火龙果•整理
uml.org.cn

INFORMATICA

避免大数据“一败涂地”

借助于 Informatica 9.5 平台，实现大数据的最大回报

白皮书

火龙果•整理
uml.org.cn

本文档含有 Informatica Corporation 的保密、专有信息和商业秘密信息（“机密信息”），事先未经 Informatica 的书面同意，不得进行拷贝、散发、复印或以任何其它方式复制。

尽管我们尽最大努力确保本文档中信息的准确性和完整性，但仍可能存在一些印刷错误或技术误差。如因使用本文档所含信息而造成任何损失，Informatica 概不负责。本文档中包含的信息随时可能更改，恕不另行通知。

Informatica 自行决定将这些材料中讨论的产品属性纳入其任何软件产品的发布或升级中，并自行决定任何此类发布或升级的时间安排。

受下列一项或多项美国专利保护：6,032,158；5,794,246；6,014,670；6,339,775；6,044,374；6,208,990；6,850,947；6,895,471；或受下列正在申请的美国专利保护：09/644,280；10/966,046；10/727,700.

此版本发布于 2012 年 4 月

火龙果•整理
uml.org.cn

目录

重塑计算的世界 2

三“V”要点：容量（Volume）、种类（Variety）
和速度（Velocity） 3

发挥大数据的业务价值 4

大数据：“一败涂地”的风险 4

Informatica 9.5 助您实现大数据的最大回报 5

提高数据的业务价值 6

降低数据成本 8

实现大数据的潜能 9

结论 11

火龙果•整理
uml.org.cn

重塑计算的世界

三大技术趋势正在重塑 IT 世界。云计算、社交计算和移动计算正在不断提供令人惊叹的新价值。但是，这些技术亦在迫使企业在 IT 部门采用新的方法，同时调整他们的传统流程和基础设施以适应新的现实情况和需求。

这些大趋势正在重新定义计算机行业的“地点”、“内容”和“方法”，并推动大数据的迅猛增长 — 更大容量、更多种类及更快速度。

- 计算“地点”正在从企业内部移往云计算，创造出新一波数据碎片，改变着计算的经济学。
- 计算“内容”正在从交易处理转为“交互处理”— 通过社交媒体的用户交互与通过“物联网”的设备交互 — 从而衍生出史无前例的庞大交互数据量。
- 计算“方法”则从台式机转移到各种移动设备，为用户提供具备上下文感知功能和基于位置的服务，并因此造成设备和传感器交互数据的激增。

庞大的新型数据能够带来前所未有的商机。企业能够通过掌控大数据提高运营效率，降低 IT 基础架构和数据管理成本，并更好地管理品牌和客户关系。不过，目前有能力充分利用大数据业务价值的企业屈指可数。假如企业无法适应不断增长的数据容量、种类和速度，还将招致无法预料的开支、复杂性及风险。

火龙果•整理
uml.org.cn

三“V”要点：容量（Volume）、种类（Variety）和速度（Velocity）

就字面而言，“大数据”这一术语有一点点误导。该术语仅暗示了容量方面的含义。容量无疑是大数据的一个关键方面，会对所有行业内的从业组织产生影响。研究公司 IDC 所进行的一项调查[1]显示，企业数据中心容量将在未来十年内增长 50 倍。但容量只是大数据现象的一个方面。数据种类和数据速度的并行增长具有同等重要性。请考虑下述信息：

- **数据种类：**社交媒体的增长 — 包括 LinkedIn 论坛、Facebook 发帖、与品牌相关的推文等，使得可向企业提供的数据类型成倍增长。同样不容忽视的是日益增长的移动计算和设备到设备的交互。数据不再是由交易应用程序生成的井然有序的行和列，而是一个杂乱无章的集合体，充斥着来自大量不同数据源的、无法预料的非结构化数据。
- **数据速度：**在我们的互联世界里，数据以不可思议的速度移动。亚秒数据移动成为企业实时检测事件并做出快速反应所需的首要能力。众多行业都面临着必须适应日益提高的数据传输速度及其对业务潜在影响的挑战，而金融交易、供应链、电信和电子商务不过是其中的一小部分。

