Oracle 企业架构白皮书
2012 年 8 月

Oracle 信息架构：架构师大数据指南

免责声明

以下内容旨在概述我们产品的总体发展方向。该内容仅供参考，不可纳入任何合同。该内容不构成提供任何材料、代码或功能的承诺，并且不应该作为制定购买决策的依据。所描述的有关 Oracle 产品的任何特性或功能的开发、发布和时间安排均由 Oracle 自行决定。

引言

大多数组织都在捕获和共享来自日益增长的数据源的不断攀升的数据。如果您的组织也面临着这种状况，那么当前的挑战在于如何快速而缜密地管理大容量和高速的数据流。

大数据的处理宗旨是从海量的非结构化信息中找到有价值的内容。如今，各企业都在投资打造能够解读消费者行为、检测欺诈甚至预测未来的解决方案。McKinsey 在 2011 年 5 月发布的一份报告显示，领先企业都在利用大数据分析来获取竞争优势。据该公司估计，借助大数据，零售企业可以将利润提高 60%。

为了支持这些全新的分析方案，IT 策略如雨后春笋般涌现，最新的技术包括针对海量信息源的强力攻击以及通过专业化并行处理和索引机制来实现数据过滤。用户可以关联不同时间和不同含义的结果，并且将这些结果与传统企业数据源合并在一起。而最新的数据发现技术包括卓越的可视化工具和交互式语义查询体验。知识工作者和数据科学家可以对过滤后的数据进行分类筛选，不断提出不相关的探索性问题。随着这些支持性技术从毕业生研究课题延伸到企业 IT 领域，IT 战略分析师、规划师和架构师不仅需要了解这些技术，而且还需要将其应用到整个企业中。

规划大数据架构时不仅要理解该架构与传统架构间的差异，还要掌握将新元素集成到现有系统（包括数据库和 BI 基础架构、IT 工具以及最终用户应用程序）中的方法。通过自主发布硬件、软件以及推出全新的合作伙伴计划，Oracle 旨在改变大数据投资的经济性以及解决方案的适用性。真正的行业挑战并不是将大数据视为一个专门的科研项目，
而是将其集成到主流 IT 系统中。

本白皮书将讨论如何数将大数据功能添加到用户的整体信息架构中，如何从企业架构的角度来规划大数据的采用，此外还将描述一些重要的用例。

大数据为何与众不同？

大数据是指难以实现存储、搜索、共享、可视化和分析的大型数据集。总体而言，大数据的数量级超过了传统的数据处理能力和最大型数据仓库的容量。举例来说，航空飞机每飞行 30 分钟就会采集 10 TB 的传感器数据。而纽约证券交易所这种常规的高性能计算环境一天才采集 1 TB 的结构化交易数据。常规的结构化企业数据仓库的容量为 TB 和 PB 级。而大数据的容量则为 PB、EB 级，甚至可能很快达到 ZB 级！除了容量巨大以外，大数据分析解决方案还需应对数据内容和数据结构的不可预测或不可预知性。这些分析方案将过滤掉低价值或低密度的数据，展现出高价值或高密度的数据。因此，通常需要选用最新的专用分析技术。大数据面临着一系列重要的架构挑战。

一般而言，大数据具备大数据量、高速和多样性等特征，但大数据的独特之处在于发现数据值的方式。传统的商务智能大多通过简单汇总已知值来揭示结果，例如通过各个订单的销售额得出年初至今的销售额。与这种方式不同，大数据通过一个不断完善的建模流程来发现数据值，这个流程包括：提出假设，创建统计、可视化或语义模型，验证以及提出新的假设。这通常需要安排特定人员阐释可视化信息或执行交互式、基于知识的查询，或者开发自适应的“机器学习”算法来发现数据的含义。然而，由于这类算法具有针对性，因此生命周期可能十分短暂。

如今，随着数据传播渠道和数据种类的增加，大数据的数量与日俱增。其中一些新数据源中的数据源自社交媒体、网络和软件日志、摄像机、信息传感移动设备、航空传感技术、基因组学和医疗纪录。

企业已经意识到此类信息的竞争优势，而现在正是运用这些数据的大好时机。

大数据用例：个性化保费

首先，大数据为保险行业的传统业务流程带来了商机，即算出兼具竞争优势和高收益的保费。

为了提高竞争力，保险公司希望尽可能降低客户的保费，同时将客户范围限定在索赔可能性最小的人群中，从而最大限度地提高利润。解决此问题的一种途径就是收集关于客户驾驶习惯的更多详细信息，并对风险进行评估。

