INTELLIGENT
BUSINESS
STRATEGIES

白皮书

设计用于分析大数据平台的架构

作者：Mike Ferguson
Intelligent Business Strategies
2012 年 10 月

适 用 对 象 ：

目录

简介 4
分析新数据源的业务需求 4

工作负载复杂性的增加 5
数据复杂性的增加 5
各种数据类型 5
数据量 5
数据生成的速度 5
分析复杂性的增加 6

什么是大数据？ 7
大数据的类型 7
为何需要分析大数据？ 8
大数据分析应用 8

大数据分析工作负载 10
分析动态数据以制定运营决策 10
非模式化、多结构化数据的探索式分析 11
结构化数据的复杂分析 12
存档数据的存储、预处理和查询 13
加速结构化数据和非模式化数据的 ETL 处理 13

端到端大数据分析的技术选项 14
用于动态大数据的事件流处理软件 14
静止大数据分析的存储选项 14
分析 RDBMS 设备 15
Hadoop 解决方案 15
NoSQL DBMSs 16
哪种存储选项最为合理？ 16
静止大数据的可伸缩数据管理选项 17
大数据分析选项 18

将大数据集成到您的传统 DW/BI 环境中 20
新型企业分析生态系统 20
接合分析处理 – 工作流的力量 21
新型分析生态系统的技术要求 22

入门：企业的大数据分析战略 24
业务协调 24
工作负载与分析平台的协调 24
技能集 24
为数据科学和探索搭建环境 24
定义分析模式和工作流 25
通过集成技术过渡到大数据企业 25

供应商示例：IBM 的端到端大数据解决方案 26
IBM InfoSphere Streams – – 分析动态大数据 27
支持分析静态数据的 IBM 设备 28
IBM InfoSphere BigInsights 28
IBM PureData System for Analytics（采用 Netezza 技术） 29
IBM PureData System for Operational Analytics 29
IBM大数据平台加速器 30
IBM DB2 分析加速器 (IDAA) 30
面向大数据企业的 IBM 信息管理 30
面向大数据企业的 IBM 分析工具 31
IBM BigSheetsuda 31
IBM Cognos 10 31
IBM Cognos Consumer Insight (CCI) 32
IBM SPSS 32
IBM Vivisimo 33
这些组件如何融合在一起以实现端到端的业务洞察 33

结束语 34

简介

多年来，组织始终通过构建数据仓库来分析业务活动

多年来，企业始终通过构建数据仓库来分析业务活动，获得供决策制定者采取业务绩效提升措施的洞察。这些*传统*分析系统通常基于经典模式，即从多个运营系统中捕获数据，并对这些数据加以清理、转换和集成，随后再将其加载到数据仓库中。通常，组织将建立多年的业务活动历史，以便运用商业智能 **(BI)** 工具来分析、对比和报告长期业务绩效。除此之外，组织通常还会从数据仓库中提取这些数据的子集，并将其置入已为更详细的多维分析而优化的数据市场中。

***BI** 市场日趋成熟，但 **BI** 仍然处于 **IT** 投资的前沿*

如今，数据仓库和 **BI** 出现已有二十余年。这段时间以来，许多企业已经在其不同的业务部分中构建了众多数据仓库和数据市场。尽管 **BI** 市场日趋成熟，但 **BI** 仍然处于 **IT** 投资的前沿。这种要求在很大程度上可以归因于人们创建的数据越来越多。但是，企业也在发生变化，已经从凭直觉运营转变为根据详尽的事实信息运营。在这个动荡的市场中，随着分析关系数据库技术的发展以及移动和协作式 **BI** 的兴起，软件技术也在不断改进。

分析新数据源的业务需求

更为复杂的新数据已经出现，而且生成的速度达到了前所未有的程度

尽管传统环境不断发展，但如今出现了许多更为复杂的新数据类型，企业需要分析这些数据类型，以便充实其已知信息。此外，这些新数据的生成速度远远超过了以往的纪录。

客户和潜在客户正在社交网络和评论网站中创建大量的新数据。此外，在线新闻项目、气象数据、竞争对手网站内容，甚至是数据市场如今都已经成为可供企业使用的候选数据源。

社交网络数据、网络日志、存档数据和传感器数据都属于人们在分析中关注的新数据源

在企业内部，随着客户转变为以在线渠道作为开展商业交易及与企业互动的首选方法，网络日志也在不断增加。分析所用的存档数据再次增多，为监测和优化业务运营而部署的传感器网络和机器数量也越来越多。结果就生成了大量新数据源、快速增加的数据量和迅速增加的新数据流，需要分析所有这些新数据。

工作负载复杂性的增加

数据和分析工作负载的复杂性正在增加

观察所有这些新数据源，可以明确的是，无论是就数据本身的特征而言，还是就企业希望执行的分析类型而言，复杂性都在增加。

数据复杂性的增加

就数据而言，复杂性主要是通过三种途径增加的：

- 企业所捕获的各种数据类型
- 企业所捕获的数据量
- 数据生成的速度或速率
- 数据的精确性或可信性

各种数据类型

除了“正常”捕获主数据和事务数据之外，企业现在还会捕获新的数据类型。这其中包括：

正在捕获新的数据类型

- 半结构化数据，例如，电子邮件、电子表格、**HTML**、**XML**
- 非结构化数据，例如，文档集合（文本）、社交互动、图片、视频和声音
- 传感器数据和机器生成数据

其中大部分数据都是非模式化的

这一系列更为复杂的全新数据类型通常也被称为多结构化数据。多结构化数据的一个主要问题是这些数据往往是非模式化的，因此必须加以“探索”，才能从中得出具有商业价值的结构化数据。因此，通常必须在传统分析环境的上游对多结构化数据执行调查分析，以便识别可能充分实现有数据仓库内已存储内容的数据。此外，还可能需要对此数据（比如石油与天然气中的地震数据）执行独立高级分析研究。

必须首先通过调查分析来确定其结构，之后才能将其引入数据仓库

数据量

某些新数据源的数据量也非常庞大

除了新数据类型造成的复杂性之外，企业收集数据的速度也在加快，这造成了数据量的进一步增加。示例包括文档和电子邮件、**web** 内容、电信业呼叫数据记录 **(CDR)**、网络日志数据和机器生成数据的集合。这些数据源可能要占用数百 **TB** 乃至数 **PB** 的空间。

数据生成的速度

数据创建的速度也在不断加快

数据的生成速度也在快速增加。金融市场数据就是一个很好的示例，这些数据以极快的速度生成和发出，必须立即分析这些数据才能及时响应市场变化。其他示例包含传感器数据和机器生成数据，此时的需求与之前相同，摄像头还可能需要视频和图像分析。

分析复杂性的增加

分析复杂性也在增加

在分析复杂性方面，现在需要利用新型算法和多种类型的分析来生成解决业务问题所需的必要洞察。此外，需要对具有不同种类、数量和速度特征的数据执行这些分析。零售市场就是一个很好的示例，在这种市场中，人们往往使用移动设备保持在线，因此需要改善在线渠道的市场活动准确性和及时性。这意味着需要更详细的客户洞察。在这种情况下，可能需要：

可能需要利用多种类型的分析来解决业务问题

- 对客户人口统计数据和客户购买交易活动（结构化数据）执行历史分析和报告，以确定客户细分市场和购买行为
- 市场购物篮分析，确定可共同销售的产品，以识别各客户的交叉销售机遇
- 点击流分析，以理解客户在浏览网站内容时的在线行为和产品查看模式，从而实时生成准确的追加销售服务
- 分析用户生成的社交网络数据，比如个人资料（如 **Facebook**、**LinkedIn**）、产品评论、评分、喜好、反感、评论和客户服务交互等
- 实时分析客户手机位置服务 **(GPS)** 数据，以检测可能位于店铺附近的客户，并为其提供针对性优惠，吸引客户进入店铺

在许多情况下，确定所需的洞察现在已经成为一个涉及多种分析类型的流程

重点在于，在许多情况下，确定解决业务问题所需的洞察现在已经成为一个涉及多种不同数据源分析的流程，而且数据和分析具有不同的复杂性。此外，在任何分析流程中都可能需要分析结构化和非结构化数据，以便生成所需的洞察。必须利用数据集成来合并多模式数据，以改进可靠的洞察。

此外，考虑到某些数据源可能是非模式化的，因此分析流程中的步骤无法全部在单一分析平台上完成，而是需要利用多种基本技术来解决业务问题。

在一个分析流程中，并非所有分析始终都可以在单独一种平台上完成

尽管存在这些复杂性，但仍然亟需分析许多业务领域中不断增加的此类新数据类型以及现有的传统数据。一个常见的示例是分析社交网络数据，以便理解客户舆情、社交图和具有影响力的人员，以便补充现有客户资料或客户细分数据。

什么是大数据？

分析工作负载的范围如今极为广泛，以至于无法通过单一企业数据仓库加以处理

新数据源的兴起和实时分析包括实时数据流和大量非结构化内容在内的一切内容让许多企业认识到，他们目前已经进入了一个全新的时代：分析工作负载的范围极为广泛，以至于无法通过单一企业数据仓库加以处理。而这还不是全部。尽管数据仓库是分析环境中非常重要的一部分，但如今的业务需求表明，需要一种更为复杂的全新分析环境，以支持传统环境无法轻易支持的一系列分析工作负载。

如今需要全新、扩展的分析环境

除了数据仓库之外，这种新型环境还包含多种底层技术平台，每种平台均为特定的分析工作负载而优化。此外，它应该能够独立地为特定工作负载利用这些平台，也能结合利用这些平台来解决业务问题。如今的目标在于应对全范围的分析工作负载。这包括传统工作负载和新型“大数据”分析工作负载。

大数据是一个与传统环境无法轻易支持的新型工作负载相关的术语

因此，大数据是一个与解决过去因技术限制和/或过高的成本而无法解决的业务问题所需的新型工作负载和基本技术相关的术语。

因此，大数据不仅仅与数据量有关

因此，大数据不仅仅与数据量有关。它可能与数据量中等但数据种类（数据和分析复杂性）极高的数据相关。大数据分析的主旨在于：与数据量、数据速度和数据种类（可能包含复杂的分析和复杂的数据类型）的某种组合相关的分析工作负载。因此，大数据可能与结构化和多结构化数据相关，而不仅限于后者。正因如此，大数据分析可能包含传统数据仓库环境，因为某些分析工作负载可能需要同时使用传统平台和针对工作负载优化的平台来解决业务问题。新型企业分析环境包含传统的数据仓库和其他最适合某种分析工作负载的分析平台。大数据不能取代数据仓库。实际上，数据仓库是扩展分析环境的一个组成部分。

大数据可能与结构化和多结构化数据相关

数据仓库是扩展分析环境的一个组成部分

分析需求和数据特征将表明需要部署的技术

分析需求和数据特征将表明大数据环境内需要部署的技术。因此，可以在多种技术平台上实施大数据解决方案，包括流式处理引擎、关系 **DBMS**、分析 **DBMS**（例如，大规模并行数据仓库设备）或非关系数据管理平台（比如，商业化的 **Hadoop** 平台或者专业化的 **NoSQL** 数据存储 — 例如图形数据库）。更重要的是，它可能会结合所有这一切，以支持业务需求。关系 **DBMS** 技术当然并非不适用于大数据分析。