数据容量、种类和速度增长（三“V”要点）是大数据（也就是大交易数据、大交互数据和大数据处理中）最常见的三个方面。

- **大交易数据：**企业数据库和应用程序中的传统交易数据继续稳步增长，拖累应用程序性能并增加基础设施成本。此外，要求企业加快响应实时事件的压力始终存在。最后，随着越来越多的企业向云系统迁移核心业务流程和相关数据，增加了其大交易数据环境的复杂性，交易数据变得更加支离破碎。
- **大交互数据：**防火墙内外都存在着人与设备交互所产生的社交媒体和设备数据。除 Facebook 发帖外，此类数据还包括呼叫详细记录（CDR）、地理空间数据、网络点击流，以及源自科学、医疗、供应链和其他行业传感器的信息。所有这些因素导致数据容量和数据种类出现史无前例的迅猛增长。
- **大数据处理：**面对来势汹汹的数据世界巨变，企业求助于专门面向数据密集型处理和分析的各种技术，如 Apache Hadoop 和 MapReduce。这些低成本平台已经改变了能够以经济、高效的方式计算的数据应用程序和分析学类型。不过，这类新型的定制专用环境有可能造成有问题的数据孤岛，还需要与传统系统进行互操作的数据集成策略。

[1] IDC, “The 2011 Digital Universe Study: Extracting Value from the Chaos,” 2011 年 6 月。

火龙果•整理
uml.org.cn

发挥大数据的业务价值

大数据被誉为“下一代创新、竞争力与生产率前沿”[2]。大数据可为您的企业提供前所未有的机遇，让您提高营收和绩效，并洞察您的业务、市场和客户。社交媒体数据就是一个很好的例子。借助于可从 Facebook、Twitter、Foursquare、Google+ 和其他来源搜集信息的技术，您的企业将能优化品牌管理、打开进入市场的新渠道并加强客户关系。

传统的交易数据可提供有关客户状况的重要视图，但这一视图并不完整。由于社交媒体数据中包含了客户喜好、情感、关系、行为及其他通常唯有社交媒体才能提供的信息，因此能够提供更具洞察力、更有价值的视图，供您预测未来发展。数据集成则是用于发挥社交媒体数据潜能的支持技术。

除社交媒体之外，大数据机会还存在于各行各业中。在制造业和供应链行业中，实时传感器数据流可帮助公司优化物流、库存和生产，同时快速识别在哪些方面出现了问题。对电信业来说，通过掌控 CDR、文本短信和移动设备数据，可减少客户流失并优化网络。金融服务机构能够根据实时市场事件采取行动，更好地管理金融风险。医疗机构和其他组织则能够以来自新数据源的信息充实传统的数据仓库，由此实现更加锐利的分析洞察能力。

大数据：“一败涂地”的风险

为充分发挥大数据的业务价值，需要一个可扩展、灵活而可管理的数据基础架构，但大多数企业缺乏这一基础架构。在实施正确流程和系统上走在前列的企业将从错失这一良机的竞争对手那里赢得可观的竞争优势。如果您的企业缺乏始终如一的大数据策略，即可能面临下述风险：

- 数据碎片散布于日益复杂的 IT 环境中
- 依赖不准确或重复的数据而导致不良决策
- 高昂的人工、软件和基础设施成本
- 错失创收与客户参与机遇
- 多余的流程和重复利用的缺乏造成成本效益低下
- 承担不必要的合规性、运营效率和财务绩效风险

据研究公司 Gartner 的预测，大部分企业的表现远不如人意。“到 2015 年，超过 85% 的财富 500 强企业将无法有效地利用大数据来获得竞争优势。”Gartner 指出，“大多数企业均未做好应对大数据所带来的技术和管理挑战的准备，其直接后果就是，能够有效地利用这一趋势来获得竞争优势的企业将寥寥无几。[3]”

