2011 年分析革命：优化报告和分析普及可操作智能

IBM 的供稿人：

David Loshin
Knowledge Integrity, Inc.。
2011 年 3 月

目录

1 执行摘要

分析值的紧急识别与迅猛增长的结构化和非结构化数据量之间存在冲突。有竞争力的组织通过采用整合商业智能和分析的策略和方法正不断取得发展，这扩充了日常（有时是即时）所做决策的范围。受数据困扰的个人可能会面临分析失效，但应要求向合适的人员交付值得信任的可操作智能可以避开分析失效，促进理智和自信的决策。

在本文中，我们将考察可以促成普遍分析的引入的业务价值驱动因素，并且审查某些关于使用可操作知识来发展竞争优势的常见示例。我们将了解普遍分析的概念，包括使操作和事务型数据源流入企业分析引擎，最终用于增强和显著地改进业务流程。

然后，本文将概述实现普遍分析所需的技术构成。将进行更深入地剖析，以提供关于这些数据管理基本情况的详细信息：

- 持续数据同步；
- 聚合信息整合；
- 强大的分析架构；
- 用以处理大结构化和非结构化数据集的性能调优；
- 大规模计算；
- 综合分析服务；以及
- 报告框架、临时查询、维度分析以及可操作知识的可视化和交流。

然后，我们来看一些通向成功的关键挑战，并结合一些可支持成功企业级商业智能、报告和分析程序的建议进行总结，这些分析程序带有支持企业的端到端报告和分析需求的一套完整的工具。

2 不断增长的数据量和智慧决策

多年以来，技术分析师一直在预想数据的迅猛增长。2010 年刊登的一篇文章称数据量将继续以正常速率扩展，文章指出“最大数据仓库的容量约每两年增加两倍。”[1] 数据的爆炸性增长在全球范围内被认为是关键信息技术 (IT) 问题。根据 2010 年 10 月的一份 Gartner 研究，“在今年夏天进行的一项调查当中，有 47% 的调查受访者将数据增长列为他们的前三大挑战之一”。[2] 以迅猛数据的增长为例，零售商沃尔玛每小时执行 100 多万笔客户交易，数据库更改估计在 2.5 PB 以上。持久性不是要求，2010 年的一份报告指出，到 2013 年，每年互联网上的流量将达到 667 EB。[3]

数据库中的结构化信息仅是冰山一角；一些重要的里程碑事件也记录了非结构化数据的爆炸式增长历程： 到 2009 年年底，据估计数字信息量已经增长至近 800000 PB（或者 800000000000 GB），到 2010 年年底总量有望达到 120 万 PB！ 以这样的速率发展，到 2020 年，数字数据量可能增长至 35 ZB（1 ZB=1 万亿 GB）[4]

人们日益期望通过各种互连系统来执行复杂的业务流程。整合传感器和探测器不仅实现持续的操作性能测量，而且很多系统互联允许快速传达和永久保存这些措施。据估计，每天会生成 15 PB 的新信息，其中 80% 是非结构化数据。[5]

我们可从这几个数字信息量的快速扩展示例得出结论，即通过结构化和非结构化数据的综合分析可以开启新的令人振奋的远景，这在以前是做不到的。除了本文中探讨的很多其他机会外，通过新的和改进的数据分析方法，组织还可以识别新的业务趋势、评估疾病传播或者打击犯罪。以我

[1] Merv Adrian，“探索数据库增长的极限”IBM Data Management，2010 年 第一期
[2] Lucas Mearian，“Gartner 表示数据增长仍然是 IT 最大的调整”2010 年 11 月 2 日 Computerworld（下载地址 http://www.computerworld.com/s/article/9194283/Data_growth_remains_IT_s_biggest_challenge_Gartner_says）
[3] 经济学家，“数据，数据无处不在”2010 年 2 月 21 日（下载地址 http://www.economist.com/node/15557443?story_id=15557443）
[4] Gantz，John 和 Reinsel，David，“数字寰宇的十年 – 你准备好了吗？ IDC 2010 Digital Universe
[5] Bates、Pat、Biere、Mike、Weideranders、Rex、Meyer、Alan 和 Wong、Bill，“智慧地球的新智能，” ftp://ftp.software.ibm.com/common/ssi/pm/bk/n/imm14055usen/IMM14055USEN.PDF

们现在的情况来看，信息正成为信息输出。”[6] 并且这些新理念不仅仅潜藏在结构化数据库中。分析也必须涵盖非结构化数据工件。

但是，随着数据量的增长，查找使这些业务流程在其优化级别运行所需的这些关键信息的复杂度也会随之上升。问题在于不再需要捕获、存储以及管理该数据。而挑战则来自于需要从中提炼并在正确的时间向正确的人员交付相关知识，以增加日常出现的数以百万计的决策制定机会。

大体上，可以将此归纳为我们希望跨组织的所有职能和级别以普遍方式将可操作智能整合到策略和操作流程中。不管这意味着向高级管理层通知新出现的收入机会，实时洞察企业绩效指标，还是为了最好地解决客户服务中断而每小时重新调整现场维修团队日程，积累、转换和分析信息以在正确的时间向正确的人员提供快速、可靠分析的能力可以增加发展机会和竞争力。

3 交付可操作智能

几乎在每个业务流程中，都存在信息积累和用于帮助个人制定决策的情况。有些决策意义深远、驱动全球企业策略，而其他决策是运作性的且范围较窄，如选择成本最低、尺寸合适的箱子来最佳地适应客户的要求。并且，对每个业务流程来说，都有绩效措施，如投资回报、实现价值的时间或材料成本。

在完美的世界中，每个决策都将是最优的，即决策的结果将带来最好的整体绩效。但是，决策制定者通常无法获得制定最优决策所需的全部信息。提供更多信息不总能解决问题，当海量未经筛选的数据在组织内输送时，受数据困扰的个人会陷入“分析失效”，在等待更多一点数据，以简化（并可能支持）即将制定的决策时，会强制延迟制定决策。当信息过载被节流回正常状态，以筛选出以最优方式驱动决策流程所需的特定信息时，可缓解“分析失效”并，并允许进行特定操作。应要求向正确的人员交付值得信任的可操作智能可避免分析失效，并促进理智和自信的决策。

作为该连续过程的一段，对公司整体绩效进行审查，备选方案被视为有助于调整公司的长期价值产生策略。作为另一端，运作活动通过可以实时调整和优化活动的特定智能加以改进。在在最理想的情况下，业务流程本身合并可操作智能的购置和演示、将分析结果直接引入流程、跟踪绩效

[6] Op. cit.，经济学家

度量，并指出何时制定更好的决策。

概括起来，运行、运行分析以及对分析驱动的业务流程的修改、然后对运行业务流程和相应应用程序的更改之间存在“闭环”关系。可操作智能通知策略和运作流程，而其向组织各层的员工的普遍交付可以推动从响应过去发生的事到简化最优决策的转变。

3.1 驱动组织优化的策略

随着组织内越来越多战略家意识到报告和分析功能，人们越来越希望采用数据分析来主动管理和改进业务策略。观察环境能提供组织将如何运作的洞察，还能使分析师了解到失败、成功的方面，并区分可重复利用的策略，从而取得更大的成功。

纵向上下文内的传统报告使分析师可查看组织内出现的情况并审查性能指标如何随时间变化对趋势进行评分。按照不同的定量和定性的层次维度组织数据，如位置、组织、客户配置文件、产品类别等等，使分析师可细分数据以寻求明确的业务机遇或流程改进。交互环境（如仪表盘和混搭）可推动明智的战略决策。