大数据的类型

通常与大数据分析项目相关的数据类型包括 **Web** 数据、特定于行业的事务数据、机器生成/传感器数据和文本。

***Web** 日志和社交网络交互数据*

Web 数据包括 **Web** 日志数据、电子商务日志和社交网络交互数据，例如 **Twitter** 流。

海量事务数据

特定于行业的事务数据示例包括电信业的呼叫数据记录 **(CDR)** 和地理位置数据、零售交易数据和制药业的药品实验数据。

文本

传感器数据

机器生成/传感器数据是增长最快的领域之一。如今，传感器监控着生活的方方面面，包括移动、温度、光线、振动、位置（例如，智能手机中的传感器）、气流、液体流动和压力。此外，我们也注意到有越来越多的产品采用了生成数

据的电子元件，所有这些产品都能连接到互联网，将数据传回给收集者和指挥中心。“物联网”的时代已经到来。

在非结构化内容的世界中，文本是最普遍的分析目标。如今，许多企业已经开始认识到，文本中蕴藏着重要价值，无论是存档文档、外部内容源，还是客户交互数据中的文本都是如此。过去的技术局限性妨碍或限制了此类数据的分析。但是，随着障碍的消除，文本分析逐渐成为高优先级的分析项目，舆情分析就是一个很好的示例。**此外，企业如今通过收集数据来避免未来的债务。**例如，石油与天然气企业要收集跨度为 **20** 年至 **30** 年的数据，以捕获运营前、运营中和运营后的环境数据。

企业如今通过收集数据来避免未来的债务

为何需要分析大数据？

多结构化数据的分析可能会生成额外的洞察，充实企业已经了解的信息

企业需要分析大数据的原因多种多样。如今的技术发展已使企业能够分析完整的数据集，而非数据子集。例如，企业可以分析每一次交互，而非每一笔交易。因此，多结构化数据的分析可能会生成额外的洞察，用以充实企业已经了解的信息，进而发现过去未知的其他机遇。这意味着可能会得到更为准确的业务洞察，帮助提升业务绩效。对于许多组织来说，即便通过分析更多数据将关键绩效指标改进了 **0.5%**、**1%**、**2%** 或者 **3%**，成果也是极为可观的。此外，数据流分析的引入也能提升响应速度并降低风险。

技能的短缺和市场的混乱妨碍了大数据技术的采用

但是，仍然有一些因素妨碍着大数据的分析。其中的两个原因如下：

更多细节能提升业务洞察的准确性和响应速度

1) 具有相应技能的人才短缺

2) 对于应该使用哪种技术平台的认识混淆不清

互联网充斥着有关关系 **DBMS**、**Hadoop** 与 **NoSQL DBMS** 对比的传言，很多人并不确定应在何时为哪类分析工作负载选用某种技术而非其他技术。

大数据分析应用

与结构化和多结构化数据相关的分析应用层出不穷。下表给出了大数据分析的部分行业用例。

行业	用例
金融服务	改进风险决策 “了解您的客户” **- 360** 度全方位的客户洞察 欺诈检测 程序化交易
保险	驾驶员行为分析（智能黑盒） 经纪人文档分析，用以深化对于承保风险的洞察，从而改进风险管理
医疗保健	医疗记录分析，以便了解患者再次入院的原因 疾病监测基因组学
制造业	“智能”产品使用和运行状况监控 通过分析服务记录来改善客户服务

大数据分析的用例极为广泛

	现场服务优化 通过关联报告的服务问题检测产品质量的早期预警，同时分析传感器数据，从而优化生产和分销
石油与天然气	油井、钻探设备和管道的运行状况和安全、风险、成本管理、生产优化的传感器数据分析
电信业	根据设备、传感器和 **GPS** 输入执行网络分析和优化，以加强社交网络和促销机遇
公用事业	智能仪表数据分析，电网优化 来自社交网络的客户洞察

Web 数据、传感器数据和文本数据已逐步成为大数据分析项目的流行数据源。

*通过分析 **Web** 日志实现网站优化*
在线广告需要分析用户在线时的点击流

就 **Web** 数据而言，点击流和社交网络内容分析较为普遍。企业通常会分析 **Web** 日志数据，以了解站点导航行为（会话分析），并将此与客户和/或登录数据相连接。媒体企业通常需要分析在线广告的“点击量”情况。这方面的挑战尤为严峻，因为这需要实时分析用户在线时的大量流式数据，从而通过发布广告来动态影响其在线导航行为。社交网络分析也是一个快速发展的领域。

传感器开启了一系列全新的优化机遇

机器生成/传感器数据的分析已用于供应链/分销链优化、资产管理、智能仪表、欺诈和电网运行状况监控等方面。

需要通过文本分析确定客户舆情

在非结构化内容的领域中，文本是极为重要的分析目标。案例管理、支持现场服务优化的故障管理、客户舆情分析、研究优化、媒体覆盖率分析和竞争对手分析都是与非结构化内容相关的大数据分析应用的典型示例。

大数据分析工作负载

许多大数据分析工作负载已经超越了传统数据仓库环境

考虑到大数据环境支持的分析范围，有必要观察超越传统数据仓库范畴的新型大数据分析工作负载。这些工作负载包括：

- 分析动态数据
- 非模式化、多结构化数据的探索式分析
- 结构化数据的复杂分析
- 存档数据的存储和重新处理
- 加速非模式化数据的 **ETL** 和分析处理，以充实数据仓库或分析设备中的数据

分析动态数据以制定运营决策

分析动态数据的目的是在事件发生时对其进行分析，从而检测影响（或预计将影响）成本、收入、预算、风险、最终期限和客户满意度等方面的数据模式。发生此类事件时，即可采取恰当的措施来尽可能降低其影响，或者最大化这些事件带来的机会。

事件流处理的主旨是自动检测、分析并在必要时处理事件，以保证优化业务

这种类型的大数据分析工作负载也被称为事件流处理，往往用于支持在一个工作日内可能发生的各类事件的日常运营决策。

示例包括金融市场中的股票销售、价格变动、订单变动、订单撤销、储蓄账户的大额提现、账户关闭、鼠标点击某个网站、错过贷款偿还、产品或货盘在分销链中的移动（通过 **RFID** 标记检测）、一条微博、一个竞争对手公告、街角处的 **CCTV** 视频、心电图 **(EKG)** 监控等。无论出现了哪些事件或数据流，在业务运营过程的每一秒中，都可能会发生数以千计乃至数以百万计的此类事件。尽管企业并非关注所有这些数据，但其中许多数据都需要某种类型的响应措施，以把握机会，避免问题发生或升级。在某些情况下，这样的响应可能需要立即或自动完成；而在其他一些情况下，此类响应则可能需要人为审批。

在一个工作日的业务运营中，可能会发生数以千计的事件

事件流处理需要在将数据存储在某个位置之前执行数据分析

流处理之所以独一无二，是因为需要在将数据存储到数据库或文件系统之前执行数据分析。考虑到数据生成的速度以及流式处理通常涉及到的数据量，这也意味着人为分析往往是不切实际的。因此，必须利用各种分析方法实现分析自动化，比如预测和统计模型或声音分析，从而确定或预测这些事件的业务影响。也可能必须自动化决策制定以确保及时响应，从而保证业务优化，并保证企业顺利达成目标。操作的范围极为广泛，从警报到完全自动化的操作（例如，调用事务或关闭油井中的某个阀门）。在出现常见问题时，更有可能采取后一种措施；而在出现需要人为干预的异常情况时，则需要前一种措施。请注意，为制定自动化决策并触发自动化操作，必须设定规则。因此，规则引擎是流处理中的重要组成部分。

人们不可能注意到每一个问题

在某些行业中，事件数据的数量极为庞大

在某些行业中，流式数据的数量极为庞大。当然，并非所有此类数据都需要保存。仅有背离常规的模式是可能需要保留以备后续执行历史分析的数据，从而识别重复出现的模式、问题和机会，而这一切都可能会作为未来战术和战略决策的依据。无论如何，即便经过筛选之后，事件数据的数量也可能极为庞大。

非模式化、多结构化数据的探索式分析

非模式化、多结构化数据需要加以探索，以确定哪些数据子集具有业务价值

多结构化数据的问题在于，此类数据往往属于非模式化数据，因此需要探索式分析[1]，以确定哪些数据子集具有业务价值。完成这样的分析之后，即可提取经判断具有价值的数据，并将其纳入数据结构，以便执行进一步分析并得出新业务洞察。

常见的多结构化数据源包括 **Web** 日志和外部社交网络交互数据。

声誉管理和“客户心声”是文本分析的主要目标

近期的一次调查[2]表明，目前从社交网络分析和提取数据的组织以面向客户的组织为主，主要目的是执行文本分析活动。这次调查还凸显了推动文本分析的主要业务应用，其中包括：

- 品牌/产品/声誉管理（**39%** 的受调查者）
- 客户心声（**39%**）
- 搜索、信息访问或问题解答（**39%**）
- 研究（**36%**）
- 竞争情报（**33%**）

文本的语言和格式各有不同

质量也可能是个问题

此类数据的挑战在于，数据量可能极为庞大，可能包含使用不同语言和格式的内容。其中可能包含大量质量低下的数据（例如，拼写错误或缩写词）和陈旧过时的内容。成功文本分析的一项关键要求是在分析之前“清理”内容。但是，许多企业通常不具备相应的机制。在分析之前预处理文本涉及到提取、解析、纠正和检测数据含义（利用注解器），以理解在分析文本时应考虑的上下文。这些问题突出表明了多结构化数据的复杂性。

多结构化数据难以分析

多结构化数据难以分析，以社交交互数据为例。分析用户生成的社交交互数据可能需要多次执行分析，以确定所需洞察。例如：

[1] 通常由数据科学家执行

[2] Text/Content Analytics 2011，Grimes，由 Alta Plana 于 2011 年 9 月发布

为了确定洞察，可能需要执行多次分析

- 第一次分析涉及到文本分析（挖掘），以便提取结构化客户舆情，同时提取嵌入于交互文本中、表示社交图成员的社交网络“用户”。
- 第二次分析的目的是分析提取出的数据，判断负面和正面舆情。
- 第三次分析则将社交网络“用户”加载到图形数据库中，在其中使用新型高级分析算法（例如 **N-Path**）来导航和分析链接，从而识别必要的联系人、关注者和关系，以便将此拼接成一个社交网络，判断具有较高影响力的人员。

基于搜索的分析工具有助于此类工作负载

此外，还可能需要预测分析、更复杂的统计分析和新的可视化工具。与此同时，利用多结构化数据搜索索引的基于搜索的分析工具也对此类工作负载极有帮助。

内容分析可能不仅仅限于文本分析，还包括音频和视频的分析

内容分析不仅仅限于文本分析，还可能需要处理音频、视频和图形。数字资产内容（例如，声音和视频）更难以解析，更难从中获得业务价值，原因在于此类内容并非文本。从此类内容获得洞察在很大程度上依赖于高级分析例程以及内容标记的合理程度，即能否描述内容是什么、主旨是什么。