[2] McKinsey Global Institute, “Big Data: The Next Frontier for innovation, Competition, and Productivity,” 2011 年 5 月。

[3] Gartner, “Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond,” 新闻稿，2011 年 12 月 1 日。

火龙果•整理
uml.org.cn

Informatica 9.5 助您实现大数据的最大回报

作为最新推出的 Informatica 平台，Informatica® 9.5 将助力您实现大数据的最大回报 — 无论您所制订的战略目标在于应对数据容量、种类和速度的大规模增长、从更为传统的数据集中萃取更大的业务价值，还是组合新型与常规数据以获得业务优势。

Informatica 9.5 拥有令数据变得更相关、更可信、更及时、更具权威性、更可行、更易访问、更全面及更安全的创新功能，从而可在多个方面提升数据价值。例如，可内嵌式云服务可确保向云应用程序提供所有相关业务数据。用于社交媒体主数据管理（MDM）的数据时间线则能够创建完整、权威的长期 4D 客户视图，提高客户招揽率和维系率。

Informatica 9.5 同时还以多种方式降低数据成本，减少业务、人工、软件、硬件和存储开支。例如，其 Hadoop MapReduce 处理功能可加快低成本计算平台的采用速度，从而减少硬件支出。最后，Informatica 9.5 能够通过提供创新的大数据处理技术，让您实现大数据的潜能。譬如，缺乏 Hadoop 专业技能的用户可借助于供 Hadoop 使用的统一可视化开发环境部署 Hadoop 系统，通过重复利用现有技能和资产加快采用速度。

Informatica 9.5 的所有全新功能均作为同样众所周知的 Informatica 平台的组成部分提供，让您能够扩展现有的技能集、操作规程和管理流程，同时允许您将重点放在新型数据问题上，却不会带来额外的技术复杂性。

即使在快速演进的大数据时代，拥有长达 19 年数据集成创新经验和领导才能的 Informatica 依然是业内唯一能够确保您实现数据最大回报的供应商。借助于 Informatica 9.5，您所能做到的将远不止于勉强应对大数据时代。您将能做到：

- 提高数据的业务价值
- 降低数据成本
- 实现大数据的潜能

客户快照： T-MOBILE

电信服务供应商 T-Mobile 采用 Informatica 平台在不同的联合体系架构中集成数据，其中包括一项基于 CDR、网络日志、帐单数据、社交媒体信息及更多信息来支持高级客户流失分析的 Hadoop 实施。通过融合大交易数据和大交互数据，T-Mobile 获得了更为完整的客户视图以及客户流失的原因，从而能够在短短一个季度内将流失率减半。

“到 2015 年，超过 85% 的财富 500 强企业将无法有效地利用大数据来获得竞争优势。大多数企业均未做好应对大数据所带来的技术和管理挑战的准备，其直接后果就是，能够有效地利用这一趋势来获得竞争优势的企业将寥寥无几。”

– Gartner

火龙果•整理
uml.org.cn

提高数据的业务价值

Informatica 9.5 旨在帮助您的企业在从及时性到可信性等八个方面提高数据的业务价值。

价值	功能
及时性	数据流功能可实时处理大数据事务，以保持竞争优势
可行性	主动监控功能可发出警报，最大程度地利用各种机遇并将风险降至最低
可访问性	Hadoop 解析功能可让业务部门一致地访问所有的复杂数据类型
相关性	可内嵌式云服务确保向应用程序提供所有相关数据
整体性	数据虚拟化功能可集成现有企业信息与大数据，提供全新洞察力
安全性	普遍数据隐私功能可不分位置地保护敏感数据，降低风险
可信性	整体数据治理功能可提供一致、准确的数据，激发业务部门的信心
权威性	数据时间线可对任一时段的数据变更进行跟踪，提供有关客户、合作伙伴和产品的 4D 视图