根据组织策略方向做出调整，融入关键性能指标的实时交付可让高级利益相关方审查短期窗口内这些战略决策的影响和结果，从而增加灵活性、降低风险，并且使企业能够快速响应新出现的业务机遇。更直接的说，获得组织如何运作（或无法运作！）的更深入洞察，可告知战略家应对公司运营模式做出重大调整。

3.2 整合分析和运营

实际上，以可操作智能形式制定的预测性分析的结果明确可以直接整合到运营流程中，甚至在业务消费者无意识的情况下进行。例如，通过将预测性分析结果与客户配置文件合并，该组织各个层面的决策者就可以迅速认识到新出现的机遇并对其做出反应，如实时调整呼叫中心脚本，或根据实时仓库、储藏室的库存数据、销售点数据或者甚至交通或天气数据重新路由产品交付。客户偏好、配置文件以及 Web 统计数据可驱动网站上动态内容的重新排列，以改进最终用户响应。

3.3 常用价值驱动因素

鉴于持续改进与客户和合作伙伴业务交互的共同愿望，我们可以考虑业务智能和分析如何通知与所有行业的常用价值驱动因素有关的决策流程。如今，大多数组织以如下两种方式使用数据： 交易/运营使用（“经营业务”）以及分析使用（“改善业务”）。当分析结果渗透到运营使用时，组织可利用已经发现的可操作知识来驱动改进。

通过分析管理业务转型的简单方法涉及到对简单分类计划中积极业务影响的机遇进行分类，以及列出潜在业务改善的主要类别，包括以下几个方面：

- 财务机遇，如减少运营成本、增加收入、识别新机遇、加速现金流，或减少处罚、罚金和其他费用。
- 与更精确信贷评估相关联的风险和合规机遇，降低投资风险，在资本投资和/或开发方面更具竞争力、更明智的决策，减少欺诈和泄漏，以及在审计方面遵守政府法规、行业预期或自我施加的政策（如隐私政策）。
- 基于信心和满意度的机遇，如客户满意度增加、工作条件改善和员工保留率提高、供应链简化、预测更精准、组织可信度上升、运营和管理报告更加一致，以及更好的决策。
- 工作效率机遇，如工作负载减少、吞吐量增加、制造缺陷减少以及最终产品质量提高。

此分类意在支持分析流程，并协助阐明一般业务目标和相应绩效指标。通过分析改进业务不仅仅需要安装和运行工具；关键利益相关方必须定义可实现目标，并使用这些工具来通知决策流程以及评估、测量和控制实现目标的程度。可在任意这些关键类别中围绕改善来设计和开发具体程序。考虑以下示例：

- 通过客户配置文件和目标营销的收入生成 – 客户分析可涵盖个人客户配置文件（包括有关每个人的人口统计、消费心态以及行为数据）的持续改进，以根据有区别的变量和相应价值集支持客户社区细分为各种群集。鉴于这些不同客户群集以及了解到分类方案允许开发微观营销策略，以使用类似配置文件扩大针对小群集客户的活动。“激光式”营销可使用客户分析结果直接聚焦个人。

- 借助对欺诈、滥用和泄漏进行鉴别的风险管理，欺诈包括有意窃取知识的行为，这种行为或举动可能导致不正当获益，通常通过入侵系统常景实施。欺诈检测是一种在某些既定情景中寻找运营模式的相关频率的分析。同时，在其合同/协议上下文内向顾客所提供产品和服务的综合分析可能会突显泄漏。这些风险都是可分析的，并且能够引起适当内部机构的关注，以采取弥补措施。

- 通过配置文件、个性化和客户生命周期价值分析提高的客户满意度 – 客户生命周期值分析计算客户在关系生命周期期间的盈利能力指标，纳入与管理该关系相关的成本以及期望来自客户的收入。使用客户配置文件结果不仅仅可通过定制材料或内容的演示来增强客户体验。客户配置文件可直接整合到所有用户交互，特别是入站呼叫中心，其中客户配置文件可提高客户服务代表接洽客户、加快问题解决的能力，甚至可以增加产品和服务销售。

- 通过开销分析改进的采购和收购工作效率 – 开销分析包含产品采购和供应商数据的收集、标准化和分类，以选择最可靠的供应商，简化 RFP 和采购流程、降低成本、改进高价值供应链的可预测性，以及提高供应链的可预测性和效率。

这只是众多示例中的一小部分，其中分析可用于优化所有行业中普遍使用的价值驱动因素。

3.4 行业示例

一方面，改进机遇在不同行业会有所不同，而另一方面，不管是哪个行业都有可改进的通用运营维度。使用相同的价值驱动因素层次结构，特定行业内的公司可从为该行业特别定制的报告和分析中获益，如以下“纵向”示例：

- **医疗保健** – 监控业务流程绩效渗透到了医疗质量的所有方面。例如，了解某些医务人员在处理某些状况会更成功的原因可以改进医疗质量。分析有助于发现致使某一方法比其他方法成功的因素，并了解这些成功案例是依赖于从医人员能控制的变量，还是不能控制的因素。改进的诊断方法可降低对高成本诊断资源的需求，如映像仪器，并且更好的治疗可减少病人的住院时间、腾出病床、提高吞吐量，并实现高病床使用率。

- **物流/供应链 –** 运输和物流管理的整合分析可洞察对有效供应链多方面的评估。例如，业务智能用于根据一系列地理范围、人口统计和消费心态分析特定产品的使用模式。可预测性成了一个神奇的词 – 了解哪个地区的哪类人群在特定时间段会购买何种范围的产品，可帮助更准确的预测（并从而满足）需求。因此，制造商可路由合适数量的产品来减少或消除库存缺货。同时，了解不同时期内按区域的需求会导致更准确的规划，以交付包装、方法和安排。可从起始点就距离映射产品销量，如果某些位置的销量比在其他位置低，那么可能表示供应链有故障，需要审查并实时实施补救。

- **电信服务 –** 在这个持续与客户打交道的行业中，增加客户的业务承诺会促成更长的客户生命周期。例如，检查客户手机使用情况可帮助确定每个人的核心网络。如果客户很少打家庭座机电话或个人移动电话，则表示客户可能更倾向于“朋友和家人”服务计划，降低对频繁拨打的号码的收费。确定核心网络内的家庭关系可通过整合移动帐户、或者交叉销售其他服务（如座机服务、互联网和其他娱乐服务）实现服务捆绑。另一方面，如果客户个人移动电话大多数是业务电话号码，并且持续时间为半小时到一小时，那么最好向该客户提供业务电话服务关系，该关系与其他移动连接服务呼叫捆绑。

- **零售 –** 大量销售点数据使其成为分析的成熟资源，零售点一直寻找优化产品定位以增加销量，同时减少开销来增加利润的方法，特别是当市场篮子不能通过附属卡直接与个人相关联时。了解实体店位置和其周围居民的类型的关系将帮助商店管理者根据商店分类选择产品。策略产品定位（如中间架子或顶端）可为那些驱动盈利能力的产品保留，并且可通过商店根据按客户细分的产品销售与客户旅游模式图进行组合）。产品位置不局限于物理位置；大量的 Web 日志可用于分析客户行为，以帮助动态地对网站上的服务定位重新排列，以及根据被弃置的购物车分析、通过协作筛选或根据客户自己的偏好分析鼓励产品向上销售。