实际上，一些保险公司已经开始利用客户汽车上的传感器来采集客户的驾驶习惯数据。这些传感器可捕获驾驶数据，例如行驶路段、行驶里程、行驶时间和刹车紧急程度等。这些数据可用于评估驾驶员风险。保险公司将个人驾驶状况与其他统计信息（例如所在地区驾驶员的平均驾驶里程以及道路的高峰时段）进行对比，随后将驾驶员风险及保险精算信息与策略及档案信息相关联，从而计算出具有高竞争性、高收益的保险费率。这样，保险公司便能为客户提供个性化的保险方案。大数据分析带来的独特功能正在推动保险行业的变革。

为满足大数据分析的要求，企业需要采集、存储和关联大量的连续数据。Hadoop 是采集和精简汽车传感器数据的绝佳选择。主数据和特定的参考数据（包括客户档案信息）可以存储在当前的 DBMS 系统中，同时，NoSQL 数据库用来捕获和存储更加动态化、格式多样和变更频繁的参考数据。Project R 和 Oracle R Enterprise 可用于分析私有保险公司的数据以及从公共数据源中捕获的数据。最后，将 MapReduce 结果载入到现有的 BI 环境中，用户即可对数据进行进一步的挖掘和关联。借助这些新工具，企业将能够满足对存储、检索、建模和处理的需求。

在本例中，传统的业务流程和支持性数据（主数据、事务数据和分析数据）可为盈利模型添加与统计相关的信息，从而推动行业的创新。

每个行业都有相关的案例。大数据源的形式多样，包括日志、传感器、文本、空间数据等。作为 IT 战略分析师、规划师和架构师，我们清楚企业多年以来一直在尝试从所有这些非结构化数据中查找相关信息。但直到最近，可满足企业需求的经济性方案才得以面世。

您采用何种方式为业务线提供完善的数据服务，以便在实现数据查找的同时改善和创新核心业务流程？它们对您的信息架构、您的企业、您的分析能力有何影响？

信息架构运营模式对比

大数据在许多方面有别于其他数据种类。下表对大数据以及 Oracle 信息架构框架 (OIAF) 中描述的其他数据种类进行了特性比较和对比。

数据种类	结构	数据量	描述	示例
主数据	结构化	低	对组织具有战略价值的企业级数据实体。通常具备非易失性和非事务性。	客户、产品、供应商和位置/地点。
事务数据	结构化和半结构化	中 — 高	在业务运营和业务流程中捕获到的业务事务。	采购纪录、询价和付款。
参考数据	结构化和半结构化	低 — 中	内部管理的或来自外部的事实数据，可以支持组织高效处理事务、管理主数据和提供决策支持功能。	地理数据和市场数据。
元数据	结构化	低	定义为“关于数据的数据”。作为标准化描述和运营的抽象层。如集成、智能和服务。	数据名、数据维度或单元、数据实体定义或量度的计算公式。
分析数据	结构化	中 — 高	业务运营和事务数据的衍生，用于满足报告和分析需求。	位于数据仓库、数据集市和其他决策支持应用程序中的数据。
文档和内容	非结构化内容	中 — 高	文档、数字图像、地理空间数据、多媒体文件。	索赔申请表、医学影像、地图和视频文件。
大数据	**结构化、半结构化以及非结构化**	**高**	**难以实现存储、搜索、共享、可视化和分析的大型数据集。**	**用户和机器生成的内容，包括源自社交媒体、网络和软件日志、摄像机、信息传感移动设备、航空传感技术的数据和基因组学数据。**

表 1：数据种类定义（Oracle 信息架构框架）

这些不同的特性影响了信息的捕获、存储、处理、检索和保护方式。随着大数据产品的发展，您可以在现有投资中寻找协同优势，充分利用专业化组织的技能、设备、标准和治理流程，从而最大限度地降低架构风险。