数据流 | 及时性

随着在线设备的不断增多，企业面临的实时事件和处理数据呈指数级增长。在多种能够通过搜集来自设备、应用程序、数据存储器和传感器以及高速交易系统的数据流获益的行业中，金融服务、制造业、公用事业、交通运输和供应链企业就是其中的范例。利用 Informatica 9.5 平台的数据流功能，企业可拥有一个无代理的消息收发体系架构，辅以具备变更数据捕获功能的实时数据复制技术，对流数据进行捕获和分发。

主动监控 | 可行性

随着数据容量和复杂性的增长，发生可影响到业务运营与服务层协议（SLA）错误的风险也在节节攀升。主动监控可收集实时与历史信息，识别兴趣模式，并向个人、团体、仪表板或其他系统分发通知，以便迅速解决问题。这些功能帮助企业最大程度地利用创收机遇，并在从 IT 领域到客户服务的广泛范围内尽量减轻各类问题造成的影响。

Hadoop 解析 | 可访问性

当 Hadoop 成为企业的一部分，许多组织发现他们不得不使用昂贵而繁琐的脚本方式处理 Hadoop 中的数据并将这些数据集成到传统的应用程序内。Informatica 9.5 引入了第一种专为 Hadoop 优化的数据转换环境，从而实现了复杂或非结构化数据与关系格式之间的相互转换。 开发人员可运用 Hadoop 解析功能发掘用于 Hadoop 的复杂数据类型的价值，这包括事件数据、CDR、网络、应用程序和网络日志及更多其他类型。

可内嵌式云服务 | 相关性

对于云应用程序提供商来说，他们应该将专业技能和才干集中用于应用程序本身，而不是担忧如何将数据与他们的应用程序集成。Informatica 提供了能够在云应用程序内无缝使用的可内嵌式数据集成云服务。这些服务通过即开即用的连接功能与预建模板处理棘手的数据集成难题，让云应用程序提供商得以将重点放在其专长的核心领域，同时依然能够向他们的业务用户提供相关信息。

火龙果•整理
uml.org.cn

数据虚拟化 | 整体性

由于复杂的基础设施和多个系统，尽管对于及时、完整数据的业务需求日益升高，但企业仍然难以按时交付数据。数据虚拟化可提供一个掩盖基层数据源复杂性的抽取层面，以及用于数据集成的全面联合环境。数据虚拟化通过为各个系统提供可重复利用的数据服务，促进商业智能，为及时进行销售和市场营销提供支持，并为 Hadoop 与传统系统之间架起沟通之桥。

普遍数据隐私 | 安全性

传统系统和更加新型的离岸、云及Hadoop 环境内的数据蔓延加大了保护敏感信息的风险和必要性。普遍数据隐私功能将可让企业确保遵守隐私法规，降低数据侵害或无授权访问生产或非生产数据的风险。自动发现敏感数据、用于定义隐私政策的统一控制台以及针对生产和非生产系统中的屏蔽数据进行审计跟踪，将能够缩短应用程序的部署时间并确保合规性。

整体数据治理 | 可信性

大数据可增加数据管理流程的复杂性，导致难以进行监控和管理，从而增进数据治理的复杂性。整体数据治理功能将赋予技术人员针对数据和元数据的可视性与控制力，提供可重复利用的数据质量、隐私和保留规则，以及一个用于交付准确、一致、安全和及时信息的单一解决方案。数据管理工作流、数据治理仪表盘、数据沿袭和影响分析均有助于简化治理过程，而语义元数据功能则能够增强业务与 IT 部门的合作关系。

数据时间线 | 权威性

数据随业务变化而变化。为主要业务实体跟踪和预测这些变化的能力，是有效管理运营活动并降低风险的关键。Informatica 9.5 平台的数据时间线功能可用于在任一选定的时间点（不论是过去、现在还是未来）查看和管理实体版本及层次结构。该功能还提供对随着时间推移所产生的所有变更的可视性，包括执行操作的时间、执行人以及受到影响的数据。

