文件的真伪，并找到个人的信用评分，这是用传统方法难以完成的艰巨任务。

在分析过程中，对所有这些指标的搜索可能会在庞大数量的数据上同时发生。每一个指标都被加权。所有指标的加权总值表示预测欺诈的准确性和严重程度。

在分析完成后，可以向相关利益方发送警报和通知，并且可以生成报表，以显示分析结果。

此模式适合于需要使用大数据进行高级分析的企业，包括进行复杂的预处理，以利用先进的技术（如特征提取、实体解析、文本分析、机器学习和预测分析）可以进行分析的形式存储数据。这种模式不涉及采取任何行动或根据分析的输出提供建议。

**解决方案模式：采取下一个最佳行动**

在关于获得高级业务洞察的解决方案模式中所做的欺诈预测通常导致应采取特定行动，例如拒绝索赔，或暂缓赔偿，直至收到进一步的澄清和信息，或报告它，以采取法律行动。在这种模式中，为预测的每个结果定义了行动。这个行动对结果（action-to-outcome）的表被称为行动决策矩阵。

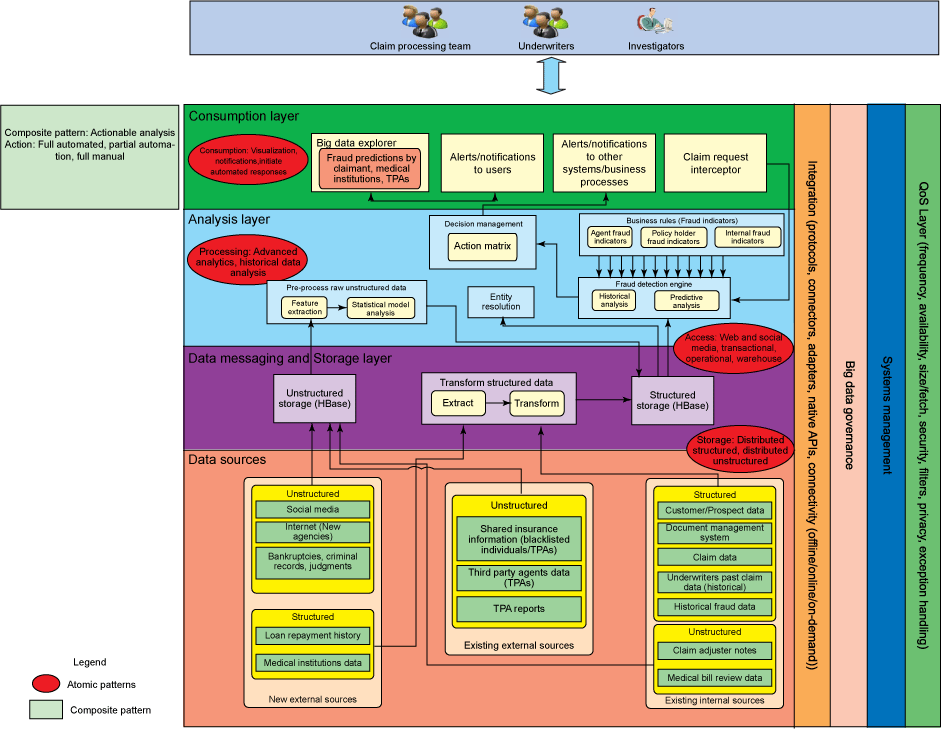


图 3. 下一个最佳行动的解决方案模式

逻辑层图显示了用于获得下一个最佳行动的解决方案模式

点击查看大图

图 3 使用：

1、非结构化数据存储

2、结构化数据存储

3、经过转换的结构化数据

4、经过预处理的非结构化数据

5、实体解析

6、欺诈检测引擎

7、业务规则

8、决策矩阵

9、数据探索工具

10、对用户的警告和通知

11、索赔请求拦截器

12、对其他系统和业务流程组件的变更和通知

通常可以采取 3 种行动：

* 可以向利益相关者发送通知，以采取必要的行动，例如，通知用户采取针对申索人的法律行动。
* 系统在采取进一步行动之前通知用户，并等待用户的反馈。系统可以等待用户响应任务，也可以停止或暂缓索赔处理事务。
* 对于不需要人工干预的场景，系统可以采取自动行动。例如，系统可以向流程发送一个触发器，以停止理赔程序，并将有关申索人、代理人和审批人的信息通知给法律部门。

此模式适合于需要使用大数据进行高级分析的企业。此模式使用高级功能来检测欺诈行为，通知并提醒相关利益方，启动自动工作流，根据处理的结果采取行动。

**形成大数据解决方案骨干的产品和技术**

下图显示了大数据软件如何映射到第 3 部分中所描述的逻辑架构的各个组件。这些产品、技术或解决方案可以在大数据解决方案中使用；必须根据您自己的需求和环境来决定选择用于部署的工具。