用户可采用的功能如下：

数据种类	结构	存储和检索	建模	处理和集成	使用
主数据、事务数据、分析数据、元数据	数据库、应用程序和用户访问	RDBMS/SQL	预定义的关系建模或维度建模	ETL/ELT、CDC、复制、消息	BI 和统计工具、运营应用程序
参考数据	平台安全性	XML/xQuery	灵活、可扩展	ETL/ELT、消息	数据使用基于系统
文档和内容	基于文件系统	文件系统/搜索	自由格式	操作系统级文件移动	内容管理
大数据 **— 网志** **— 传感器** **— 社交媒体**	文件系统和数据库	分布式文件系统/noSQL	灵活（键值）	Hadoop、MapReduce、ETL/ELT、消息	BI 和统计工具

表 2：数据种类特性（Oracle 信息架构框架）

大数据架构功能

下面，我们将简要概述大数据的功能及其主要技术：

存储和管理功能

Hadoop 分布式文件系统 (HDFS)：

- Apache 的开源分布式文件系统，详情请参阅 http://hadoop.apache.org
- 预计可在高性能商用硬件上运行
- 可跨三个节点进行高可伸缩性存储和自动数据复制，从而实现容错功能
- 可跨三个节点进行自动数据复制，无需备份
- 一次写入，多次读取

Cloudera Manager：

- Cloudera Manager 是适用于包含 Apache Hadoop 的 Cloudera 发行版的一款端到端的管理软件，详情请参阅 http://www.cloudera.com
- Cloudera Manager 实时展现整个集群内节点和正在运行的服务的情况；提供一个中央位置来将配置更改应用到整个集群；同时引入全方位的报告和诊断工具来帮助优化集群性能和利用率。

数据库功能

Oracle NoSQL：（点击此处查看更多信息）

- 动态和灵活的模式设计。高性能键值对数据库。键值对是预定义模式的替代方案，适用于不可预测的动态数据。
- 可以在无行列结构的情况下有效地处理数据。采用主键 + 次键模式，可以在单次 API 调用中实现多次记录读取。
- 高度可伸缩的多节点、多数据中心、容错、ACID 运营
- 简洁的编程模式、随机索引读取和写入
- 不只是面向 SQL。简单的模式查询和自定义开发的解决方案可以访问 Java API 等数据。

Apache Hbase：（点击此处查看更多信息）

- 支持随机、实时读取/写入访问
- 严格一致的读取和写入
- 自动和可配置的表分片
- 支持在区域服务器间实现自动故障切换

Apache Cassandra：（单击此处查看更多信息）

- 数据模型提供的列索引具有日志结构更新、物化视图和内置缓存等特性
- 为每个节点提供量身打造的容错功能，可跨多个数据中心进行复制
- 可为每次更新选择同步或异步复制

Apache Hive：（点击此处查看更多信息）

- 提供的工具可以对直接存储在 Apache HDFS 中或其他数据存储系统（如 Apache HBase）中的文件进行轻松的数据提取/转换/加载 (ETL) 处理
- 采用类似于 SQL 的简单查询语言，即 HiveQL
- 通过 MapReduce 执行查询

处理功能

MapReduce：

- 由 Google 在 2004 年定义。（点击此处查看白皮书原文）
- 将问题细分为更小的子问题
- 可以在成千上万个节点中进行数据负载分配
- 可通过 SQL 和基于 SQL 的 BI 工具公开

Apache Hadoop：

- 领先的 MapReduce 实施方案
- 高度可伸缩的并行批处理
- 高度可定制的基础架构
- 可在集群内写入多个副本，从而实现容错功能

数据集成功能

Oracle Big Data Connector、Oracle Loader for Hadoop、Oracle Data Integrator：（点击此处了解 Oracle 为大数据打造的数据集成解决方案）

- 将 MapReduce 结果导出至 RDBMS、Hadoop 和其他目标
- 将 Hadoop 连接至关系数据库进行 SQL 处理
- 内置图形用户界面集成设计器，可生成专用于迁移和转换 MapReduce 结果的 Hive 脚本
- 通过并行数据导入/导出优化处理
- 可安装在 Oracle 大数据机或通用的 Hadoop 集群上

统计分析功能

开源的 Project R 和 Oracle R Enterprise：

- 用于统计分析的编程语言（点击此处了解 Project R）
- Oracle 数据库的 SQL 扩展功能，可高效执行数据库内统计分析（点击此处了解 Oracle R Enterprise）
- 借助 Oracle R Enterprise，用户可以重复利用已有的 R 脚本，而无需进行任何修改