客户快照：US XPRESS

货运公司 US Xpress 实施了一个名为“不放过任何数据”的长期计划，从数以万计的货车运输系统中收集 900 项数据元素，其中包括轮胎、耗油量和引擎运行的传感器数据、车队跟踪状态的地理空间数据，以及在货车司机博客上的抱怨帖子。US Xpress 利用 Hadoop 和 Informatica 技术对这种大数据进行处理和分析，优化车队使用率、减少怠速时间和燃油消耗，每年节省数百万美元。

火龙果•整理
uml.org.cn

降低数据成本

Informatica 所提供的广泛数据技术可帮助您的企业将业务、人工、软件、硬件和存储成本降至最低。

重点领域	功能
业务成本	数据验证可通过防止坏数据避免做出不良业务决策
人工成本	数据发现可通过自动发现业务实体和敏感数据提高生产率
软件成本	混合型 IT 平台可让您在获取云经济优势的同时，维持对所有数据的控制权
硬件成本	Hadoop 处理功能可加快低成本平台的采用速度，减少硬件支出
存储成本	智能分区可降低用于 Hadoop、设备和关系数据的基础设施成本

数据验证 | 业务成本

不一致、不准确的数据可能导致错误的业务决策并错失商机，造成具有负面财务影响的不必要开支和收入损失。数据验证通过确保数据的可信度和完整性以制订出最佳决策，提高盈利能力并降低业务成本。开发人员可轻松地创建强大的自动数据测试与验证规则。Informatica 9.5 提供了可在将更新数据载入生产系统之前对其进行核对的强化功能，从而消除了坏数据，确保提供满足业务需求的信息。

数据发现 | 人工成本

随着数据源和复杂性的持续增长，企业经常不得不采用手动的人工密集型数据发现和探查方法，以了解企业信息及其与业务的关联性。Informatica 9.5 所拥有的企业数据发现技术可在数千份图表中自动发现数据和元数据结构、关系及域，只需手动方式成本的几分之一就能更快地实现业务价值。数据发现能够突出显示结构化和非结构化信息中的异常情况，检测差异，降低风险。

混合型 IT 平台 | 软件成本

如果企业必须通过多余的软件和资源分别管理企业内部和云系统中的数据，则非但无法享受到云计算的经济优势，还可能承担产生更多不一致的数据孤岛的风险。Informatica 为在企业内部和云环境组合中无缝集成、清洗和管理信息提供了单一的混合型 IT 平台，帮助企业治理数据并充分利用云模式所具备的成本效益。

Hadoop 和 MapReduce 处理 | 硬件成本

传统的网格计算基础设施不仅昂贵，而且在其扩展至太字节甚至拍字节的数据量时，可能大幅增加网络成本。Informatica 9.5 凭其在 Hadoop 系统中进行本地数据探查、处理和转换以及自动转化规则并将其映射到 MapReduce 语言的能力，开辟了一条低成本 Hadoop 和 MapReduce 系统的采用途径。企业可避免不必要的硬件投资并最大程度地提高 Hadoop 分布式框架的价值。

火龙果•整理
uml.org.cn

智能分区 | 存储成本

随着大数据在传统环境、Hadoop 环境和云环境中的剧增，对数据增长的有效管理成为控制存储成本和提高绩效的关键。Informatica 9.5 所具备的智能分区功能是一个全面信息生命周期管理解决方案的一部分，用于自动实施复杂分区、应用程序和数据库归档策略。数据库和应用程序管理员能够制订连贯的信息生命周期政策，对分区和归档操作进行集中治理，从而降低存储、基础设施和维护成本，并确保达成应用程序 SLA。

实现大数据的潜能

通过采用 Informatica 9.5，您的企业将能实现大数据的潜能，以更快速、更明智的方式开展业务。

重点领域	功能
大交易数据	通用数据复制功能可向 Hadoop 分析处理提供交易数据，以降低处理成本
大交互数据	社交媒体 MDM 将社交客户资料与现有的客户数据集成，提高客户招揽率和维系率
大数据处理	Hadoop 本地转换可充分发挥 Hadoop 的能力，加快实现业务价值的速度
大数据处理	专供 Hadoop 使用的统一可视化开发环境能够创建数据业务逻辑，无需专业的 Hadoop 资源