- **金融服务/保险 –** 在保险和银行业中，识别风险和管理隐患对提高盈利能力至关重要。提供一系列金融服务的银行将开发与识别其他风险变量的客户活动和配置文件相关联的精确模型。例如，分析信用卡的大量购买与抵押失败有关，可能为在特定购物中心购物或在特定类型的快餐店吃饭的人展示增加的缺陷风险。反过来，识别缺陷风险的指示性行为可能有助于银行预测缺陷事件，并且接触到那些把备用产品留在家里的个人、降低缺陷风险并长期改进贷款现金流的可预测性。

- **制造 –** 工厂性能分析对保持可预测的和可靠的工作效率至关重要；跟踪生产线性能、机器停机时间、生产质量、进行中的工作、安全事故，以及交付运营绩效指标及管理升级链的测量，以便在合适的环境、合理的时间范围处理不利事件。

- **服务业 –** 连锁酒店评估客户配置文件和相关旅游模式，并且了解特定客户可能在其他竞争者中划分其年度“晚间分配”。通过分析客户旅游偏好和首选位置，公司可通过忠诚度计划来提供刺激性服务，以获得该客户更多的晚间分配。

这些行业示例相似处在于，以为分析范围包括关键业务绩效指标的简单报告，以及探索优化组织运营模式或改善与客户和其他业务合作伙伴交互的机遇。业务流程的调查以及任意行业的绩效度量将产生专门从报告和分析获益的方法建议。

4 为什么要普遍分析？

“业务智能”和“分析”的概念包括支持组织中一群用户社区的工具和技术，原因是要收集和组织大量（和多元）数据集以支持运营、战略以及策略层面的管理和决策。通过数据收集、聚合、分析以及演示，可交付可操作智能，以便最好地服务广泛的目标用户。数据仓库程序已经变得成熟的组织，可让用户从企业信息资产提取可操作知识，并快速实现业务价值。

但是，传统数据仓库架构支持业务分析查询和封装的报告或高级管理仪表盘、信息洞察的综合程序，且智能可增强在大量策略、战略和运营角色中各种类型的员工的决策流程。甚至在直接运营

环境内整合相关信息成为了成败的关键因素。脱机用户分析提供一般销售策略是一个方面，但实时可操作智能可根据客户的交互历史为该销售人员提供特定的备选方案，以最佳的方式同客户交流，同时优化企业盈利能力和销售人员佣金。最大化所有相关方的整体利益，最终改善销售、增加客户和员工满意度，以及提高响应速率，同时降低货物销售成本 – 实现对所有人的真正的共赢。

广泛的分析功能都有助于针对一系列价值日益上升的问题提供解决方案方面的建议：

- **发生了什么？** – 预定义报告将为运营管理者提供解决方案、详细描述在组织内部发生了什么以及剖析查询结果的各种方法，以了解业务活动的基本特征（例如，计数、总和、频率、位置等等）。传统 BI 报告提供 20/20 事后认知 – 告诉您发生了什么，可提供关于已发生事件的聚合数据，甚至指示个人就已发生事件采取具体的措施。
- **原因？** – 更多的突发查询与一系列时间内测量和指标的审查相结合，使我们更关注审查。深入报告维度使业务用户得到更多尖锐问题的解决方案，如任意报告问题源，或者比较跨相关维度的特定绩效。
- **如果怎样？** – 更多高级统计分析、数据挖掘模型以及预测模型可让业务分析师考虑不同的措施和决策可能会如何影响结果，从而得到改进业务的新想法。
- **接下来是什么？** – 通过在预测、规划和预测模型内评估不同选项，高级战略家可衡量可能性并做出战略决策。
- **这如何实现？** – 通过考虑组织绩效优化的方法，高级管理者可采用业务策略来更改组织经营业务的方式。

信息分析使解决这些问题变得可能。改进的决策流程有赖于辅助业务智能和分析能力，这些能力将增加交付可操作知识在很大范围内的复杂性及价值（如图 1 所示）。随着分析功能的复杂性增加，业务客户可获得更多优化机制的洞察。统计分析将帮助隔离任何所报告问题的根本原因，以及提供一些预测功能（如果现有模式以及趋势在没有调整的情况下继续）。捕获过去模式的预测性模型有助于预测“假设”场景，为组织的高绩效指导战略和策略。

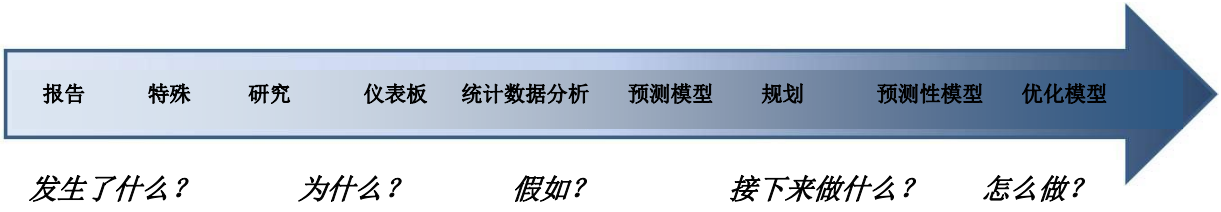

图 1： 种种有益于分析各种客户的技术。

智能分析和业务智能正逐渐成熟，转变为帮助优化业务的工具。这些工具是否确实用于帮助

- 满足战略目标的C 级别主管审查选项，
- 寻求精简业务线的高级管理者，或者
- 以认为不可能的方法做出的运营决策。

这些分析整合数据仓库、数据挖掘、多维分析、流以及混合应用以提供敏锐的愿景，该愿景可实现对新出现机遇的即时反应，同时允许评估一段时间的环境以发现改进和扩展业务的方法。

5 了解技术考虑事项

早期数据仓库架构师力争设计和执行有效的方法，以从源系统提取数据，将该数据转换为适合分析的格式，并且将其加载到数据仓库。一旦数据已在数据仓库内被组织，就利用报告和分析工具（如联机分析处理 (OLAP)）和查询与报告工具将数据交付给知识客户。

尽管这些方法的某些核心方面仍没有变动，但自二十世纪九十年代中期以来，该技术基础架构的细节已经演变，以便满足来自更多源的更多数量级的数据可用性上的日益增长的需求。相应的，知识用户池的多样性和期望值增加了在类型、可扩展性和可操作知识交付机制上的需求。如今，在正确的时间将正确的信息交付给正确的人员是数据管理最佳实践的终极阶段，并且组织与技术基础架构组合旨在指导大量稳定数据通道进入可在实时约束范围内交付可靠结果的高性能分析平台。这取决于以下技术领域：

- 来自多源的数据持续同步以提供连贯一致的视图；
- 支持海量高质量数据的聚合信息整合被协调并保持一致以实现有效分析；

- 收集、排序以及管理大型数据集的强大架构，用于分析；
- 性能调整特性从硬件改进到调整编程以及执行模型，以处理大量未结构化数据集；
- 大规模计算，包括处理和分析从很多数据源积累的大量数据集，通常带有实时期望结果；
- 全面分析服务集（包括非定向分析、预测性模型以及文本分析），减少或消除成为强大用户的需要以从可操作智能获取利益；
- 报告、突发查询、维度分析以及可视化和可操作知识通信的框架。