非模式化数据的探索性分析需要一个过程

因此，非模式化、多结构化数据的探索性分析本身需要一个过程。大数据分析工作负载涉及到以下任务：

- 获得必要的非模式化数据
- 清理数据
- 探索数据以识别价值
- 通过探索式分析生成一个模型（结构）
- 解释或分析该模型以生成洞察

结构化数据的复杂分析

数据挖掘是结构化数据的复杂分析的一个典型示例

此类大数据分析工作负载的目标可能是从数据仓库或其他数据源（例如，运营事务系统）获取的结构化数据，具有对此类数据执行复杂分析的特定目的。此类分析可能是必要的，确保超级用户能挖掘数据以生成预测模型，从而在日常业务运营中使用。一个示例是构建为客户风险评分的模型。这些模型可用于推荐服务，确保在客户进行交互的所有业务渠道中使用相同的推荐，例如网站、联系中心、销售人员、合作伙伴、信息亭和实际门店（如分支机构、店铺等）。另一个示例是支持有关支出数据的详细统计分析，以生成模型，预测支出在何时可能超出预算，从而提前发出预警，确保掌控支出。

可以构建预测和统计模型，以便在数据库或实时操作中部署

某些垂直行业为复杂分析注入大量投资，意在降低风险

石油和天然气提供了另一个示例，即油井完整性方面的复杂分析，这对于管理环境风险、运行状况和安全，同时保证产量极为重要。灾难将对企业品牌产生极为严重的负面影响，因此与多个油井相关的多个文件中的详细数据将载入分析 **DBMS**，以便查找可能的问题，并对比油井之间的读数。此外，还可能需要将实际油井数据与地震数据进行对比，对比实际地质情况与来自地震调查的模拟地质情况。这有助于更好地识别钻探机会。

如今的企业尤为关注大数据环境中的存档数据存储和分析

存档数据的存储、预处理和查询

人们逐渐将大数据系统视为存储存档数据的廉价替代方案。组织的存档数据也在日益增加。这方面的原因多种多样，其中包括：

合规性、审计、电子取证和数据仓库存档都是企业希望执行此类操作的原因

- 将数据存储多年，以确保符合法律和/或行业法规。这包括与业务事务、主数据、淘汰的遗留应用程序、文档、电子邮件和音频文件相关的数据
- 需要在线存储存档数据，以支持审计
- 需要收集和存档结构化及多结构化数据，以避免未来的债务。
- 随着事务量、数据源以及需要更低细节级别的业务用户的增加，数据仓库中数据量的持续增加，因此需要管理数据仓库的成本和性能。这是通过存档价值已经降低的陈旧数据实现的。

尽管就目前而言，在大数据环境内存储此类数据是可以实现的，但如果需要恢复存档数据以响应法律挑战、支持审计要求或为特定的分析目的重新处理数据，挑战就会随之出现。在这种情况下，可能需要以多种方式访问和恢复数据。可能需要“按照存档的方式”执行查询和分析，以便满足合规性，也可能需要将这些数据恢复到自执行存档以来已改变的结构之中。此外，可能需要将数据仓库内当前存储的数据与存档的历史数据进行对比，同时保留多个版本的层次结构，以实现更为精准的对比分析。

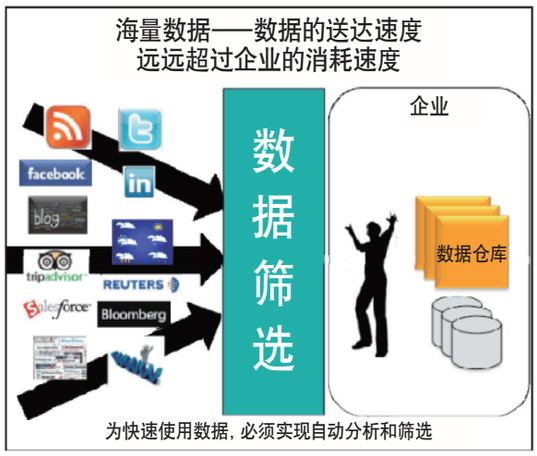
多结构化数据平台与数据仓库之间可能需要实现集成，以处理和分析数据

因此，这种大数据分析工作负载可能要求在多结构化数据平台与结构化数据仓库平台之间实现集成。这可能就要利用 **ETL** 处理，或者对存储在其他位置的存档仓库数据执行数据仓库预测和统计模型，将存档数据重新配置到传统环境中。音频文件或电子邮件和附件等多结构化存档数据也可能需要分析处理。所有这一切都强调了跨不同类型的存档数据管理混合工作负载和查询的需求。

加速结构化数据和非模式化数据的 ETL 处理

最后的一种大数据分析工作负载需要加速非模式化数据的筛选，以充实数据仓库或分析数据库设备中的现有数据。图 **1** 展示了这种工作负载。由于企业目前可用的新数据源如此之多，数据的送达速度远远超出了企业的消耗速度，因此必须将分析下推至 **ETL** 处理，以便自动分析非模式化数据，从而快速提取和使用有价值的数据。这种做法的目的在于加快利用非模式化数据充实现有分析系统的过程。这能提高敏捷性，开启更及时地生成业务洞察的新途径。

*目前，需要将分析下推至 **ETL** 处理，以便自动分析非模式化数据，从而更加迅速地使用所需数据*



端到端大数据分析的技术选项

传统环境需要添加新技术来支持大数据分析工作负载

根据大数据分析工作负载的定义，下一个问题是有哪些可用于支持这些工作负载的技术需要引入传统数据仓库环境，以便扩展此类环境，支持端到端的大数据分析？

用于动态大数据的事件流处理软件

流处理软件支持专为持续优化业务运营而设计的实时分析应用程序

流处理[3]软件用于支持以实时或接近实时的方式自动分析动态数据。其目的在于识别一个或多个数据流中有意义的模式，并尽快触发操作来做出响应。因而，此类软件提供了构建实时分析应用程序的能力，以持续优化业务运营的不同部分。这些应用程序必须能自动分析包含多结构化数据（例如，**Twitter** 流或视频流）和/或结构化数据的事件数据流。实时分析工作流中部署的预测和/或统计模型提供了流处理软件中的这种自动分析功能。此外，还需要利用一种规则引擎来自动化决策制定和操作的执行。

软件必须能应对高速的“事件风暴”— 事件无序、高速地出现

此类软件的难题之一是扩展以识别超高速“事件风暴”中的事件模式。如果模式中的事件未能依次出现，同一时间序列内有多个模式实例并存，那么就存在这样的需求。此外，即便来自同一模式的多个实例的事件混杂、无序地出现，也必须能够识别事件模式的每个实例。

另一个难题是集成多模式数据。一个示例是对来自街角摄像头的流式视频数据执行面部分析，然后将此数据与 **GPS** 信息相集成，将所关注的嫌疑人的所在位置告知距离最近的执法人员。另一个示例是使用多变量挖掘模型，比如回归或集群模型，以分析购买交易，并将此与 **Facebook** 或 **Twitter** 源集成。这些信息可用于将离家较远的汽油购买行为（通常会触发“信用卡可能被窃”）与在离家较远的地点发布的微博相关联，以正确判断这是否属于合理的购买行为，而非欺诈性购买。

静止大数据分析的存储选项

除了动态数据之外，本文讨论的所有其他大数据工作负载都与静止数据相关。换句话说，这些数据需要在执行分析之前进行存储。请务必牢记，前述工作负载与传统数据仓库内运行的工作负载同时存在。在这种全新的扩展分析环境中，有多种存储选项可支持对静止数据执行大数据分析。这些选项包括：

有多种存储选项可支持对静止数据执行大数据分析

- 分析 **RDBMS**
- **Hadoop** 解决方案
- **NoSQL DBMS**，比如图形 **DBMS**

此外，还可能需要分析 **RDBMS** 与 **Hadoop Map/Reduce** 集成的混合选项。

*分析 **RDBMS** 设备是专门为分析处理而优化的硬件/软件产品*

[3] 与复杂事件处理 (CEP) 相似，但侧重于分析（而非单纯的规则）以及分析多结构化数据的能力。

分析 RDBMS 设备

分析 **RDBMS** 平台是通常在专门为分析处理而优化的专用硬件上运行的关系型 **DBMS** 系统。这种硬件与专用 **DBMS** 软件的结合通常被称为*设备*，属于一种工作负载优化系统。多年以来，分析 **RDBMS** 不断增强，以改善已经得到充分认识的结构化数据的可伸缩性、查询性能和复杂分析。改进包括：

*多年以来，分析 **RDBMS** 设备不断增强*

- 引入固态硬盘 **(SSD)**，以加快 **I/O** 处理
- 特殊处理器，筛选贴近磁盘的数据
- 列存储和数据库查询处理
- 数据压缩
- 扫描共享，以使查询能够“附加”其他查询提供的数据
- 数据库内分析，用于以更接近数据的方式运行分析函数，充分发挥硬件和大规模并行处理的力量
- 内存中数据
- 跨 **SSD** 和转轴式磁盘划分多温度带数据

Hadoop 解决方案

__Hadoop__ 体系允许批量分析应用程序使用数千个计算节点处理分布式文件系统内存储的数 __PB__ 的数据

Apache Hadoop 是一种开放源码软件体系，旨在支持数据密集型分布式应用程序。它允许批量分析应用程序使用数以千计的计算机或计算机节点处理数 **PB** 的数据。除了开放源码 **Apache** 版本之外，市场中还有多个 **Hadoop** 商业发布版，其中许多都要在专用硬件设备上运行。**Hadoop** 体系的组件如下：

组件	描述
Hadoop HDFS	分布式文件系统，跨多台机器分区大文件，以便提高数据访问吞吐量
Hadoop MapReduce	一种分布式批处理编程框架，可跨多台服务器分布大型数据集
Chukwa	一种分布式数据（日志）收集和分析平台

Hive 是一种面向 Hadoop 的数据仓库系统，提供了预计 Hadoop 数据结构的机制

Hive 提供了一个接口，可将 SQL 转为 Map/Reduce 程序

Mahout 提供了完整的分析库，可发挥 Hadoop 集群的全部实力

Hive	一种面向 **Hadoop** 的数据仓库系统，能促进数据汇总、即席分析和 **Hadoop** 兼容文件系统内存储的大型数据集的分析。**Hive** 提供了一种预计此类数据的结构，并使用类似于 **SQL** 的语言 **HiveQL** 查询此类数据的机制。**HiveQL** 程序将转换为 **Map/Reduce** 程序
HBase	一种开放源码、分布式、支持版本控制、面向列的存储，根据 **Google** 的 **Bigtable** 建模
Pig	一种高级数据流语言，用于表示分析大型 **HDFS** 分布式数据集的 **Map/Reduce** 程序
Mahout	一种可伸缩的机器学习和数据挖掘库
Oozie	一种工作流/协调系统，用于管理 **Apache Hadoop** 作业
Zookeeper	一种面向分布式应用程序的高性能协调服务

Hadoop 通常不会与 **RDMS** 技术形成竞争关系。它扩展了使用更广泛内容的机会。因此，**Hadoop** 极为适合多结构化数据的探索式分析，但也可以在这种环境中分析结构化数据。