通用数据复制 | 大交易数据

为发挥企业信息的价值，在传统和新型大数据系统内实时复制数据至关重要。Informatica 9.5 为所有类型的数据复制提供了单一平台，在包含 Hadoop、云平台和应用程序的混合环境中实现了实时数据仓库和分析。通过利用通用复制功能向 Hadoop 交付大量交易数据，Hadoop 可作为一个信息汇聚点，从而减少数据仓库的存储规模和成本。

社交媒体 MDM | 大交互数据

社交媒体上的消费者互动为寻求单一客户视图以支持销售和市场营销添加了一道新环节。Informatica 9.5 提供了社交媒体 MDM 功能，用于集成源自社交媒体数据的客户资料并将此类社交资料与现有客户数据整合。企业可因此获得唯有社交媒体上才能得到的对客户喜好与行为的洞察力。该功能与数据时间线相辅相成，使得 Informatica 9.5 能够提供完整的 4D 客户视图，以便提高客户招揽率、维系率及顾客份额。

客户快照： HEALTHNOW

位于纽约州的医疗保险和服务提供商 HealthNow 运用 Informatica 的数据虚拟化功能精简其横跨 16 个企业数据源、以服务为导向的体系架构，降低了数据基础设施的复杂性，并提供单一、可信的视图。这使得该公司能够充分利用包括源自社交网络的数据以及索赔单和病历中的非结构化数据在内的大数据，为针对医保问题的预测分析提供支持。

火龙果•整理
uml.org.cn

“西太平洋银行正开始利用社交媒体数据来提升市场促销活动的成功率，以及提高客户招揽率和维系率。Informatica 9.5 平台为通过 Hadoop 管理和处理社交媒体数据以增进对客户的了解并实现大数据的最大回报提供了多种全新功能，我们对此备感振奋。”

– Westpac Information Technology 信息管理系统经理 Torrance Mayberry

Hadoop 本地转换 | 大数据处理

对 Hadoop 开发人员来说，在 Hadoop 集群上定制和调整数据转换逻辑十分困难、耗时且昂贵。Informatica 9.5 为开发人员提供了即开即用的 Hadoop 本地转换功能，其中包括身份识别和数据解析、探查、匹配及清洗。企业可运用这些创新技术来降低开发成本并快速释放 Hadoop 数据供全企业使用，实现所有信息的最大回报。

专供 Hadoop 使用的统一可视化开发环境 | 大数据处理

在 Hadoop 系统上实施大数据处理可能需要高度专业化且价格昂贵的开发资源。Informatica 9.5 平台所具备的专供 Hadoop 使用的统一可视化开发环境让 Informatica 开发人员能够在无需专业培训的情况下，轻松地在 Hadoop 上大展身手。借助于该技术，开发人员可使用熟悉而直观的 Informatica 开发工具定义业务逻辑、转换和映射规则，从而降低成本、优化资源利用率并充分发挥大数据的潜能。