图 4 显示了大数据设备，如 IBM PureData™ System for Hadoop 和 IBM PureData System for Analytics，它们跨越多个层次。这些设备具有内置的可视化、内置的分析加速器，以及单一系统控制台等特性。使用设备有很多优点。（请参阅 参考资料，了解有关 IBM PureData System for Hadoop 的更多信息。）

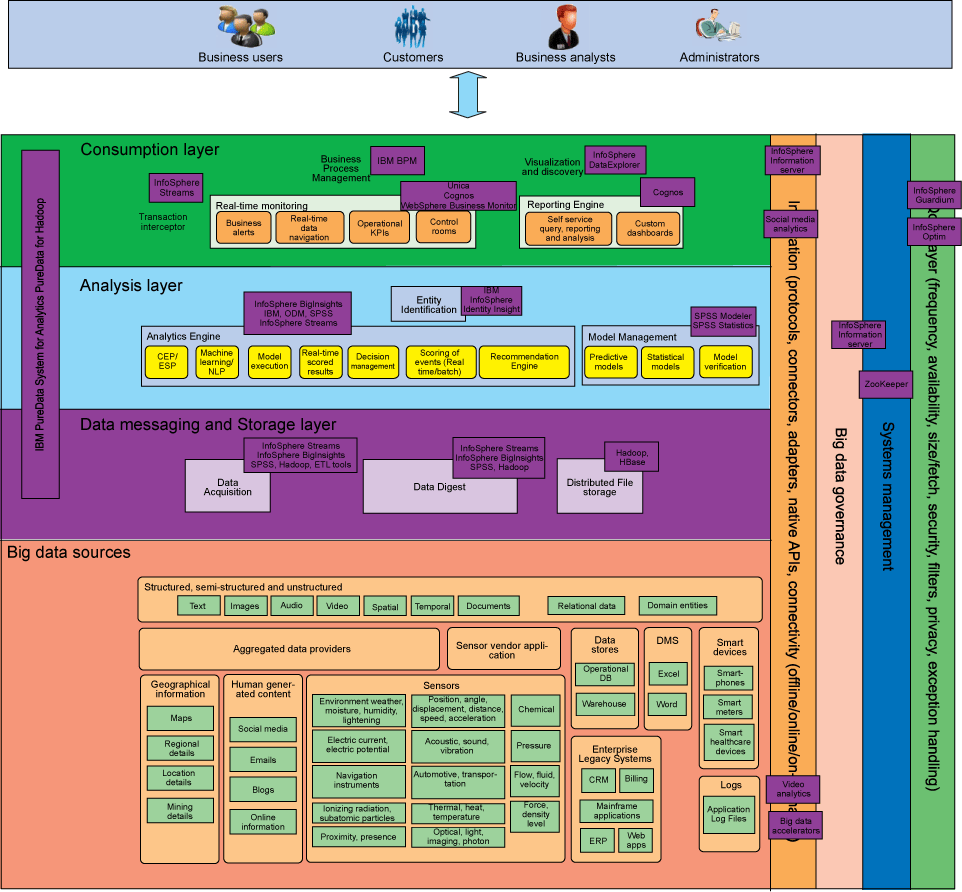


图 4. 映射到逻辑层次图的产品和技术

逻辑层图显示了一些产品

**在欺诈检测中使用大数据分析的好处**

与传统方法相比，使用大数据分析来检测欺诈具有多种优势。保险公司可以构建包含所有相关数据源的系统。一个包罗万象的系统有助于检测不常见的欺诈案件。预测模型等技术可以深入分析欺诈实例，筛选明显的案例，并参照低发欺诈案件执行进一步分析。

大数据解决方案还可以帮助建立整个企业的反欺诈工作的全局视角。通过链接组织内的关联信息，全局视角往往导致更好的欺诈检测。欺诈行为可以在多个起源点发生：理赔、保险退保、缴费、申请新的保单，或者与员工相关的欺诈或第三方欺诈。各种来源的数据相结合可以实现更好的预测。

分析技术使组织能够从非结构化数据中提取重要信息。虽然大量结构化信息存储在数据仓库中，大多数关于欺诈的关键信息都是非结构化数据，比如第三方报告，它们很少被分析。在大多数保险机构中，社交媒体数据没有被正确地存储或分析。

**结束语**

利用基于保险行业中的识别欺诈用例的业务场景，本文介绍了复杂性各不相同的几种解决方案模式。最简单的模式解决来自不同来源的数据的存储，并执行一些初步的探索。最复杂的模式涵盖如何从数据中获得洞察，并根据分析采取行动。

每一个业务场景都被映射到组成解决方案模式的相应原子模式和复合模式。架构师和设计师可以应用解决方案模式来定义高级的解决方案，以及相应的大数据解决方案的功能组件。

作者: Divakar等  来源: DeveloperWorks

End.