大数据架构

在本节中，我们将进一步讨论大数据的总体架构。

传统信息架构功能

为便于理解大数据的高级架构层面，我们首先回顾一下结构化数据格式良好的逻辑信息架构。在插图中，大家可以看到两种数据源，它们利用集成（ELT/ETL/更改数据捕获）技术将数据传输至 DBMS 数据仓库或运营数据存储中，随后再通过多种分析功能来获取所需信息。这些分析功能包括：信息板、报告、EPM/BI 管理软件、汇总和统计查询、文本数据语义解释以及针对高密度数据的可视化工具。此外，一些组织还实施了跨项目的监管和标准化，并且可能通过企业级管理完善了信息架构的功能。

图 1：传统信息架构功能

关键的信息架构准则包括将数据视为资产对其进行价值、成本和风险分析，从而确保数据的及时性、质量和准确性。同时，EA 监管职责将建立和维持一种平衡的治理方式，包括利用卓越中心来实现标准管理和培训。

增加大数据功能

大数据架构的处理功能需要满足对大容量、高速度、多样化和价值的需求。独特的分布式（多节点）并行处理架构可对这些大型数据集进行解析。各种不同的技术策略可满足实时和批处理的需求。而实时的键值数据存储，如 NoSQL，可以实现基于索引的高性能检索。对于批处理，一种称为“Map Reduce”的技术可以根据特定的数据发现策略来执行数据过滤。发现过滤好的数据后，用户可直接对数据进行分析，将其载入到其他非结构化数据库中，发送至移动设备或合并到传统数据仓库环境中并与结构化数据相关联。

图 2：大数据信息架构功能

大数据是新型的非结构化数据，除此之外，它还有两个与众不同之处。首先，鉴于数据集的大小，原始数据不能直接迁移到数据仓库中。然而，在 MapReduce 处理后，我们可以将“化简结果”集成到数据仓库环境中，这样便可利用传统的 BI 报告、统计、语义和关联功能。理想方案是利用分析功能，将传统 BI 平台与大数据可视化和查询功能相结合。其次，可以创建沙盒环境来简化 Hadoop 环境中的分析。

对许多用例而言，大数据需要捕获不断变化和不可预测的数据。为了分析这类数据，企业需要一种全新的架构。在零售行业，一个很好的示例是捕获实时客流量，以便进行店内促销。为了追踪楼层展示和促销的有效性，企业必须借助可视化或查询工具对客户移动方式和行为进行交互式探索。

在其他用例中，只有将大数据与其他企业数据（结构化的数据）关联在一起，企业才能更好的完成分析过程。在消费舆情分析案例中，捕获正面或负面的社交媒体评论具有一定的价值，而将评论数据与利润贡献最高或最低的客户信息关联在一起则可进一步扩大数据的价值。因此，大数据 BI 需要上下文和理解功能。借助强大的统计和语义工具，用户可以从海量数据中找到所需信息并对未来进行预测。

总的来说，大数据架构面临的挑战是在将大数据与其他数据相关联的同时满足快速使用和快速理解数据的需求。

值得特别注意的是，尽管关键信息架构准则是相同的，但应用这些准则的策略可以不同。举例来说，如何将大数据视为资产？我们一致认为，大型的高密度数据集具有隐藏价值。但我们该如何对大数据进行评估？如何进行优先级排序？关键在于对最终目标的考量。专注于业务价值，理解大数据对支持业务决策的重要性以及未知隐藏模式的潜在风险。

应用架构准则的另一个示例是数据治理。大数据的质量和准确性需求有时差异巨大。对用户舆情数据采用严格的数据精确规则将过滤掉大量有用信息，而数据标准和通用定义对欺诈检测而言仍十分重要。

在此重申，应用核心信息架构准则和实践固然十分重要，但在应用时仍需注重与大数据之间的相关性。此外，大数据的 EA 职责仍然不变，包括提高成功率、集中培训以及建立标准。

集成的信息架构

与现有 BI 生态系统的集成度不足是阻碍企业采用 Hadoop 的一个重要因素。.目前，传统 BI 与大数据生态系统相互独立，导致集成数据分析成为一个难题。这样一来，大多数业务用户或高管都无法采用这种分析方式。

早期采用大数据的企业通常通过编写自定义代码来将大数据处理结果迁回至数据库，或通过开发自定义解决方案来报告和分析处理结果。这些方式对企业 IT 来说可能并不具备经济性和可行性。首先，它将导致一次性编码和标准数量激增。其架构也会影响 IT 经济性。独立的大数据方案会增加冗余投资风险。此外，大多数企业的员工和技能水平尚不足以完成此类自定义开发任务。