火龙果•整理
uml.org.cn

结论

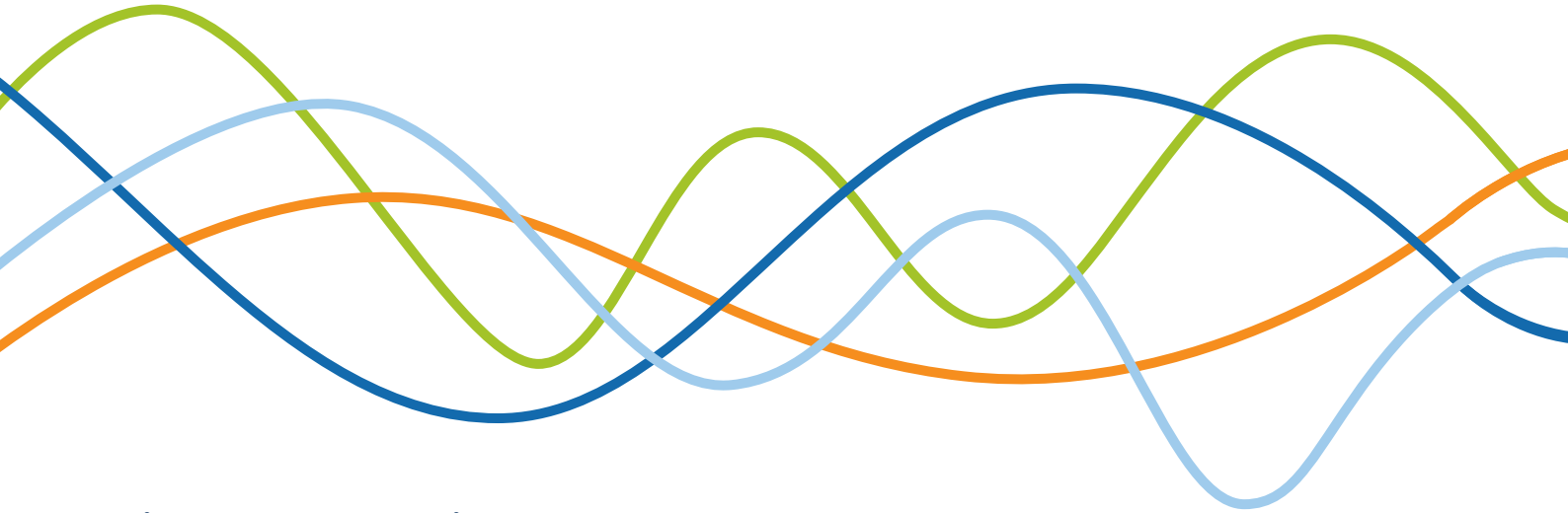
变化是永恒的主题。由云计算、社交计算和移动计算等三大趋势推动的大数据正在重塑业务流程、IT 基础设施以及我们对于企业、客户和机器生成信息的捕获与使用方式。但企业成功地适应大数据时代带来的新环境却绝非理所当然之事。

为确保您的企业取得成功，您需要对数据挑战和机遇做出战略性评估，并向一个能够根据持续变化的业务需求灵活扩展的领先数据基础设施过渡。Informatica 9.5 所拥有的全新功能和更高的集成度能够提供您的企业提升价值、降低成本及充分发挥大数据潜能所需的创新技术。

关于 INFORMATICA

全球领先的独立企业数据集成软件提供商 Informatica Corporation（纳斯达克代码：INFA）助力世界各地的企业获得最大化的数据投资回报，从而赢得竞争优势。全球有超过 4,630 家企业依赖 Informatica 平台，使用及管理其在企业内部、云中和社交网络上的信息资产。有关详细信息，请致电 +86-10-5879-3366（在中国也可拨打 1-800-653-3871）或登陆 Informatica 公司网站 www.informatica.com/cn。欲了解更多Informatica 信息，请点击：http://www.facebook.com/InformaticaCorporation、http://www.linkedin.com/company/informatica 和 http://twitter.com/InformaticaCorp。

火龙果•整理
uml.org.cn



INFORMATICA®

www.informatica.com.cn

北京办事处
地址：北京市朝阳区建国门外大街乙 12 号，LG 双子座大厦
东塔 19 层 1906 室　　邮编：100022
电话：86-10-5879 3366　　传真：86-10-5879 3130

上海办事处
地址：上海市浦东世纪大道 201 号渣打银行大厦 5 楼
邮编：200120
电话：86-21-6182 6825　　传真：86-21-6182 6755

© 2012 Informatica Corporation。 保留所有权利。 美国印刷。Informatica、Informatica 徽标、The Data Integration Company 是 Informatica 公司在美国和全世界司法区域的商标或注册商标。 所有其它公司和产品名称都可能是它们各自所有者的商业名称或商标。

IN09_0412_01942