本节，我们考察某些现代分析环境必需的关键技术考虑事项，以及其结果可如何全面整合到成百上千的日常业务流程中。鉴于对技术成分的了解，我们将开始研究环境中异构技术相结合以支持分析程序而出现的某些挑战。

5.1 持续可用性： 数据同步和一致

不断出现的传统数据仓库框架挑战涉及在报告或分析与其交付之间铰接需要延长的周转时间。延迟源于当前数据的可操作知识交付使及时决策变得困难，并且长数据延时将影响在合理时间框架内评估那些决策结果的能力。从本质上来说，数据同步已经成为整合策略和运营决策活动的基础架构内的关键成分。及时和当前数据的关键性不容低估。例如，特定区域内的产品类别销量增加可表明物流资源的立即重新分配需求不断增加，以此来防止库存缺货并满足该需求。依赖过去一周的订单信息是不够的，相反，当前架子和仓库库存信息结合供货资源分配可立即实现决策，保持稳定的产品流给客户。

普遍业务智能和分析要求高级别的数据同步，这意味着环境必须

- 降低或消除数据延时；
- 保持用于跨企业分析的数据一致；
- 提供及时和当前信息；
- 向类似要求提供一致、决定性的结果。

在组织内（也来自外界）多种可用数据源间保持合理的同步和一致程度，要求针对持续数据可用性的技术策略不对环境施加压力。部分策略包括：

- 为复制数据更改数据捕获 (CDC)，包括从一个源到另一个或多个目标数据系统的数据的复制管理、同步。CDC 是为捕获来自源数据集更改的事件驱动机制，并通过各种通道，直接面向目标数据库或者通过后续编程的信息查询，传播那些更改。通过 CDC 修改同步数据实现运行系统和分析系统的一致，实时启用可操作机遇的发现，同时保持各报告系统的一致性。

- 数据联合，实现了对异构（通常是物理分布）数据类型、平台和源的透明访问，并且格式众多，不要求暂存区域或集中式存储库（见图 2）。联合对捕获大型或极大的分布式数据集的子集是有效的方法，且当数据是异地数据、格式较老时频繁使用，或很少使用。例如，数据联合框架将允许应用程序访问数据库、XML 数据、平面文件或者数据服务或使用统一访问机制的数据流。相应的，联合服务器自动访问数据源并返回同步结果。联合简化整合来自多个源的数据，实现信息的“交叉影响”以更好的发现机遇，
  并且对在数据到达目标之前加入相异数据非常有效。

- 信息流编程，实时提供应用程序持续访问流数据。连接数据流和永久数据源支持复杂事件编程，以及根据新出现的知识活动的机遇的实时发现，如基于天气的商品交易或防止零售库存缺货的立即交付。

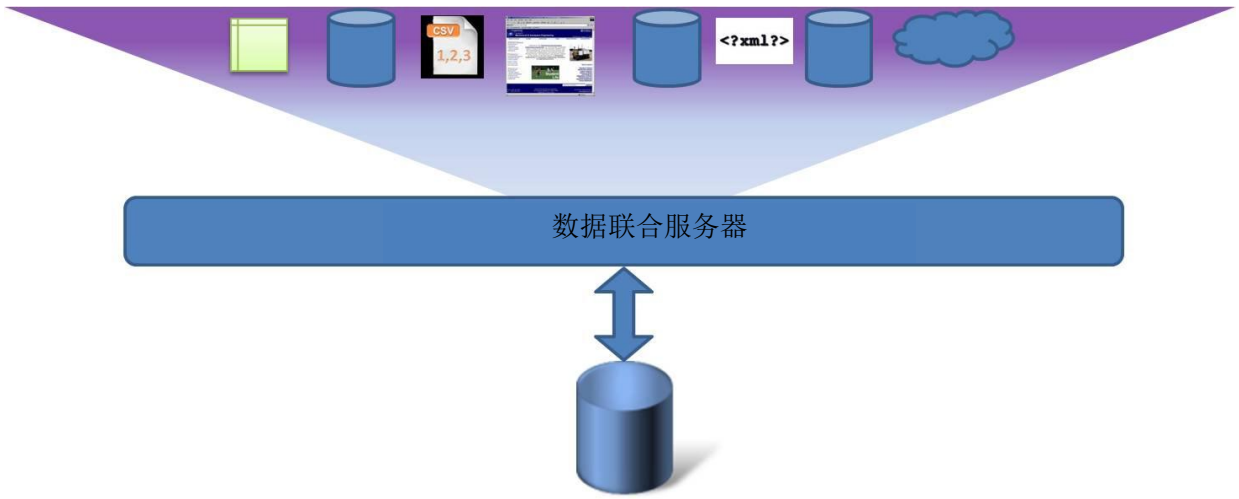

图 2： 数据联合服务提供对各种数据源动态、透明的访问。

5.2 一致的信息整合

因为大多数分析的数据都是为了特定事务或运行目的从分析点进行上游创建，所以有必要提供一致信息整合框架来促进信息获取和共享。换句话说，为分析准备企业数据要求协作技术从多个源获取数据，管理表示形式不同的相似数据概念的表示，然后与不同应用程序的多个客户共享信息。

- 数据整合和 ETL – 像数据联合一样，数据整合依赖于无缝地访问来自许多不同源数据的能力，并且将该数据交付到各种目标，但是关于解析提取数据、将其规范为标准格式、清除以及将数据转换为适合加载到数据仓库，以及报告和分析的任何后续使用方面其包含的内容要更多。ETL 是数据仓库的关键使能因素，提供基本信息流共同实现业务智能。

- 主数据管理 – “主数据”对象就是那些在不同数据孤岛上表示并在各组织不同的业务应用程序上使用的核心业务概念，还有与其相关联的元数据、特性、定义、角色、连接以及分类和层次结构。某些常见的示例包括客户、供应商、部件、产品、位置、协议、联系机制。主数据管理 (MDM) 整合了业务应用程序、信息管理方法和实施策略的数据管理工具、程序、支持捕获的基础架构、整合、以及准确、及时、一致和完整的主数据的后续共享使用。MDM 可管理对每个跨应用程序基础架构的主数据实体的独特可识别表示的访问。提供

由多数据源组装的核心数据概念的“统一视图”对减少客户、产品、供应商或其他关键数据资产类别的复制很有帮助。

- 元数据管理 – 仅有适量的元数据对支持数据提取、整合以及确保意义的一致性分析要求是必需的。元数据整合了业务定义（与数据概念、业务条款相关联以及相关的定义和语义）、概念引用数据域、其相应值、数据元素定义、数据元素格式、结构以及数据类型还有实体模型、实体关系、以及支持数据监管的元数据如信息使用图、数据质量规则以及访问控制。管理数据定义、引用表、数据映射促使信息重新使用，并且从各个系统整合时帮助同步数据语义。

- 数据配置文件 – 像所有资产一样，它对库存资产很宝贵 – 决定它是什么、属于谁、用于何处以及其易维护性。数据配置文件能这样做 – 它支持项目审查和源数据集的经验分析的组合，旨在了解对分析至关重要的数据元素元数据的特性。数据配置文件可提供与实际数据值相关联的“地面实况”，以及为元数据提供一致的证据，并将提供候选源的适用性的洞察以满足目标分析需要。结合消耗应用程序预期审查使用数据配置分析所进行数据质量评估的结果，可能不涵盖在元数据存储库内管理的数据质量规则。