Hadoop 极为适合非模式化、多结构化数据的探索式分析

通常情况下，非模式化数据存储在 **Hadoop HDFS** 文件系统中，并在其中执行探索式分析，以判断结构，并将结构存储在 Hive 中以备后续分析。数据科学家使用 **Java**、**Python** 和 **R** 等语言开发在此环境中运行的批处理分析应用程序，此类应用程序使用 **MapReduce** 编程风格。这允许将程序复制到存储数据的数以千计的计算节点，从而支持程序并行运行。此外，**Mahout** 中的 **Hadoop** 内分析还可使用贴近数据的方式并行运行，从而充分发挥 **Hadoop** 集群的力量。**SQL** 开发人员和/或工具也可以利用 **Hive**，使用 **HiveQL** 语言访问 **Hadoop** 内的数据。此外，分析 **RDBMS** 供应商纷纷发布外部表函数和实用工具，为 **SQL** 社区提供了在 **Hadoop** 中处理多结构化数据的能力。

Mahout a 分析可应用于 Hadoop 数据，结果可存储在 Hive 中

Hive 具有为 SQL 开发人员和工具提供数据的界面

NoSQL DBMS

除了 **Hadoop HDFS**、**HBase** 和 **Hive** 以外，还有其他可作为分析数据存储的 **NoSQL DBMS** 选项。其中包括键值存储、文档 **DBMS**、列式 **DBMS**、图形数据库和 **XML DBMS**。一些 **NoSQL** 数据库不可以进行大数据分析。而其他 **NoSQL** 数据库则专门针对大数据分析，或者面向特定类型的分析。一个很好的示例是尤为适合社交图（网络）分析的图形 **DBMS**。请注意，**NoSQL** 市场中尚无标准可言。

图形数据库是一种 NoSQL 数据存储，尤为适合社交网络链接分析

哪种存储选项最为合理？

总体而言，大数据工作负载的数据和分析特征表明了最佳的解决方案。下表展示了在确定大数据分析工作负载应将数据存储在何处时，可用作指导准则的标准。

分析 **RDBMS**	**Hadoop / NoSQL DBMS**
数据分析和报告或复杂分析	在数据探索完成后即执行分析，这是 **NoSQL DBMS** 极为适合的一种分析类型，例如图形数据库中的图形分析
已经充分理解了数据	**尚未充分理解**数据
模式已定义并且已知	模式未定义，情况多变
批处理和联机分析	批量分析及一些通过 **Hive** 或 **Lucene** 实现的联机功能

Hadoop 适合批量分析非模式化数据或大量结构化数据

分析 RDBMS 适合结构化数据的复杂分析，以及不包含繁重混合工作负载的数据仓库

通过生成 **SQL**、可在数据库内运行预测/统计模型的 **BI** 工具进行访问	使用 **Java**、**R**、**Python**、**Pig** 等语言开发 **MapReduce** 应用程序
在专门构建的 **MPP** 集群上，可伸缩到数百 **TB**	在专门构建的设备或云中，可伸缩到数 **PB**

查看静止大数据的工作负载，下表尝试将各工作负载与恰当的数据存储平台相匹配。

请务必注意，应保证数据特征和分析工作负载与技术相符，以便选择最佳平台

大数据分析工作负载	大数据存储平台
非模式化、多结构化数据的探索式分析，例如 **Web** 日志、非结构化内容、经过筛选的传感器数据、电子邮件	**Hadoop**
结构化数据的复杂分析，或者拥有“轻量级”混合工作负载的数据仓库的分析	分析 **RDBMS** 设备
存档数据的存储和重新处理	**Hadoop**
加速结构化和非模式化数据的 **ETL** 处理	混合：**Hadoop** 和分析 **DBMS**
社交图链接分析	**NoSQL** 图形 **DBMS**

静止大数据的可伸缩数据管理选项

在扩展的分析环境中，必须找到一种跨所有分析数据存储实现信息管理的通用工具套件

在这种全新的扩展分析环境中，一个关键的成功要素是跨多个分析数据存储的始终如一、高质量的数据，包括数据仓库 **RDBMS**、分析 **RDBMS** 设备、**Hadoop** 集群/设备和 **NoSQL DBMS**。数据管理的选项多种多样，包括针对不同平台的各种数据管理工具，以及能够为所有平台提供数据的通用工具套件。理想情况下，后者更为合理。但是，有时在分析过程中也可能需要在不同平台之间移动数据，包括：

- 将主数据从 **MDM** 系统移动到数据仓库、分析 **DBMS** 或者 **Hadoop**
- 将得到的结构化数据从 **Hive** 移动到数据仓库
- 将筛选后的事件数据移动到 **Hadoop** 或分析 **RDBMS** 之中
- 将数据仓库中的维度数据移动到 **Hadoop**
- 将 **Hadoop** 中的社交图数据移动到图形数据库
- 将图形数据库中的数据移动到数据仓库

如今，所有这些移动操作都是必不可少的，如图 **2** 所示。

信息管理需要整合数据，以便加载分析数据存储，同时在数据存储之间移动数据

*信息管理套件需要集成 **Hadoop**、**NoSQL DBMS**、数据仓库、分析 **RDBMS** 和 **MDM***

图 2

为了实现这一切，信息管理软件需要支持 **Hadoop**（多结构化数据）和/或分析 **RDBMS**（结构化数据）上的 **ELT** 处理，并与事件处理结合，摄取经过筛选的事件流数据，将数据载入 **Hadoop** 和 **NoSQL DBMS**，在 **Hadoop** 中分析数据，在 **Hadoop** 中清理数据，生成 **HiveQL**、**PIG** 或 **JAQL** 以便在 **Hive** 或 **Hadoop HDFS** 中处理多结构化数据，对 **Hadoop** 中的数据执行自动化分析，最终从 **Hadoop** 和 **NoSQL DBMS** 中提取数据。它必须支持主数据管理。

大数据分析选项

为了在这种全新的扩展分析环境中分析数据，根据数据的存储位置，有多种选项可用。这些选项如下所示：

有多种选项可用于分析静止大数据

- 使用“**Hadoop** 内”的自定义或 **Mahout** 分析的自定义 **Hadoop MapReduce** 批量分析应用程序
- 生成 **MapReduce** 应用程序的基于 **MapReduce** 的 **BI** 工具和应用程序
- 对分析 **DBMS** 的数据库内分析
- 除了数据仓库和多维数据集之外，传统 **BI** 工具还会分析 **Hadoop Hive** 和分析 **RDBMS** 中的数据。
- **Hadoop** 和分析 **RDBMS** 中基于搜索的 **BI** 工具
- 对事件数据流中动态数据的动态分析

*自定义构建 **map / reduce** 应用程序，以分析 **Hadoop** 中的数据*

第一个选项是构建自定义 **map/reduce** 应用程序，分析 **Hadoop HDFS** 中的多结构化数据。示例包括文本、点击流数据和图像等。例如，参与探索式分析、使用 **R** 语言编写自己的分析的数据科学家可能会选择这种选项，在其应用程序内使用预先构建的 **Mahout** 分析库的数据科学家也可能会选择这种选项。

*Hadoop 中预先构建的 **Mahout** 分析*

*Hadoop 中使用 **map/reduce** 的预构建分析应用程序*

*在 **Hadoop** 中生成 **map/reduce** 作业的新 **BI** 工具*

另外，还有其他一些预构建的分析应用程序解决方案和新 **BI** 工具可用于生成 **MapReduce** 应用程序，利用 **Hadoop** 中的并行机制分析多结构化数据，比如海量内容或客户交互数据。

*分析 **RDBMS** 中的数据库内分析*

数据库内分析是在分析 **RDBMS** 内部署自定义构建或预先构建的分析，以分析结构化数据。这是结构化数据的复杂分析的一个典型示例。

基于 SQL 的 BI 工具通过 Hive 访问 Hadoop 数据库，或者访问 RDBMS

基于搜索的 BI 工具和应用程序使用索引分析 Hadoop 和/或分析 RDBMS 内的数据

除了访问分析 **RDBMS** 和数据仓库来分析和报告结构化数据的传统 **BI** 工具之外，部分此类工具现在还支持 **Hive** 界面，允许将所生成的 **SQL** 转为 **MapReduce** 应用程序，从而在 **Hadoop** 中处理多结构化或结构化数据。

最后，考虑到分析的文本数量，现在兴起了全新的基于搜索的 **BI** 工具（如图 **3** 所示），允许对 **Hadoop** 和/或数据仓库设备内的多结构化和结构化数据执行自由形式的分析。这些工具可以抓取分析 **RDBMS** 中的结构化数据，同时利用 **MapReduce** 为 **Hadoop** 中的数据建立索引。随后，即可在这些索引的基础之上构建分析应用程序，支持对多结构化数据和/或结构化数据执行自由形式的探索式分析。这些工具可以利用 **Hadoop Lucene** 搜索引擎索引或者其自身存储在 **Hadoop** 内的其他索引。