一种更好的方案是将大数据结果整合到现有的数据仓库平台中。信息的影响力取决于数据关联的能力。我们需要将不同数据源、处理需求相结合，从而实现及时和有价值的分析。

以下是将传统信息架构与大数据架构相结合的 Oracle 整体功能图：

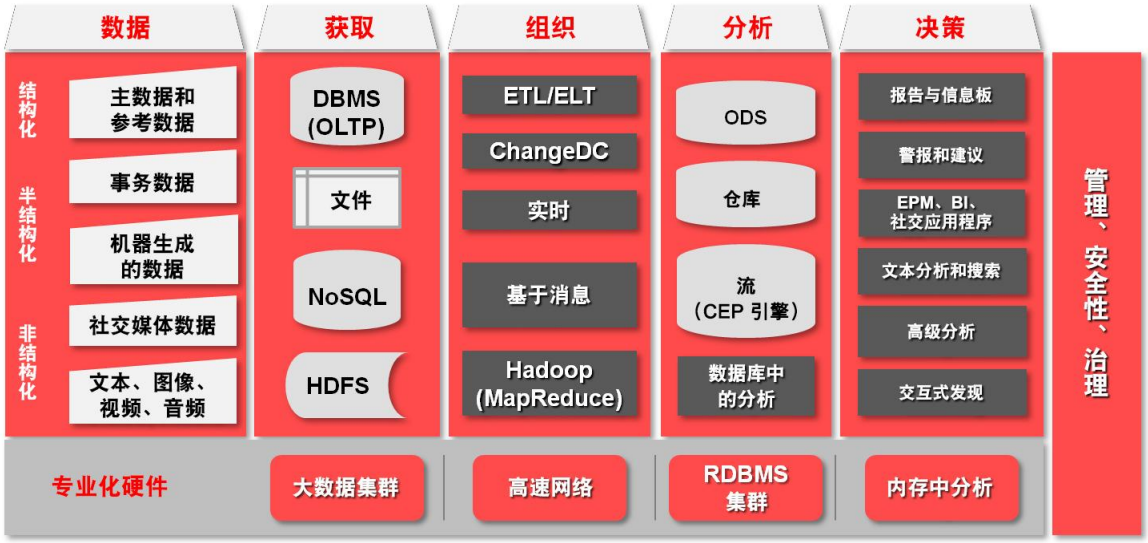

图 3：Oracle 集成信息架构功能

捕获到各种数据之后，企业可在传统的 DBMS、简单文件或分布式集群系统（如 NoSQL 和 Hadoop 分布式文件系统，即 HDFS）中存储和处理这些数据。

从架构上来说，打破分裂局面的关键组件是中间的集成层。集成层必须贯穿所有数据类型和数据域，并弥合传统与新型数据捕获及处理框架之间的差异。数据集成功能需要满足对速度和频率的各种需求，要应对日益增长的海量数据，同时，还要消除数据结构间的差异。企业需要找到支持双向集成 Hadoop/MapReduce 与数据仓库和事务数据存储的技术。

在下一层中，企业将“化简结果”从大数据处理输出载入到数据仓库中，以便实施进一步的分析。

同时，在处理大数据以查找模式（如检测欺诈活动）时，企业还需要能够访问其结构化数据，例如客户档案信息。

与对事务数据的处理相同，大数据处理输出将会载入到传统 ODS、数据仓库和数据集市中，以便于实施进一步的分析。该层的另一个组件是 Complex Event Processing 引擎，专用于实时分析流式数据。

除了报告、信息板和查询等传统组件以外，BI 层还包含高级分析、数据库内统计分析和高级可视化组件。

治理、安全性和运营管理同样也覆盖了整个企业级数据和信息架构。

借助该架构，业务用户将不会再看到传统数据与大数据间的差别，他们甚至不需要了解传统事务数据与大数据之间存在着差别。当用户浏览各种数据和信息集、检验假设、分析模式以及制定决策时，数据和分析流程都将实现无缝衔接。

制定大数据架构决策

信息架构或许是 IT 所面临的最为复杂的领域，它是投资的终极回报。在当今的经济形势下，业务必须由有价值、准确和及时的信息驱动。此外，大数据的出现增加了这一问题的复杂程度。但是，要经济高效地获取数据
和信息，企业必须对业务和 IT 进行权衡。