- 数据清除 – 数据配置分析将暴露数据中的潜在异常情况和错误，而作为分析处理的输入使用的错误数据将影响其结果的可信度。如果源数据系统所有者能够解决会引入数据问题的设计或流程缺陷，则可以消除根本原因以及错误。但是，如果源数据不受组织控制，则流程和技术基础架构必须到位，以解析、标准化、增强/强化和清除数据以满足下游分析需求。数据清除通常在数据转用时要求，尤其是在新用途有更为严格的数据质量预期时。数据清除会增加数据的价值，而一致的高质量数据会令决策制定流程值得信赖。

- 数据验证 – 或者，常常会把新错误引入数据，而在信息生产流的早期识别和消除潜在的数据缺陷将减少下游的不确定因素和不一致性，并提高整体操作效率。但是更为重要的是，通过定义的数据规则验证提供的数据将有助于交付值得信赖的数据。

- 身份识别 – 由于整个企业内部署的大量操作系统的有机增长，因此在表示不同的系统中的多个数据示例以多种方式表示同一现实实体并不少见。或者，常常会出现两个现实实

体共享同一识别属性的情况，这样就难以区分它们。这两类问题都反映出相同的核心挑战： 评估记录对之间的相似性以及确定它是否表示相同的事物的能力。身份识别通过计算任意两条记录之间的相似程度并打分来解决这些问题。当该得分超过特定的阈值后，将假定这两条记录匹配；如果低于另一阈值，则将其视为不匹配。身份识别用于在存在不确定因素或不完整的属性时来匹配记录，或者确定两条记录是否真正表示不同的实体。身份识别是主数据管理的重要组成部分，并且在实体匹配中提高精度将减少数据重复，同时支持高质量的报告和分析。

- 普遍的交付机制 – 该电子表格不再是交付分析结果的唯一方式。报告的自助服务配置和来自业务应用程序的查询结果将消除信赖 IT 人员的支持引起的瓶颈，并且基于 Web 的交付将简化使用各种直观的交互式可视表示对象（如图形、热点地图）的结果的可用性。通过自动化发出可以对各种接口定制的事件驱动的通知（范围从台式机到手持 PDA），行动情报可以在需要时直接提供。

5.3 使用强健而可扩展的架构管理知识

虽然大多数组织的主要利益相关者关注我们的每一个高级别价值驱动因素，每个公司的业务领域以及特定于其业务线的外部压力会驱动对解决常见价值驱动因素改进的框架的需求，同时通知每个行业所特有的业务流程以及相应的报告和分析需求。这意味着与这些常用主数据域（如“客户”、“产品”或“协议”）整合的数据仓库模型，可以在增强的模型中实现灵活性来解决行业特定的难题。考虑以下示例：

- **银行业**，有对关注企业风险管理平衡、客户可见性监督以及产品盈利能力的模型的需求。从风险角度看，这会结合管理各种政府层面的法规法令（Basel II、萨班斯-奥克斯利法案、国际财务报告）及与之关联的法定报告的合规性，以及管理管理公司财务、交易对方和信用风险。同时，这些企业将有对各种报告和客户互动分析的需求，以帮助保持客户忠诚度、增加保留以及度量渠道的有效性。
- **保险业**，有对解决审查和管理索赔、每成员/每月报告、中介业绩、合规性（如 Solvency II 要求）、风险管理、基本生活养老金保险精算以及企业年金合规性相关难题的模型的需

求，但仍要求对客户管理和保留、产品性能以及市场营销和销售有效性具有可见性。

- **零售业**，有对客户管理、推销管理、产品和服务管理、商店运营管理、公司财务管理、供应链、多销售渠道管理以及法规遵从可见性进行报告和分析的需求。此外，实时预测性分析的结果还可以回流进入运营市场营销和销售流程，以改善创收并降低客户损耗。
- **电信业**，他们希望监控市场营销绩效、广告管理、收入保障、补给品供应、评级、计费、信用风险、客户策略和计划分析，以及报告语音和非语音通信的使用情况、客户保留和损耗以及系统和网络绩效和使用情况分析（及减少的呼叫数！）来帮助制定决策，以便在电信基础架构方面实现重大改进。

正如您可以看到的，所有这些示例都要求强健的架构，以支持收集、聚合、排序以及管理非常大的数据集。而且，虽然所有这些行业示例都有大量关于市场和销售效力或客户关系管理的共同目标，但是，每一个示例都有报告和/或分析工作负载的不同混合形式，可以从强健的行业数据仓储模型中获益。

5.4 绩效提升技巧

鉴于数据量以惊人的速度增长，保持高级别系统性能的问题将推到前台，尤其是对于类别广泛的业务智能和分析用户社区，比如

- 标准的商业消费者，他们使用领域特定的报告以及其自身的特殊查询和直接的互动分析；
- 偶然的消费者，他们通常审核通过记分卡或仪表板呈现的初步设计报告中总结的职能或运营绩效指标；
- 运营分析消费者，其业务流程通过运营分析得到的嵌入结果来通知；
- 扩展的企业消费者，包括外部各方、客户、管制机构、外部业务分析师、合作伙伴、供应商或者对战术决策制定的报告信息有需求的任何方；以及
- 高级用户以及使用各种工作和技巧来分析海量数据的高级分析师，其结果将通知决策制定流程。

通常，不同的消费者及其典型使用反映对报告和分析的一般混合需求。结果，因为对这些不同类型分析结果的不同类型的需求，所以业务智能架构必须支持集中和预设计的报告、高级

用户准备的报务和分析、特殊查询、互动式分析以及到支持数据集的交互式深入的混合工作负载。

准备海量数据以及支持混合的报告工作负载、互动式分析以及复杂分析数据意味着不只是对问题下“猛药”。而是，一个人必须考虑信息产生流程的多少个方面可能会受到性能低下的影响，从数据流入分析环境的点，到海量数据必须可供报告、最终用户交互、切割和分块使用的点。

对性能的需求必须根据信息产生流程的四个方面进行评估： 在数据整合点，以及加载到数据仓库，因为数据要接受查询和算法分析，以交互方式将结果传送到用户社区和永久存储。我们可以考察这些用于提高性能的特定技巧：

- 分析数据库应用程序，经过系统设计，配有旨在提高处理速度的内置架构增强功能；
- 备选的编程模型（如 Hadoop），可以在算法分析与数据库分析之间实现协作；压缩，旨在减少存储量以缩短加载和查询的响应时间；以及
- 内存层次结构增强，如使用固态存储设备替代旋转磁盘。

5.4.1 大规模数据分析

最终用户不应仅限于访问过滤的和缩减的结果。将海量非结构化数据用来支持分析流程时，最终用户可能需要往回钻取到原始来源。但是，缺少上下文可能会导致不一致和混乱。通过维护关于其关系的上下文中收集的每个数据项目（如网页、演示文稿或视频）与其他项目的相关性和内容的历史信息，可以为分析提供一个维度集。此外，通过标记的意义基于定义的语义层次强化数据项目，将为审查提供额外的上下文，从而提供额外的维度。

这意味着分析环境必须能够将配置解决方案的各个方面应用于“大”问题。这包含对以下两者的组合：管理对海量数据项目的快速访问，将计算分析（如使用 Hadoop 进行的分析）的结果与数据仓库内用于报告和分析的结构化数据进行整合。这些功能将共同支持推动洞察力和可用知识发现的新兴高级分析服务类型。