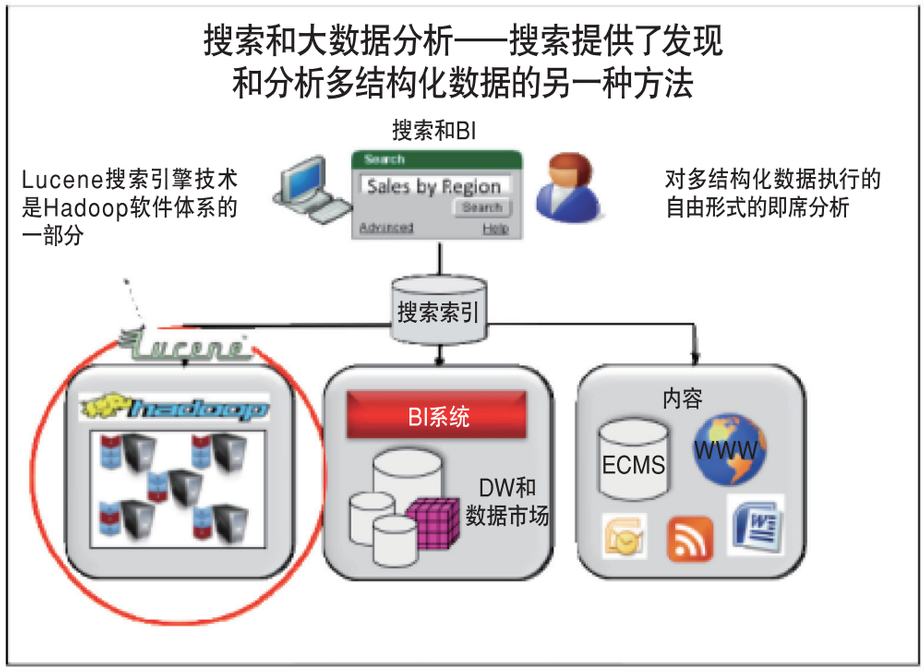


图 3

将大数据集成到您的传统 DW/BI 环境中

了解了大数据技术之后，最显而易见的问题是：我们所讨论的所有这些技术组件如何与传统数据仓库环境相集成，从而将此类环境扩展为支持所有传统和大数据分析工作负载？

新型企业分析生态系统

需要扩展传统数据仓库环境以支持大数据分析工作负载

图 4 展示了支持本文所述大数据分析工作负载以及传统数据仓库即席查询处理、分析和报告所需的扩展的端到端分析环境。有些人将这种新型环境称为“企业分析生态系统”或“逻辑数据仓库”。通过这种架构可以看出，可以对传感器数据或其他任何事件数据源（比如金融市场）执行动态数据的事件处理。如果事件数据中出现变化，事件处理软件将分析其业务影响，并在必要情况下采取措施。随后，信息管理软件可获取筛选后的事件，将其载入 Hadoop 以备后续进行历史分析。如果使用成批的 map/reduce 分析处理生成了任何额外的洞察，那么随后会将洞察传送到数据仓库中。对于非模式化、多结构化数据，此类数据可利用信息管理软件直接载入 Hadoop，便于数据科学家在其中利用定制的 map/reduce 应用程序或生成 HiveQL、Pig 或 JAQL 的 map/reduce 工具执行探索式分析。此外，还可利用基于搜索的 BI 工具，通过 map/reduce 实用工具使用 Hadoop 内置索引分析数据。例如，如果多结构化数据是 Twitter 数据，则将提取 Twitter 用户，并将其载入 NoSQL 图形数据库，以便执行进一步的社交网络链接分析。信息管理软件可以管理社交网络链接数据从 Hadoop 到 NoSQL 图形 DBMS 的转移，从而支持此类分析。如果数据科学家获得了任何宝贵洞察，那么它还能将洞察载入数据仓库，充实其中已经包含的结构化数据，让传统 BI 工具用户同样能利用这些洞察。

信息管理在确保这种环境实现集成的过程中扮演着重要角色

结构化数据的复杂分析是利用数据库内分析在分析 **DBMS** 设备中执行的。同样，如果产生了任何洞察，或者创建了任何新的预测/统计模型，即可将其移动到数据仓库中，供信息使用者在报表、仪表板和记分卡中使用。存档数据的存储和重新处理可以在 **Hadoop** 中利用批量 **map/reduce** 应用程序或上文提到的用于分析此类数据的前端工具管理。**Hadoop** 内分析（定制构建或 **Mahout**）可按需使用。最后，在加速结构化和非模式化数据的 **ETL** 处理方面，可以利用信息管理工具，以便利用 **Hadoop** 分析和/或分析 **DBMS** 设备内的数据库内分析。传统数据仓库工作负载也会照常继续。

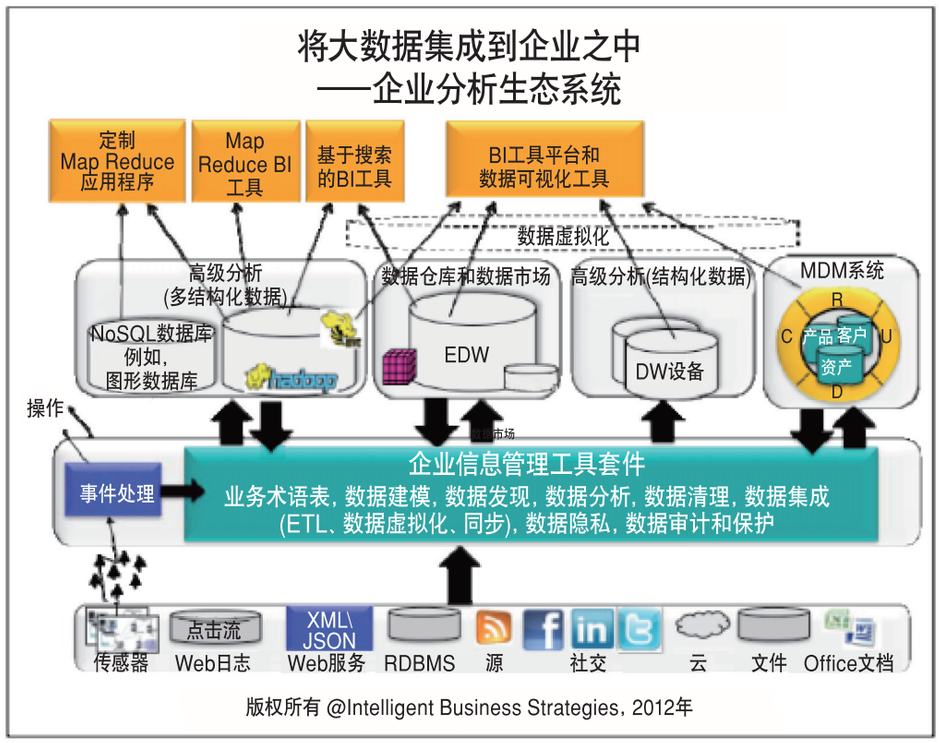


图 4

数据虚拟化能简化对多种分析数据存储的访问

主数据管理为所有分析平台提供一致的主数据

这种新型扩展分析环境混合使用了传统数据仓库和大数据工作负载优化系统，因此也需要为使用传统前端工具的用户简化这些数据存储的访问。这是通过数据虚拟化软件实现的，此类软件能让数据看上去就像位于一个数据存储中一样。在后台，它使用下推优化和其他高级优化技术答复业务查询。此外，还可以利用主数据，为所有分析环境提供一致的维度数据。

接合分析处理 - 工作流的力量

信息管理工作流可以转为在整个分析生态系统中工作的分析流程

企业信息管理软件支撑着整个扩展分析环境。这种软件的主要优势之一是能够在图形工作流中定义数据质量和数据集成转换。但是，考虑到分析现在可以下推到分析数据库和 **Hadoop** 之中，因此分析、规则、决策和操作现可添加到信息管理工作流中，以便创建自动分析流程。随后，可在扩展分析环境内跨*多个平台*运行这些分析流程。这也就是说，可以定期构建和重用工作流，以对结构化和非模式化、多结构化数据进行一般分析处理，从而加速组织使用、分析和处理数据的速度。所有这一切都有可能自动化。这种概念的强大力量令人敬畏，也激励着我们将信息管理工作流转变为可跨企业创建的成熟分析流程。此外，这些分析流程利用最适合构成此类工作流的分析工作负载分析平台。如果这还不够，整个工作流将由服务支持，让其可按需使用。例如，社交网络交互数据可载入 **Hadoop**，调用文本数据来提取舆情数据和社交“用户”，用户更新主数据，将舆情与客户和评分关联，从而将舆情评分添加到数据仓库之中。随后，即可执行分析来确定存在不满意情绪的*宝贵*客户，并采取措施来留住这些客户。此外，可以将社交用户载入图形数据库，深度分析社交网络链接，以识别具有较高影响力的用户和开启交叉销售机遇的其他关系。

这能加速组织使用、分析和处理数据的速度

可以利用强大的新型分析工作流留住客户，加强竞争优势

新型分析生态系统的技术要求

为了支持所有这一切，技术解决方案需要满足以下需求。以下列表不分优先顺序。

除了企业数据仓库之外，还有多种分析数据存储

- 支持多种分析数据存储，包括：
  - **Apache Hadoop** 或 **Hadoop** 的商业发布版，支持经济高效的存储和非结构化数据的索引
  - **MPP** 分析 **DBMS** 提供预先构建的数据库内分析，支持并行运行以多种语言（例如，**R** 语言）编写的定制构建分析，以便改进分析海量数据时的性能
  - 数据仓库 **RDBMS**，支持面向业务、可重复的分析和报告
  - 图形 **DBMS**
- 支持流处理，以分析动态数据
- 信息管理工具套件，支持加载 **Hadoop HDFS** 或 **Hive**、图形 **DBMS**、分析 **RDBMS**、数据仓库和主数据管理
- 信息管理工具套件生成 **HiveQL**、**Pig** 或 **JAQL** 的能力，以充分利用 **Hadoop** 处理的能力

信息管理工具与所有分析数据存储和事件流处理的集成

- 流处理与信息管理工具集成，以便获取筛选后的事件数据，并将其存储在 **Hadoop** 或分析 **RDBMS** 之中，以备进一步分析
- 支持跨多种 **SQL** 和 **NoSQL** 数据存储的无缝数据流

简化数据访问的数据虚拟化

- 数据虚拟化，以隐藏多种分析数据存储的复杂性

最适合的查询优化和数据存储内分析

- 查询重定向，以在最适合所需分析的分析系统内运行分析查询，即数据仓库、分析 **RDBMS**、**Hadoop** 平台和事件流处理引擎等

在多种分析数据存储和事件流处理中部署模型

- 开发预测和统计模型，并将其部署在一个或多个工作负载优化的系统以及数据仓库（例如 **Hadoop** 系统）中，支持实时预测分析的分析 **RDBMS** 和事件流处理工作流

分析工作流及全面的决策管理

- 在信息管理工作流中运行数据库内和 **Hadoop** 内分析的能力，以便在数据转换和数据移动过程中自动化分析
- 信息管理工作流之间的集成，以及支持工作流执行过程中的自动化决策的规则引擎
- 嵌套工作流，以支持多道分析查询处理

	• 在 **ETL** 处理过程中利用 **Hadoop** 并行集群
	• 创建沙盒的能力，以支持跨一个或多个底层分析数据存储的探索式分析
支持数据科学家执行探索式分析的沙盒	• 支持数据科学项目管理以及在沙盒中开展工作的数据科学家间的协作的能力
	• 将数据源和信息管理工作流与沙盒相关联，从而确定数据来源、加载和准备数据以支持 **MPP** 沙盒环境中海量内容探索式分析的能力
	• 将第三方分析工具集成到特定沙盒中的能力
治理整个生态系统，包括沙盒和数据科学家	• 控制沙盒访问的能力
	• 开发可在 **Hadoop** 平台、图形数据库或分析 **RDBMS** 内部署为 **SQL MR** 函数的 **MapReduce** 应用程序的工具
	• **Hadoop** 集群或实时流处理集群中文本分析的并行执行
并行工作以解决业务问题的新工具和传统工具	• 跨新型企业分析生态系统内的所有分析数据存储构建搜索索引的能力，以促进基于搜索的结构化和多结构化数据分析
	• 通过 **Hive** 将传统 **BI** 工具连接到 **Hadoop** 的能力
	• 跨整个分析生态系统的端到端单一控制台系统管理
	• 跨整个分析生态系统的端到端工作负载管理

入门：企业的大数据分析战略

业务协调

大数据项目需要与业务战略相协调

从某种意义上来说，开始利用大数据分析与其他分析项目并无差别。必须有相应的业务缘由，确保所生成的洞察能帮助组织实现业务战略中设定的目标。对于分析来说，与业务战略保持一致是成功的关键要素，在大数据方面也是如此。

识别备选大数据项目，并根据业务收益为其划分优先级

要实现此目的，组织应该创建备选大数据分析应用程序的列表，涵盖结构化和/或多结构化的动态数据和静止数据。显而易见的示例包括 **Web** 日志、文本、社交网络数据、金融市场和传感器数据的分析。表明存在相应需求的信号还包括：当前技术制约了分析数据的能力；数据或分析复杂性导致数据分析无法正常完成。就像在传统数据仓库中一样，随后需要根据投资回报率为这些备选应用程序指定业务优先级。