关键驱动因素

下面概述了企业在制定架构决策时需要考虑的业务和 IT 驱动因素。

业务驱动因素	IT 驱动因素
• 更好的洞察力	• 降低存储成本
• 更快的处理周期	• 减少数据迁移
• 准确性和及时性	• 加快上市速度
	• 标准化的工具集
	• 简化管理和运营
	• 安全性和治理

三个用例的架构模式

大数据处理和分析的业务驱动因素普遍存在于各行各业中，并且将很快成为新服务交付和流程分析的核心。在某些情况下，它们将利用我们熟悉的新型数据源（例如移动设备上的应用程序、位置服务、社交网络和电子商务）带来新的商机。设备生成的数据中的细微差别有助于远程监视患者情况、个人健身活动、驾驶习惯，实现基于位置的存储迁移和预测消费者行为。但是，大数据同样也可以对传统企业业务流程加以改进，例如通过销售和服务网站以及呼叫中心功能实现的基于文本和舆情的客户交互、人力资源简历分析、产品生命周期管理中从检测到性能增强的工程变更管理、制造执行系统中的工厂自动化和质量管理等。

在本节中，我们将探讨以下三个用例，并简要描述架构决策和技术组件。用例 1：零售业 — 网志分析。用例 2：金融服务业 — 实时事务检测。用例 3：保险业 — 非结构化和结构化数据关联。

用例 1：初始数据挖掘

第一个用例与零售业相关。在圣诞节期间，某零售业巨头的网络渠道销售额非常不理想，于是期望利用其在线购物网站来改善客户体验。潜在调查范围包括购物者的网志和产品/网站评论。了解导航模式，尤其是与购物车丢弃相关的导航模式十分有益。在进行大型投资前，企业需要明确这些数据的价值与相关成本。

该零售商的 IT 部门面临着以下方面的挑战：缺乏相应的技能来满足新数据集和处理能力的需求，无法处理大容量数据。传统 SQL 工具通常是业务和 IT 的第一选择，但将所有数据载入到关系数据库管理平台中在经济上是不可行的。