5.4.2 架构改良： 分析数据库设备

由于数据仓库支持交易的需求极其有限，因此，数据仓库设计者愿意牺牲交易的性能，而提高复杂查询的性能。这一理念催生了“分析数据库设备”这一概念，该设备是一个系统，与

高性能计算资源、高速 I/O 和网络、数据库管理以及为实现混合工作负载报告和分析而预先安装、配置和优化的其他可能的工具进行了整合。

分析数据库设备常常包括层次处于海量并行处理系统顶部的经特殊优化的软件，尽管底层存储架构可以用不同方式配置。下面基于其存储配置来介绍三种常见的方法：

- 完全不共享，每个独立的处理单元连接到其自己的内存，并直接与其自己的磁盘系统进行通信。
- 共享磁盘，每个独立的处理单元连接到其自己的内存，但与完成不共享方法的不同之处在于所有处理单元都访问一个公共磁盘系统。
- 完全共享，每个处理单元均有权访问共享内存系统以及共享磁盘存储。

分析数据库设备也针对优化的数据分配进行了工程处理。不是传统的面向行的数据布局，而是某些设备使用柱状布局来组织和存储其数据。因为每一列均可单独存储，所以对于任何查询，系统可以评估哪些列将要被访问并因此将仅检索这些列请求的值。索引也可以得到简化，因为每列中的值可以进行排序以及用于建立索引，这不仅会调整数据库占用空间的扩展，而且还可以减少与次级存储之间的数据流量，从而可以动态提高查询性能。

5.4.3 利用交替编程模型整合复杂算法模型

进行高级分析的高级用户关注“大问题”，如分布式数字捣弄、复杂的统计分析、使用数据发掘工具开发模型以及大规模过滤和缩减（作为一种评估海量数据集的方法，通常是结构化数据和非结构化数据的混合，通常与操作系统集成）。这些用户可能采用各种用于不易并行的任务的分析技术，例如将数据仓库与数据发掘算法或精心制作的应用程序相结合。而这些应用程序的结果可能仍将需要通过传统的业务智能方法进行报告，因此，任何 BI 环境都应启用与标准数据仓库模型结合的复杂算法分析，尤其是在海量并行处理的架构中。

Hadoop 是此类交替编程环境的最佳示例，该环境经过结构化，可利用数据分布和大规模并行处理。Hadoop 是一个开放源代码的框架，最主要由两种服务组成。第一种是可靠的分布式文件系统，第二种是基于名为“MapReduce”的方法的并行编程模型。MapReduce 编程模型由 Google 研究员推出并进行描述，以进行涉及海量数据集（从数百 TB 到数百 PB）的并行分布式

计算。与常见的程序/命令式语言（如 Java 和 C++、MapReduce 的编程模型模仿功能语言（最显著的是 Lisp 和 APL））相反，主要应归于其对应用于数据值对集合或列表的两种基本操作的依赖性：

- 映射，描述应用于一系列输入键/值对，以生成一系列中间键/值对的计算或分析，以及
- 缩减，在此操作中，与由映射操作输出的中间键/值对关联的一系列值进行了组合，以提供结果。

凭借应用于海量数据集的一些应用程序，理论上，在映射阶段对每个输入键/值对应用的计算彼此独立。融合数据独立性和计算独立性意味着数据和计算可分步在多个存储和处理单元中，并可自动并行化。此可并行性允许程序员利用可扩展的海量并行处理资源，以提高处理速度和性能。算法的结果又可进而与数据仓库的维度模型重新保持一致，以便将算法分析和传统报告整合到一起。

5.4.4 利用压缩缓解存储需求

虽然磁盘存储的平均成本持续下降，但所有类型的数据爆炸式增长速率仍超出成本的相对减少。更糟糕的是，商业智能环境的存储需求超出了存储在数据仓库中的需求。由于索引、作业数据存储、中转区、测试环境和开发环境的使用，需要寻求减少存储占用空间，而不损失信息的方法。

实现此目标的一个通用技巧是压缩。压缩是通过识别共同模式并以较小型数据项进行替代来降低存储需求的过程。压缩在向您的数据仓库环境中加载数据时尤为重要。随着结构化数据集大幅增长，普通输入/输出 (I/O) 通道受限的带宽对及时将数据移动到分析框架中形成了束缚。或者依赖于多表连接的用户查询也可因受限的 I/O 带宽而受到束缚。通过压缩和应用创造性的数据布局和过滤减少存储占用空间，可减少对数据移动的需求，从而提高性能、降低成本并增强可持续性。

具有多种不同的压缩算法；一些示例包括行程编码（在此算法中，重复的数据项目与其相应的计数一起表示），以及按行对数据值应用 Lempel-Ziv 压缩。在后面的示例中，提取了共同模式并以压缩的数字值表示，然后在其每次出现时用该模式替代，从而降低存储需求。良好的压缩技巧可将存储降低 50%，并且常常可实现存储降低多达 80%。

5.4.5 通过改进内存层次结构提高性能

不同的计算架构通常在多个系统部分（如 CPU、缓存、核心内存、临时磁盘存储区域和永久磁盘存储）对最大化系统性能的贡献程度有所不同，尤其是在混合商业智能工作负载环境中。其中，这些存储模型组成“内存层次结构”，硬件架构师在不同的配置中采用此结构来找到不同大小、成本和速度的内存设备的正确组合，以通过减少响应日益复杂的询问和分析的延迟来提供最优的结果。

性能驱动型应用程序尝试使用内存层级结构，通过智能数据访问和缓存策略最小化延迟。这些改进的价值可因策略性地使用硬件而翻倍。例如，响应不同类型的查询要求写入临时内存空间，最常使用旋转磁盘作为介质。使用固态闪存进行临时存储可加快访问速度、减少 I/O 延迟、减少查询响应时间，并因此提高查询的吞吐量。

5.4.6 流计算

决策制定流程越来越多地涉及到种类不断增多的输入，这些输入中基于结构化数据的内容越来越少。或者，决策制定者消耗来自不同来源的多种不同种类的非结构化数据。手动扫描这些输入将对生产力形成重大瓶颈。相反，需要寻求不仅可快速扫描众多数据，还可连续实时分析流动数据的方法，以便可以过滤和减少适当的零散信息，并将其转换为在适当的时候向适当的个体提供的可用知识。随着更多信息源可用，这些方法必须适应不同来源、格式，甚至不同类型（包括文本、音频、图片和视频）的异构数据。

本质上，这要求有效的方法来分析数据流和允许分析发送的所有数据，以便在可推测出关键信息时可即时采取行动。还可选择性地保存数据，以进行后续处理。此类信息流分析高效地提供层积在查询引擎顶部的“数据联合”功能，其中的查询引擎应用于分析来自许多异构来源的海量动态数据。由数据流计算提供的分析可在刷新数据源后于加锁步骤中更新查询结果集来实现对事件和不断变化的环境的快速响应。本质上，可审查、过滤、清除和聚合多种（不同的）数据源，或遵循其他转换方式，以满足众多下游数据用户的需求。

5.5 新兴高级分析服务

除了可能被称为“主流”的分析外，还具有正快速整合到环境中，以支持普遍操作智能的新兴技术，如：

- 数据发掘和预测建模
- 嵌入式预测分析模型
- 实体识别和实体提取
- 文本分析
- 观点分析

5.5.1 数据发掘和预测建模

数据发掘和其他高级统计技术使分析师能够构建在一定程度上复制一些思维过程的模型，执行该思维模型可识别成功模式。业务分析师已应用数据发掘技术，而不预期识别担保以进一步确定业务价值的模式。