工作负载与分析平台的协调

将分析工作负载与最适合相应工作的分析平台相匹配

确定上述方面之后，组织应设法将分析工作负载与最适合相应工作的平台相匹配。例如，非模式化、多结构化数据（比如社交网络交互数据）的探索式分析可能更适合采用 **Hadoop** 平台，而结构化数据的复杂分析可能更适合采用分析 **RDBMS**。组织应将工作负载与恰当的平台相匹配，并不能盲目地将所有数据都存储在一个企业数据仓库之中。

技能集

数据科学家是企业急需招募的新型人才

在技能方面，新技能集也在不断兴起。数据科学家的技能就属于此类新技能集。为了研究分析项目，必须借助数据科学家的力量，而在数据仓库和 **BI** 工作中，传统数据仓库开发人员和业务分析师不可或缺。在这种新型分析环境中，所有这些人员都需要协同工作。

数据科学家是积极主动、善于分析的人才，具有深厚的数学背景，对数据充满热爱

数据科学家往往在统计和数学建模技术方面拥有深厚的背景，能够利用编程语言探索、分析和可视化多结构化数据。例如，在石油与天然气行业中，他们可能是能够熟练地挖掘、筛选、分析和可视化海量复杂数据的地震工程师。他们需要能自由分析非模式化数据，并生成可用于充实传统环境中数据的洞察。

*传统 **ETL** 开发人员和业务分析师需要拓宽其技能领域，以适应大数据平台和数据仓库*

传统数据仓库 **ETL** 开发人员需要拓展信息管理工具的使用范围，需要将数据载入 **Hadoop** 环境，以支持数据科学家开展分析；还需要将数据科学家认定具有较高业务价值的 **Hadoop** 中的数据载入数据仓库。此外，业务分析师应增加对 **BI** 工具的使用，以利用 **Hive** 界面访问 **Hadoop** 数据和数据仓库中的数据。

为数据科学和探索搭建环境

希望对大数据执行研究分析的数据科学家需要利用治理妥当的沙盒

数据科学家需要一种治理妥当的环境，以便探索非模式化数据和/或对海量结构化数据执行复杂分析。创建项目环境，以支持小型数据科学家团队在 **Hadoop** 和/或分析 **RDBMS** 设备上的“沙盒”中工作及开展协作是极为重要的一步，沙盒的创建和访问需要得到控制。此外，数据科学家还要能够搜索待分析内容，并迅速突出显示所需数据源。随后，需要利用信息管理工具将数据载入沙盒、从沙盒中提取数据，从而治理进出沙盒的数据。此外，还需要选择构建定制 **Map/Reduce** 应用程序和/或挖掘数据的工具，以发现洞察、开发分析模型。随

后即可将这些工具部署到流处理中，从而分析动态数据；或者部署到分析 **RDBMS**、**Hadoop** 和数据仓库当中，分析静止数据。

定义分析模式和工作流

*事件流处理和基于 **Hadoop** 的分析通常在数据仓库的上游执行*

为了最大限度地提升这种新型分析生态系统的业务获益，必须定义模式。一种显而易见的模式包括将事件流处理和 **Hadoop** 定位为处于数据仓库“上游”的分析系统。这也就意味着，需要将流处理中经过筛选和分析的流数据传递到 **Hadoop**、**NoSQL** 和/或数据仓库中，以便执行后续分析。此外，**Hadoop** 或 **NoSQL** 图形数据库内探索式分析生成的洞察也需要引入数据仓库，以充实已知内容。

通过集成技术过渡到大数据企业

为了允许您的组织建立涵盖大数据和传统 **DW/BI** 分析工作负载的扩展分析生态系统，必须对技术加以扩展和集成。其中涉及到的工作如下：

利用大数据洞察充实数据仓库内的数据

为了创建迎合所有分析工作负载需求的新型企业分析生态系统，需要添加额外的技术，并将其与传统数据仓库相集成

- 向传统环境和存在相应业务需求的位置添加新的分析平台和流处理，支持大数据分析工作负载
- 将信息管理工具的使用扩展到数据仓库和数据市场以外，以便：
  - 为 **Hadoop**、**NoSQL** 图形数据库、分析 **RDBMS** 设备、数据仓库和 **MDM** 系统提供数据
  - 在 **Hadoop** 与数据仓库和/或分析 **RDBMS** 设备之间移动数据
  - 将 **MDM** 系统中的数据移动到 **Hadoop**、数据仓库和/或分析 **RDBMS** 设备
  - 集成流处理，以便将经过筛选或分析的流数据加载到 **Hadoop** 和/或分析 **RDBMS** 之中
- 将预测分析与 **Hadoop**、**DW** 设备和流处理引擎相集成，从而在工作负载优化系统中部署模型
- 如有可能，将现有 **BI** 工具与 **Hadoop Hive** 相集成
- 将基于搜索的 **BI** 工具与 **Hadoop** 和数据仓库相集成
- 在所有分析系统的基础之上引入数据虚拟化
- 将前端工具集成到门户中，以创建单一用户界面

供应商示例：IBM 的端到端大数据解决方案

上文已经给出了大数据的定义，了解了大数据分析工作负载以及新型扩展企业分析生态系统的技术和其他需求，本部分将着眼于一家供应商（即 **IBM**）如何应对交付必要技术和集成以支持所有这一切的挑战。换句话说，也就是让组织能够支持传统、运营和大数据分析工作负载，实现端到端的分析处理。

IBM 提供了一系列技术组件，支持对动态数据和静止数据执行端到端的分析

IBM 提供了一系列集成技术组件，支持对动态数据和静止数据执行端到端的分析。这些组件包括：

其中包括一个数据仓库和一系列分析设备

- 一种流处理引擎，支持实时分析动态数据
- 一种数据仓库平台，支持对静止的结构化数据执行传统分析和报告
- 一系列分析设备，专门为特定的大数据高级分析工作负载而优化
- 一种允许加速运营分析查询处理的设备
- 一种集成化自助服务 **BI** 工具套件，支持即席分析和报告，包括对于移动 **BI** 的支持
- 基于搜索的技术，支持构建分析应用程序，允许对多结构化数据和结构化数据执行自由形式的探索式分析
- 面向模型开发和决策管理的预测分析
- 针对内容分析的应用程序和工具
- 预先构建的模板，用于快速开始对流行的大数据源进行分析处理

用于治理和管理数据的信息管理工具

- 一种集成的信息管理工具套件，用于治理和管理此类新型扩展分析环境中的数据

IBM 大数据平台包含所有这些组件

这一组技术结合在一起，构成了 **IBM** 大数据平台，如下图所示。该平台包含三种分析引擎，支持大多数企业所需的大量传统和大数据分析工作负载。这其中包括：

IBM 大数据平台中的三种分析引擎

- 流计算
- **Hadoop** 系统
- 数据仓库（可以是一个或多个数据存储）

该平台同时具备可扩展的特点，能够支持额外的第三方分析数据存储，例如，非 **IBM** 分析 **RDBMS** 和 **NoSQL** 数据存储。

I IBM 大数据平台是 IBM 的企业分析生态系统

IBM INFOSPHERE STREAMS – 分析动态大数据

***IBM InfoSphere Streams** 支持对动态数据执行持续的实时分析*

IBM InfoSphere Streams 是 **IBM** 大数据平台中支持构建和部署持续实时分析应用程序以分析动态数据的技术组件。这些应用程序将持续不断地寻找数据流中的模式。检测到模式后，将分析模式的影响，并即时制定实时决策，从而加强竞争优势。示例包括分析金融市场交易行为、分析 **RFID** 数据以实现供应链和分销链优化，监控传感器数据以支持制造流程控制，新生儿重症监护室监控，实时欺诈防范和执法中的实时多模式监测。**IBM InfoSphere Streams** 能同时监控多个外部和内部事件流，无论它们是由机器生成的，还是手动生成的。该解决方案支持海量结构化和非结构化流式数据源，包括文本、图像、音频、语音、**VoIP**、视频、**Web** 流量、电子邮件、地理空间数据、**GPS** 数据、金融交易数据、卫星数据、传感器和其他任何类型的数字信息。

***IBM InfoSphere Streams** 产品附带预先构建的工具包和连接器，能加速实时分析应用程序的开发*

为帮助加速实时分析应用程序的开发，**IBM** 还提供了预先构建的分析工具包和面向流行数据源的连接器。**IBM** 合作伙伴还提供了第三方分析库。此外，还包含一种基于 **Eclipse** 的集成化开发环境 **(IDE)**，允许组织构建自己的定制构建实时分析应用程序，支持流处理。还可以在 **InfoSphere Streams** 分析应用程序工作流中嵌入 **IBM SPSS** 预测模型或分析决策管理模型，以预测事件模式的业务影响。

IBM InfoSphere Streams 可用于持续不断地将数据摄取到 IBM BigInsights Hadoop 系统当中，以支持进一步分析

在为实时分析而优化的多核、多处理器的硬件集群上部署 **InfoSphere Streams** 应用程序能确保可伸缩性。此外，还可以筛选出业务部门关注的事件，将其传送到 **IBM** 大数据平台内的其他分析数据存储，以便执行进一步分析和/或重放。因此，可以利用 **InfoSphere Streams** 持续不断地将关注的数据摄取到 **IBM BigInsights** 内，以执行分析。也可以汇总海量数据流，将其路由到 **IBM Cognos Real-Time Monitoring**，随后在仪表板中显示，以便进一步执行手动分析。 .

支持分析静止数据的 IBM 设备

IBM 付出了大量努力来增强 Hadoop，使之更加可靠

除了 **IBM InfoSphere Streams** 外，**IBM** 大数据平台还包含一个传统数据仓库，以及专为大数据分析、面向结构化数据的高级分析，以及加速运营事务数据的分析查询处理而优化的设备。具体如下：

IBM InfoSphere BigInsights

IBM InfoSphere BigInsights 是 IBM 的商业 Hadoop 发布版

IBM InfoSphere BigInsights 是 **IBM** 的商业 **Apache Hadoop** 系统发布版。它专为海量多结构化数据的探索式分析而设计，旨在获得过去不可能获得的洞察。**IBM InfoSphere BigInsights** 随附提供了标准 **Apache Hadoop** 软件。但是，**IBM** 还添加了以下功能，旨在加强该产品的能力：

IBM InfoSphere BigInsights 不但支持 IBM 自己的 Hadoop 发行版，还支持第三方 Hadoop 发行版

- 企业级可伸缩、兼容 **Posix** 的文件系统 **GPFS-SNC** [4]
- **JSON** 查询语言 **(JAQL)**，支持轻松操控和分析半结构化 **JSON** 数据
- 数据压缩
- 基于 **Map/reduce** 的文本和机器学习分析
- 存储安全性和集群管理
- 除了 **IBM** 的 **Hadoop** 发行版之外，还支持 **Cloudera** 的 **Hadoop** 发行版
- **IBM DB2**、**IBM PureData Systems for Operational Analytics (DB2) and for Analytics (Netezza)** 的连接器，支持在大数据分析过程中通过基于 **JAQL** 的 **MapReduce** 应用程序访问结构化数据
- 作业调度和工作流管理
- **BigIndex –** 一种 **MapReduce** 工具，利用 **Hadoop** 的强大力量，为基于搜索的分析应用程序建立索引

BigInsights 提供两个版本：

- **IBM BigInsights Basic Edition –** 预先构建、可免费下载的 **BigInsights** 试用版
- **IBM BigInsights Enterprise Edition**

BigInsights Enterprise Edition 适用于预先配置的 **IBM System x** 和 **PowerLinux** 参考架构。

IBM 大数据平台内的分析系统属于 **PureSystem** 全新系列系统的一部分。**PureSystem** 系列系统是在今年早些时候发布的，该系列基于一组专注于集成专业技术、简化管理和优化性能的核心准则而构建。**2012** 年 **10** 月，**IBM** 对这个系列进行了扩展，包含专门关注数据的系统，即 **PureData Systems**。新引

[4] SNC = 无共享集群

入的系统有两种 – **PureData System for Analytics** 和 **PureData System for Operational Analytics**，下文将进一步介绍这两种系统。

IBM PureData System for Analytics（采用 Netezza 技术）

IBM PureData System for Analytics 专为针对结构化数据的高级分析和某些数据仓库工作负载而优化

采用 **Netezza** 技术的 **IBM PureData System for Analytics** 是新一代的 **Netezza** 设备，专为针对结构化数据的高级分析工作负载而优化。

IBM PureData System for Analytics 是一种紧凑、低成本的数据仓库和分析硬件设备。其用户数据容量能从 **100 GB** 扩展到 **10 TB**，并且全面兼容 **IBM Netezza 1000** 和 **IBM Netezza High Capacity Appliance**。

PureData System for Analytics 是为特定用途构建的基于标准的数据仓库设备，它将数据库、服务器、存储和高级分析功能集成在一个系统中。它能从 **1 TB** 扩展到 **1.5 PB**，包含筛选从磁盘中传出的数据的特殊处理程序，确保仅在 **RDBMS** 中处理查询相关数据。名为 **Netezza Platform Software (NPS)** 的 **IBM Netezza Analytic RDBMS** 无需索引或调优，因此更易于管理。它旨在与传统 **BI** 工具（包括 **IBM Cognos BI** 平台在内）接合，还能运行在包含海量数据的数据库中部署且由 **IBM SPSS** 开发的高级分析模型。

IBM Netezza Analytics 提供了数据库内分析功能，每个 IBM Netezza 1000 或 PureData System for Analytics 均免费提供该设备，让您能够在设备内创建和应用复杂的分析。

IBM Netezza Analytics 是一种先进的分析框架，对 **IBM PureData System for Analytics** 形成了补充。除了提供庞大的并行化高级算法和预测算法库之外，它还支持使用多种不同的编程语言（包括 **C**、**C++**、**Java**、**Perl**、**Python**、**Lua**、**R*** 甚至是 **Fortran**）创建定制分析，允许集成 **SAS**、**SPSS**、**Revolution Analytics**、**Fuzzy Logix** 和 **Zementis** 等公司提供的领先第三方分析软件产品。

** 利用 Revolution Analytics 推出的 Revolution R Enterprise 软件*

IBM PureData System for Analytics 允许您创建、测试和应用模型，在 **IBM Netezza** 设备内对数据进行评分，消除了移动数据的需要，与需要将数据提取到笔记本电脑或其他小型计算机相比，这种方法能让您访问更多的数据和更多的属性。

IBM PureData System for Operational Analytics 是一种模块化、预集成的平台，专为运营分析数据工作负载而优化

DB2 10 包含 NoSQL 图形存储

IBM PureData System for Operational Analytics

IBM PureData System for Operational Analytics 于 10 月推出，基于 IBM Power System 和 IBM Smart Analytics System。IBM Smart Analytics System 是一种模块化、预集成的实时企业数据仓库，专为 **IBM System x**、**IBM Power System** 或 **IBM System z** 服务器中的运营分析数据工作负载而优化。这些系统经过预先集成和优化，能确保快速的部署，能够加速创造价值的过程。**IBM PureData System for Operational Analytics** 和 **IBM Smart Analytics System** 系列产品均包含基于 **DB2 Enterprise Server Edition 10** 运行的 **IBM InfoSphere Warehouse 10** 软件。**InfoSphere Warehouse** 包含数据建模、数据移动和转换、**OLAP** 函数、数据库内数据挖掘、文本分析和集成化工作负载管理。**DB2 10** 包含经过改进的工作负载管理，表级、页级和存档日志压缩，全新的索引扫描和预先获取，时序数据访问和全新的 **NoSQL** 图形存储。全新 **PureData System** 解决方案中还利用固态硬盘 **(SSD)**，显著改进了自动优化数据放置功能。此外，还提供了 **IBM Cognos Business Intelligence**。

IBM 大数据平台加速器

IBM 大数据加速器旨在加速 IBM 大数据平台的开发工作

为了加速和简化 **IBM** 大数据平台的开发，**IBM** 构建了许多加速器。其中包含百余种示例应用程序、用户定义的工具包、标准工具包、行业加速器和分析加速器。示例包括：

- 数据挖掘分析
- 实时优化和流分析
- 视频分析
- 面向银行业、保险业、零售业、电信业和公共运输业的加速器
- 预先构建的行业数据模型
- 社交媒体分析
- 舆情分析

IBM DB2 分析加速器能分载在 IBM System z 上运行 DB2 混合工作负载的 OLTP 系统的复杂分析查询

IBM DB2 分析加速器 (IDAA)

IBM DB2 分析加速器是一种 **IBM Netezza 1000™** 和/或 **PureData System for Analytics** 设备，旨在分载 **IBM System z** 上 **DB2** 混合工作负载的复杂分析查询。这是通过使用预先定义的管理性 **DB2** 存储过程在 **IDAA** 上重建 **DB2** 表，随后再将 **DB2** 中的数据载入 **IDAA** 实现的。如有必要，还可以锁定 **DB2** 表，避免其在 **IDAA** 加载过程中被更新。此外，在加载过程中，可以路由查询，并由 **IDAA** 执行处理。无需对访问 **DB2** 的任何应用程序或工具做出任何更改。原因在于，**DB2** 优化器将负责确定要将哪些动态 **SQL** 查询重新路由到 **IBM DB2** 分析加速器，以执行并行查询处理。因此，无论意图和目的如何，**IDAA** 对于查询 **IBM System z** 上的 **DB2 DBMS** 的应用程序和报告工具来说都是“不可见的”。此外，由于 **Netezza** 技术没有任何索引，所有管理活动都通过预先构建的 **DB2** 存储过程完成，因此数据库管理方面的开销也能保持在最低限度。这就避免了容量升级，同时确保提高服务水平。

目前共有三种 **IDAA** 产品，支持 **8**、**16** 和 **32 TB** 的用户数据。通过数据压缩能增加容量。

面向大数据企业的 IBM 信息管理

IBM InfoSphere Information Server 和 Foundation Tools 提供了跨 IBM 大数据平台内所有数据存储的端到端数据管理

IBM 信息集成和治理平台提供了一组集成工具套件，支持数据的集成、治理和管理。其中包含确保高质量信息、将数据纳入单一视图、管理整个数据生命周期的数据、保护信息安全、将所有数据集成到通用视图之中、确保一致的理解和知识集的工具。**IBM InfoSphere Information Server** 支持连接到 **IBM InfoSphere BigInsights**、**IBM PureData System for Operational Analytics** 和 **IBM PureData System for Analytics** 数据仓库设备以及 **IBM DB2** 分析加速器。它还能与 **IBM InfoSphere Streams** 相集成，将筛选后的事件数据发送到 **IBM InfoSphereBigInsights** 之中，以便执行进一步的分析。

IBM 利用 InfoSphere Blueprint Director 创建智能工作流，治理数据清理、数据集成、数据隐私和数据移动

IBM 将 **InfoSphere Information Server** 作为 **IBM** 大数据平台中*智慧整合*的基础。智慧整合利用 **IBM InfoSphere Blueprint Director** 构建和运行工作流，运用 **InfoSphere Information Server** 上的服务来清理、集成、保护数据并将数据分发到最适合分析工作负载的恰当分析数据存储。智慧整合的目的在于提高敏捷性，批量和实时移动与集成数据，以及创建一种集成数据管理框架，从而跨 **IBM** 大数据平台内的所有分析数据存储治理和管理数据。这能隐藏复杂性、加强自动化，为工作负载路由和按需动态工作负载优化铺平道路，从而在优化计划中实时移动数据，响应 **IBM** 大数据平台的入站查询。

IBM 信息集成与治理平台还包含数据虚拟化

面向大数据企业的 IBM 分析工具

BigSheets 支持用户分析 Hadoop 中的数据

BigSheets 可将内外数据源中的数据导入到 BigInsights Hadoop

IBM BigSheets

IBM BigSheets 是一种基于 **Web** 的用户界面工具，基于电子表格模式，允许行业用户利用 **Hadoop** 分析结构化和非结构化数据，无需技能精湛的 **IT** 专家的协助。它允许用户通过对网站执行爬网、从内部服务器和桌面选择数据、使用定制导入程序获取 **Twitter** 等特定数据源中的选定数据，从而将数据导入 **BigInsights**。导入 **BigInsights** 的数据可通过 **BigSheets** 电子表格用户界面查看，用户可以在其中利用基于 **Hadoop MapReduce** 的预先构建、定制构建的分析、筛选和分析数据。可将分析结果获取到其他工作表中，以便集中精力关注洞察。一个很好的示例是分析微博，根据所得到的积极舆情分析销售线索机遇。考虑到电子表格往往会加大查看洞察的难度，**IBM** 还在 **BigSheets** 中集成了 **Many Eyes**，为用户提供以图形化方式查看洞察的能力。通过 **BigSheets** 插件功能还可以支持第三方可视化。

此外，还包含 Many Eyes，用于改进可视化

IBM 通过 Hive 将其 Cognos BI 工具套件与 BigInsights 相集成，还集成了 IBM Netezza 和现已成为 PureData Systems 系列产品的一部分的 IBM Smart Analytics System

IBM Cognos 10

IBM Cognos 10 是 **IBM** 的旗舰级 **BI** 平台。这是一种集成系统，支持即席报告和分析、仪表板创建、记分卡、生产报告、多维分析、预算制定、规划和预测。它用于访问 **IBM** 和非 **IBM** 数据仓库与数据市场内托管的结构化数据，包括 **IBM PureData System for Analytics**（采用 **Netezza**）和 **IBM PureData System for Operational Analytics** 以及非 **IBM** 数据仓库平台。

IBM Cognos 系列产品以集成、可伸缩的方式为组织提供了起点：

- **IBM Cognos Insight**（个人、桌面分析）
- **IBM Cognos Express**（工作组 **BI**）
- **IBM Cognos Enterprise**