此无指引的分析不依赖于任何先决条件，并且可以是开始开发预测模型的良好方法。分析师利用数据发掘技术开发预测模型，并经过改进和培训，然后应用于非常大型的数据集，以识别与机遇或风险对应的模式。例如，群集客户数据是基于已发现的相似之处和差异对客户分组的未受指引的过程。评估源于群集客户数据的因变量可指导性地将新客户记录分类到已有的群集中。

有多种可结合以开发预测模型的数据发掘方法和技术，如：

- 群集，对项目进行分组，以使每个组彼此截然不同，且每个组中的成员有明显的相似性。
- 联想，对一批数据实例进行分析，查找可用于基于每个实例中其他值的出现来预测一系列值的出现的规则。
- 回归，是一种分析数据集的统计方法，可将一批数据点与数学公式匹配。而该公式又可用于插入值，实现对因变量的预测。
- 购物篮分析，是寻求对象的关系的过程，这些关系共同属于业务环境内，并且其名称源于分析超级市场顾客的手推车的内容，以查看是否有任何可用于实现业务优势的“天然”亲和性的观念。
- 基于案例的推理，使用已知的情况形成分析模型。将新情况与该模型进行对比，查找最匹配的情况，然后对其进行审查，以帮助确定分类或进行预测。
- 决策树分析，查看一批数据实例和给出的结果，评估一系列变量的频率的数值分布，并

以 树的形式构造决策模型。此决策树上每一层上的每个节点都代表一个问题，而此问题的每个可能的答案都由指向下一级上的节点的分支表示。

- 神经网络，使用概率统计方法来分析培训数据，以创建采用一定数量的输入并产生一些预测输出的“黑盒子”过程。

5.5.2 嵌入式预测分析模型

使用各种数据和文本发掘算法开发的预测模型可整合到业务流程中，通过已揭示的模型预测未来的事件或帮助实现特定的目标，从而补充运营决策制定以及策略分析。例如，通过应用群集分析创建的客户描述可用于根据特定的人口统计资料和其他特征进行实时分类。这些描述可用于推荐交叉销售和向上销售的机会，从而提高收入。嵌入式预测模型可用于解决我们所有的价值驱动因素，并且在许多不同的情景下使用，包括客户保留、收购和采购、供应链改进、欺诈建模、提高预测准确性、临床决策制定、信贷分析和自动承销。

5.5.3 实体识别和实体提取

非结构化项目日益泛滥的后果是在隔离文本中的关键词组（如人员、地点和实物）方面以及在这些概念间建立联系和关系方面的挑战。“实时”实体身份识别传统上是一种批处理，但与在线评论、客户服务、呼叫中心操作相关的时间关键型操作，或者涉及安全性、银行保密法案/反洗钱或其他“相关人员”应用的更敏感的活动在可实时识别个体身份时变得更为高效。

相关挑战超出了查找顺序文本中的名称模式，而是以自然语言处理概念为基础，以揭露关系（如个人对特定慈善团体的亲和性）、因果关系（如地理区域内各产品问题的相关性），或对相同实体的多个参考（如引如“他”、“它”等代词，指代采用适当代词的命名实体，如“George Washington”）。实时身份识别可实现个体与其相关属性、特征、概况和交易历史的快速联系，并可用于实时嵌入式预测模型，以改进运营决策制定。

5.5.4 文本分析

文本分析可用于隔离半结构化和非结构化文本内的关键字、短语和概念，并对这些关键文本项目进行了语义上的分析、建模，而其源文档根据认可的概念相互关联。这暗指对概念分类标准

的需求，在此标准中，可在不同的精度级别收集和聚合同类项，如汽车制作、模型，以及源于存在或缺乏特定功能的备用版本。

就文本而言，实体认知、实体提取和文本分析的算法需要更加复杂。虽然可以通过诸如正则表达式解析等技术扫描简单的、基于模式的未结构化实体（如电话号码），但是人们开始越来越多地使用更为复杂的模式和环境敏感型技术。OASIS 于 2009 年建立了名为未结构化信息管理架构 (UIMA) 的内容分析标准，该标准用于指导文本分析组件和规则集成框架的开发，以驱动未结构化分析软件的开发。

这些组件可让您通过其他技术执行常见术语分析，确定哨兵或信号术语，创建概念层次，创建字典以及记录短语认知和概念提取的规则。一旦完成此分析，文档内包含的信息会被立即聚集、分类和组织以支持智能搜索，而流式文本的过滤概念将有助于识别可直接路由到对所供应的内容特别感兴趣的个人的重要文本项目。一旦按顺序排列、识别和提取完这些概念，可将其用于数据发掘和其他类型的解析分析，以帮助工作人员从可操作的信息中得出结论。

5.5.5 情绪分析

社交媒体网站、联机网络和博客为广泛的人群提供张贴主观看法、产品评论、服务评级、体验和其他观点的充裕机会，这些通常都会影响到作者网络中的广泛的其他个人。与闲散地只是反映所呈现的内容相反，各组织通过快速处理负面情绪或利用积极情绪，力争利用这些持续增长的有影响力的网络。

情绪分析通过分析未结构化的文本来审查和评估该材料的作者态度上的主观性，从而将文本分析带到另一个层级。例如，产品制造商可能会分析呼叫中心经常出现的负面交互（如产品出故障）的零部件名称的报告。这样做有助于确定通用故障模式，从而可以在零部件出故障之前将前瞻性措施延伸至产品所有者。情绪分析还会带来其他商机，如确定新兴消费趋势、确定客户偏好或发现不满意的客户。这可让企业通过重点突显积极观点同时降低负面观点的影响来管理其网上声誉。

情绪分析是本白皮书中所讨论的大量技术的至高点 - 分析术语频率、推断分类标准和层次、用相应的概念标识标记文档项目、根据备用结构化数据组织概念，以及应用数据发掘分析找到模式、关联、因果关系和其他类型的关系。

5.6 分析交付服务

分析服务的范围支持整个组织数据使用者的多样性。

- **报告和特殊查询 –** 源自用户规格的标准静态报告提供业务特定方面的一致视图，批量生成并且通常通过标准 (web) 接口按计划的时间交付。报告的静态性质将驱动其他洞察备用方法的需要。一个方法是将所报告的数据提取到其他数据操控电子表格，同时允许特殊查询收集用于分析的其他数据。标准报告可以向广大消费者提供知识，即使这些消费者必须具备上下文知识才能识别关键标识并采取行动。然而，考虑到数据向 PB 字节的增长，标准报告正快速生成例外报告。

- **记分卡和仪表板 –** 如果需要受培训的人员从已扫描的报告扫描关键性能指标，简化关键性能指标的呈现可能会更好地使工作人员进行如下过渡：从看到已发生的事情到了解改善业务流程所需的改变。记分卡和仪表板会定制性能指标摘要的最新呈现，允许全天候的持续监控。普遍交付机制可将仪表板推向宽泛的渠道，从传统的基于浏览器的格式到手持移动设备。凭借仪表板的互动性质，工作人员可以研究有关任何新出现商机的关键指标，以及通过集成的流程流和通信引擎采取行动。