IBM Cognos 可作为原生应用程序扩展到移动环境。

IBM Cognos RTM 能够分析由 InfoSphere Streams 筛选并发送给它的事件数据，以支持实时异常监控

在大数据方面，**IBM Cognos BI** 平台已扩展为通过 **ODBC** 支持基于 **MapReduce** 关系数据库的报告，比如 **EMC GreenPlum**、**Teradata Aster** 和其他产品；通过 **Hive** 适配器支持 **IBM InfoSphere BigInsights** 中的报告。

IBM Cognos 还提供了 **IBM Cognos Real-time Monitoring (RTM)**，用于实时监控和可视化业务事件，让业务用户能够在需要立即采取措施时制定明智的决策。

IBM Cognos RTM 完全适合大数据环境，因为它能监控 **IBM InfoSphere Streams** 发送给它的经过筛选的事件。

实时业务洞察可与历史和预测智能同时集成到运营和管理仪表板之中

IBM Cognos RTM 和 **IBM Cognos Enterprise** 集成提供了个性化 **BI** 工作区中对运营和管理仪表板的实时感知。这种集成支持在单一用户界面内同时查看历史、实时和预测信息。此外，如果用户希望获得实时警报，还可以定义监视点和阈值。

IBM Cognos Consumer Insight (CCI)

IBM Cognos Consumer Insight 是一种专门构建的社交媒体分析应用程序，它利用 IBM 的 BigIndex 分析消费者和客户交互数据

IBM Cognos Consumer Insight 是一种专门构建的社交媒体分析应用程序，能分析从公共网站和内部数据库中存储的客户交互中收集到的海量内容。该应用程序利用 **IBM InfoSphere BigInsights** 的搜索和文本分析功能，分析非结构化社交媒体数据。目的在于判断舆情、文字关联性和其他有关社交行为的洞察。**IBM Cognos Consumer Insight** 包含与 **IBM Cognos BI** 平台相集成的预构建报告。**IBM Cognos Consumer Insight** 得出的分析结果还可整合到 **IBM SPSS** 预测模型之中，以确定在响应特定客户行为和意图时应采取怎样的措施。

IBM SPSS

IBM SPSS 用于在 IBM 大数据平台中构建和部署高级分析

IBM SPSS 是 **IBM** 用于构建和部署高级分析、开发自动化决策管理应用程序的工具套件。利用 **IBM SPSS**，超级用户即可设计和构建预测和统计模型，自动化分析数据。这些模型可部署在以下产品之中：

- **IBM InfoSphere Streams** 应用程序，以分析动态大数据
- **IBM PureData System for Analytics**（采用 **Netezza** 技术）和 **Netezza Platform Software**，以便对静止的结构化数据执行数据库内高级分析
- **IBM PureData System for Operational Analytics** 上的 **IBM InfoSphere Warehouse DB2 DBMS**，以实现数据库内高级分析和运营 **BI**。

IBM 正在对 SPSS 进行扩展，以利用 map/reduce 扩展 BigInsights 上的分析

除了 **Hadoop Mahout** 高级分析库之外，**IBM** 还会在 **IBM InfoSphere BigInsights** 内提供 **SPSS** 开发的预测分析，从而自动分析 **Hadoop HDFS** 和 **Hive** 中的海量多结构化数据。

IBM Vivisimo

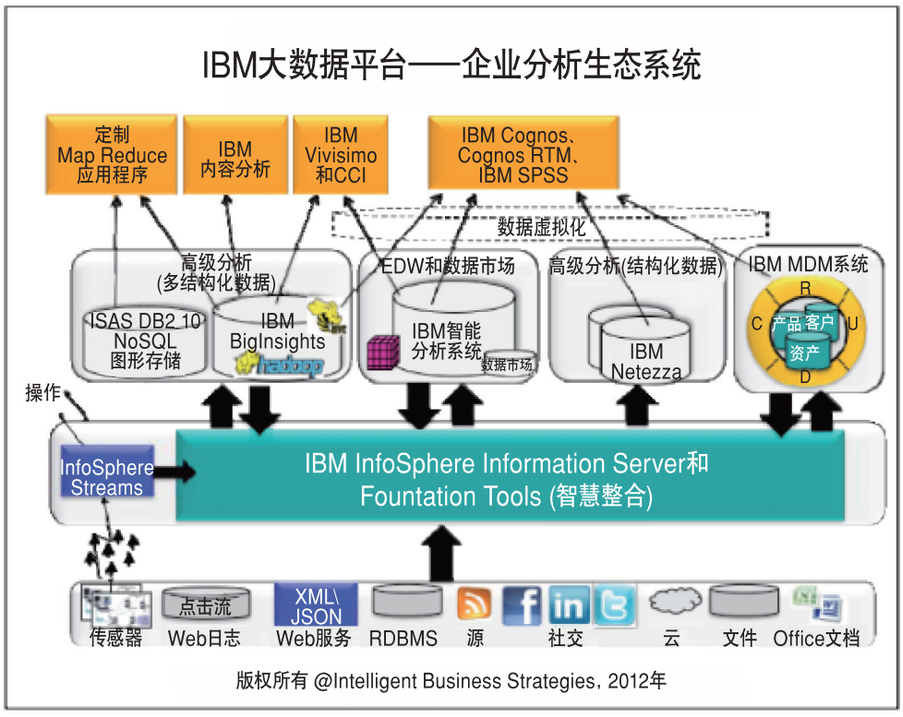
IBM Vivisimo 是一种基于搜索的平台，用于开发自由形式的分析应用程序，分析大数据平台上的数据

IBM Vivisimo 是 **IBM** 大数据平台的最新成员。该产品能加速创造价值的进程，它允许组织联合海量结构化数据和非结构化数据的探索和浏览，同时将数据保留在原本的多个数据源之中。这是通过爬取结构化和非结构化大数据源以构建搜索索引来实现的。**IBM Vivisimo** 还提供了工具，支持在这些索引的基础之上开发应用程序，让组织能够执行探索式分析和基本方面的浏览，从而发现宝贵洞察。在问题的解决方案尚不明朗，需要无限制地自由探索和浏览多结构化数据时，这种产品极为有用。此外，如果在业务运营中需要迅速响应特殊问题，但尚未开发任何报告或数据结构，那么该产品也极为适合此类情况下的即席分析。

这些组件如何融合在一起以实现端到端的业务洞察

所有这些组件融合在一起，扩展了传统数据仓库环境，建立起一种能处理传统和大数据工作负载的企业分析生态系统

下图展示了所有这些组件如何融合在一起。



结束语

如今的企业需要更为强大的分析能力，以分析全新的结构化和多结构化数据源

如今，许多企业现在都处于这样一种状态：业务用户总是需要分析大量全新的结构化和多结构化数据源，以生成新洞察，占据竞争优势。数据本身的复杂性不断增加，包括数据创建的速度、涉及到的数据量以及需要分析的新数据类型。分析复杂性也在不断增加，促成了对于新分析技术和工具的需求。此外，社交互数据等新数据源的数据质量也逐渐成为问题。为了响应这种更高的复杂性，供应商也在构建专为某些分析工作负载而设计的新分析平台。大多数组织都认为，将所有这些新数据纳入企业数据仓库并非正确方法。需要识别新的大数据工作负载，并选择恰当的分析平台。需要扩展传统 **BI** 环境以支持企业数据仓库，以及这些全新的工作负载优化分析系统，同时为所有分析数据存储提供一致的数据。除此之外，企业还需要为用户屏蔽这种较为复杂的分析环境，使用户能轻松访问多个分析系统中的数据。查询需要自动指向最适合分析工作负载的恰当系统，数据需要轻松在分析系统之间移动，以便在最佳位置予以处理。此外，组织还需要集成用户界面，使之能轻松使用洞察，并轻松查看全部洞察。

支持特定分析工作负载的新技术纷纷涌现

现在需要扩展传统数据仓库环境以支持这些全新的大数据分析工作负载

现在的挑战在于找到一家能够交付新型分析生态系统的技术合作伙伴，以全面支持业务用户需要的传统和大数据分析工作负载。此外，这一切都必须在保证数据得到治理、一切得到集成的前提下实现。

IBM 大数据平台直接应对创建这种新型分析环境的挑战

纵观市场，**IBM** 已经确定了这项需求。其大数据平台是一种综合全面的产品，全面支持所有分析工作负载。其中包含强大的 **IBM InfoSphere BigInsights Hadoop** 产品、**IBM Smart Analytic System** 或带有内置的 **NoSQL** 图形存储的 **IBM PureData System for Operational Analytics**（运行 **InfoSphere Warehouse**），采用 **Netezza** 技术处理复杂结构化数据分析的 **IBM PureData System for Analytics**，对动态数据执行实时分析的 **InfoSphere Streams** 以及加速运营分析的 **IBM DB2 Analytics Accelerator (IDAA)**。此外，支持整个 **IBM** 大数据平台的 **InfoSphere** 信息管理工具套件也为所有分析数据存储和 **IBM MDM Server** 提供了一致的可靠数据。它还支持数据虚拟化，从而简化数据访问。**InfoSphere Blueprint Director** 能够在所有平台之间移动数据，并在工作流执行过程中调用数据库内分析。**InfoSphere Streams** 可以将事件推送到 **IBM Cognos RTM**，**IBM Cognos RTM** 本身可与 **IBM Cognos 10** 相集成，支持查看实时洞察。**IBM Cognos** 还能连接到 **BigInsights** 和所有关系型分析数据存储。**IBM SPSS** 可将分析推送到 **PureData System for Analytics**、**PureData System for Operational Analytics** 和 **InfoSphere Streams** 以获得大数据性能，**IBM Vivisimo** 还支持基于自由式搜索的分析。所有这一切都让 **IBM** 成为任何大数据竞争环境中最后候选人名单中的主要竞争者。

IBM 大数据平台让 IBM 成为支持端到端分析工作负载的主要竞争者

关于 Intelligent Business Strategies

Intelligent Business Strategies 是一家研究和咨询公司，主要目标是帮助企业了解和利用商业智能、分析处理、数据管理和企业业务集成领域的新发展。这些技术彼此结合，能帮助组织转变成为*智能企业*。

作者

Mike Ferguson 是 **Intelligent Business Strategies Limited** 的管理总监。作为分析师和顾问，他专攻商业智能和企业业务集成。在超过 **31** 年的 **IT** 从业经验中，**Mike** 为数十家企业提供了有关商业智能战略、大数据、数据治理、主数据管理、企业架构和 **SAO** 的咨询服务。他是活跃在全球各地各类活动中的发言人，撰写了无数篇文章。他的许多文章和博客文章都提供了有关行业的深入见解。在此之前，他是关系模型的发明者 **Codd and Date Europe Limited** 的总裁和联合创始人，还担任过 **Teradata** 公司的 **Teradata DBMS** 首席架构师和独立分析机构 **Database Associates** 的欧洲管理总监。他在广受欢迎的大数据分析、商务智能和数据仓库新技术、企业数据治理、主数据管理和企业业务集成高级讲习班中授课。

Water Lane, Wilmslow
Cheshire, SK9 5BG
England
电话： (+44)1625 520700
Internet URL：www.intelligentbusiness.biz
电子邮件：info@intelligentbusiness.biz

设计用于分析的大数据平台的架构