- **混合 –** 混合将仪表板带到了另一层级，可让知识使用者自身具有以下能力：确定特别适合其自身业务需要和目标的自身的分析和报告与外部数据流、新输入、社交网络和其他可视化框架中的 Web 2.0 资源的组合。混合框架提供集成数据流和业务智能与交互式业务应用程序的“黏性”。

- **多维分析和联机分析处理 (OLAP) –** 由 OLAP 工具提供的多维分析有助于分析不同变量（其自己的层次内）之间的“细微”关系，如“按时间，企业的收入是多少？”或“按地点和供应商，产品的可用性怎么样？” 使用“按”这个字表明查看数据所围绕的中心点，可让您查看先按时间段后按地区分类的销量，或者相反，先按地区后按时间段分类的销量。OLAP 可让分析人员以不同的维度反复研究层次，以便发现隐藏在层次内的依赖关系。

6 挑战

成熟的业务智能和分析程序将具备对这些技术组件的完整补充，以在整个分析过程支持知识使用者。在许多环境中，分析程序已有机地成长，拥有各种已获工具、内部开发的解决方案，与不同硬件、网络和软件（如数据库管理系统）选择集成，成为一个可操作的（甚至最高效的）解决方案。

大量供应商在支持互操作性（以便他们所有都能发挥良好）方面煞费苦心时，已设计 BI 系统很长一段时间，其拥有不同供应商提供的异构组件，较少考虑到组件集成的复杂性，更不用说性能或优化。事实上，伴随着速度要求的增长，我们认识到当各组织尝试“自制”其综合分析基础架构时存在一些潜在的挑战，包括以下因素：

- **有机发展和异构性 –** 发展的有机性质意味着已在“按需”基础上合并分析应用程序，既不需要综合程序计划，也不需要评估整个企业的业务需要。这导致基于发展决策（不与满足业务需要有关）的技术性依赖性，并且这些依赖性到时可能会阻碍灵活的端到端分析解决方案的成熟。

- **灵活性和可拓展性 –** 尽管许多供应商都在尝试实现互操作性，但是都受制于其与（通常发布的）产品版本（这些产品版本存在已公布的规格）之间良好协作的能力。事实上，这会施加严格的集成限制，可能会阻碍顾客启用所有可用的产品功能。例如，如果所选数据清除工具只对版本 5.7 的所选 ETL 产品起作用，消费者就不得升级到版本 6 的 ETL 产品，直到数据清除供应商增强其产品，以支持 ETL 产品升级。此外，随着业务需要、要求、分析预期或消费者数量发生变化，基本的分析基础架构将必须适应这些变化。这表明需要可轻松将功能添加到业务智能基础架构的能力。

- **数据质量 –** 即使数据管理和质量管理的浓缩度提高，仍然有可能引入数据缺陷。间断数据验证、不同数据质量工具（用于解析、标准化和清除）和冲突的规则设置仍会诱发数据缺陷和不一致。

- **实现价值的时间 –** 安装、测试和验证各种组件，并确保各组件都能良好运转，需要在计划、设计、实施和部署方面投入大量时间和资源。提升的实施和部署复杂性将会增加实现系统可被高效使用所需的时间。

- **性能和可伸缩性 –** 许多 BI 系统都受到其自身成功的贻害；随着用户数量的增加，查询负载也随之增加，或者随着要分析的数据量增加，系统的适当伸缩能力导致性能下降。此可伸缩性挑战只会在互操作性限制人为节制任何集成组件的性能潜能时增加。

- **快速解决方案 –** 必须重新设计相同（或类似）的报告和分析，将会耗尽工作效率和资源，并延长实现价值的时间。回顾过去，如果各组织合并了工具和技术，以支持常用的应用程序（如顾客洞察力和盈利能力、市场分析、开销分析和其他共同价值驱动因素），他们会获益良多。

7 总结： 应对挑战

当把所有这些挑战放在一起考虑时，将会出现一条共同的线索： 由于从大量来源集聚的零碎技术组件所带来的效率低下，最终会降低组织在必要时间向适当个人交付可操作性智能的能力，尤其是分析需求增加时，无论是由于增加的嵌入式操作智能、更大的数据量、增加的用户数量，还是（更可能是）这些需求的组合。

如果这些风险因素的根源是各种技术组件，那么降低这些风险需要考虑创建端到端解决方案的选项，这些解决方案被设计为最充分利用补充组件。一套完整的支持整个组织的报告和分析需要的工具需要应对这些挑战：

- 端到端解决方案 – 借助设计良好的架构，程序团队可以通过所有组件都设计为相互适合的综合解决方案，明确阐述满足业务需要的策略。从单一供应商选择一个完整的解决方案不仅会简化实施和部署，还会在简化采购流程的同时降低异构环境风险。
- 灵活性和可拓展性 – 单一供应商会提供更大的灵活性、特别是当以如下方式同步升级和发布时：确保功能改善不会人为受到产品版本控制的限制。此外，顾客可以通过在有策略程序计划的加锁步骤中部署新的升级或模块，按需引入功能。
- 数据质量 – 标准化数据验证、清除和增强工具以及这些工具的用法，会从其初始输入（或创建）到大量下游消费者提供预测水平的企业数据的一致性。
- 实现价值的时间 – 通过统一完整解决方案平台降低系统复杂性，会简化收购流程、降低有关实施、培训和部署的资源要求，从而加快实现价值的时间。
- 打包解决方案 – 在解决一般客户挑战方面有丰富经验的供应商能够将最佳实践集成到其打包解决方案，同时利用其技术组件来处理常用价值驱动因素，如通过交叉销售和向上销售的收入生成或者通过开销分析的成本降低。在既定业务部门或行业多年与客户打交道可让供应商定制其解决方案，以利用主题知识构建特定行业内满足业务需要的解决方案，如零售业务的定位、路由物流优化或保险行业的危险区域评估。

- 性能和可伸缩性 – 当设计的端到端解决方案是在特定硬件上运行时，开发人员可以利用大量直接集成到硬件和软件平台的优化，如工作负载管理、任务和流程安排、负载平衡、并行 I/O 渠道或高可用性。优化的分析数据库管理服务允许高性能的分析数据仓库，受到并行的数据集成和高速联合服务的支持。可以将增加的查询数量卸载到备用处理单元或路由到内存数据库，从而在降低 DBMS 负载的同时加快响应速率并增加吞吐量。

通过从建立于大量技术组件的企业业务智能和分析环境的有机演变到在策略上设计的端到端解决方案的转变，组织可以通过实时、集成的分析快速地实现价值，从而实现在正确的时间向适当的决策人员交付具有优势的智能。

关于作者

David Loshin 是 Knowledge Integrity, Inc, (www.knowledge-integrity.com) 的主席，也是公认的思想领导者和数据质量、主数据管理和业务智能方面的专家顾问。David 在 BI 最佳实践（专家渠道 www.b-eye-network.com）方面是个多产的作家，有大量有关 BI 和数据质量的书籍和文章。他的书《Business Intelligence: The Savvy Manager's Guide》（2003 年 6 月）被誉为可让领导者"了解业务智能、业务管理原则、数据仓库以及所有这些如何协同工作"的资源。他是《Master Data Management》 的作者，该书受到数据管理行业领导者的赞誉，最近出版的《The Practitioner's Guide to Data Quality Improvement》主要讲述改善信息实用性的实践过程。有关数据质量的更多洞察，请访问 http://dataqualitybook.com，有关主数据管理的观点，请访问 http://mdmbook.com。

可通过 loshin@knowledge-integrity.com 与 David 接洽。