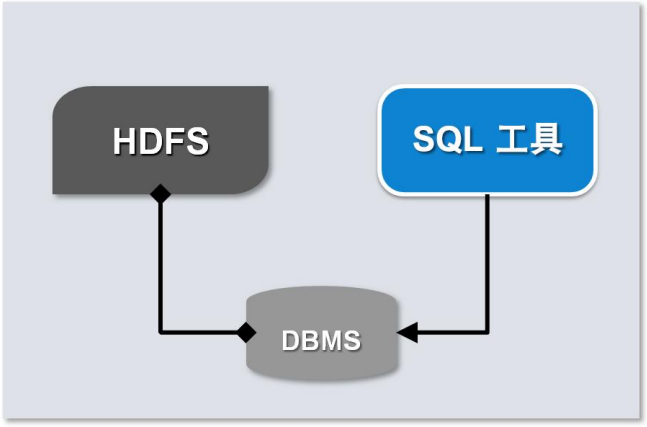

图 4：用例 1：初始数据挖掘

如上图所示，此设计模式需要通过 DBMS 系统来挂载 Hadoop 分布式文件系统，以便利用传统 SQL 工具来探索数据集。其主要优势包括：

- 无需迁移数据
- 利用对数据库和 SQL 或 BI 工具的现有投资以及相关技能

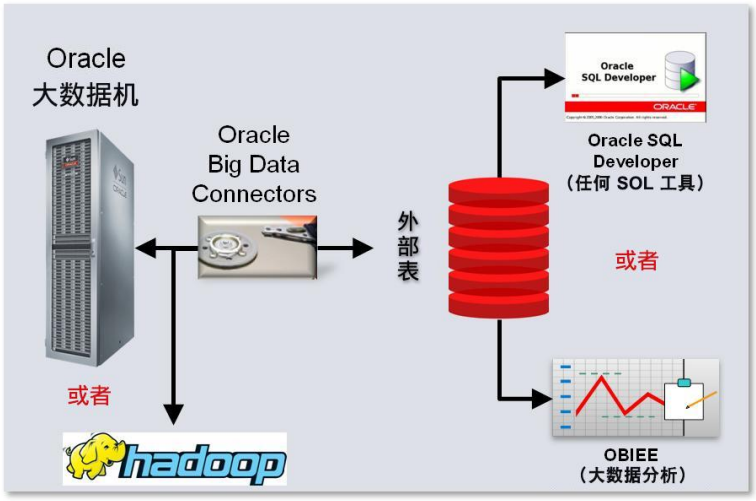

图 5：用例 1：架构决策

上图展示了通过 Oracle 产品来满足这些标准的逻辑架构。

该架构中的关键组件包括：

- Oracle 大数据机（或其他 Hadoop 解决方案）：
  - 依托包含 Apache Hadoop 的 Cloudera 发行版 (CDH) 的完整版本，用于存储日志、评论和其他相关大数据。
- Oracle Big Data Connectors：
  - 创建优化的数据集，确保可在 Oracle Database 11g 和 Oracle Enterprise R 中实现高效的加载和分析。
- Oracle Database 11g：
  - 外部表：Oracle 数据库的一项特性，可以表格形式呈现文件系统中存储的数据，并且可以透明地在 SQL 查询中使用。
- 传统 SQL 工具：
  - Oracle SQL Developer：具有图形用户界面的开发工具，允许用户通过 SQL 访问存储在关系数据库中的数据。

- 还可以使用 OBIEE 之类的商务智能工具来访问 Oracle 数据库中的数据。

总的来说，该场景中的关键架构决策的目的就是避免数据迁移、最大限度降低处理需求和减少投资。借助该架构，上述零售商可以通过数据库和 SQL 界面直接访问 Hadoop 数据，从而完成初始数据挖掘。

用例 2：使用大数据进行复杂事件处理

第二个用例与金融服务行业相关。大型金融机构在检测金融犯罪和恐怖活动过程中发挥着至关重要的作用。然而，IT 部门是否能够应对以下挑战将严重制约金融机构满足上述要求的能力：

- 扩充反洗钱法，以涵盖日益增长的犯罪活动，如赌博、集团犯罪、贩毒和资助恐怖主义活动
- 捕获、存储和评估日益增长的信息
- 关联来自多种数据源、具有不同格式的数据

其 IT 系统必须能够自动采集和处理来自多种数据源的大量数据，这些数据源包括货币交易报告 (CRT)、可疑活动报告 (SAR)、可转让票据日志 (NIL)、基于互联网的活动和事务等。

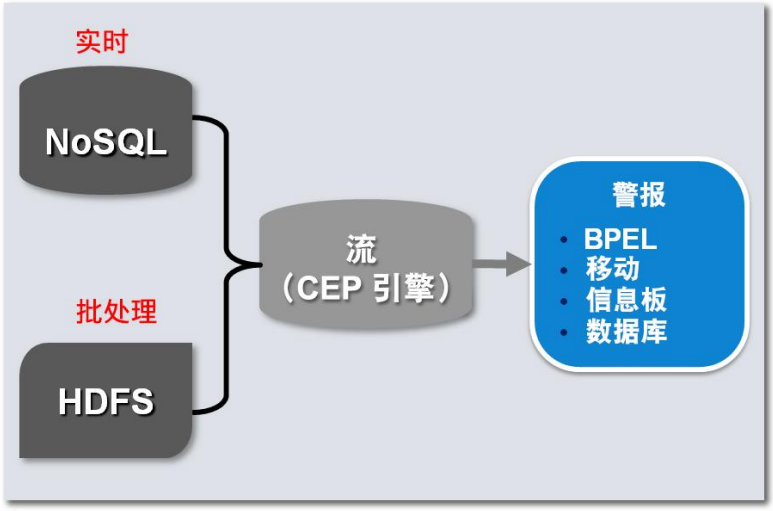

图 6：用例 2：使用大数据进行复杂事件处理

最理想的场景会涵盖所有历史档案变更和交易纪录，以便从多个汇总级别和层级精准地确定每个账户、客户、交易对象和法律实体的风险率。然而，由于受到处理能力和存储成本的限制，传统方式并不可行。而借助 HDFS，用户可以整合所有详细数据点，计算风险概况并将其发送至 CEP 引擎，从而为风险模型奠定基础。

在此场景中，NoSQL 数据库将从数据结构灵活的多个数据源中捕获并存储大量低延迟数据，并借助 CEP 引擎来提供实时数据集成，从而实现自动警报和信息板，并触发业务流程来采取相应的措施。

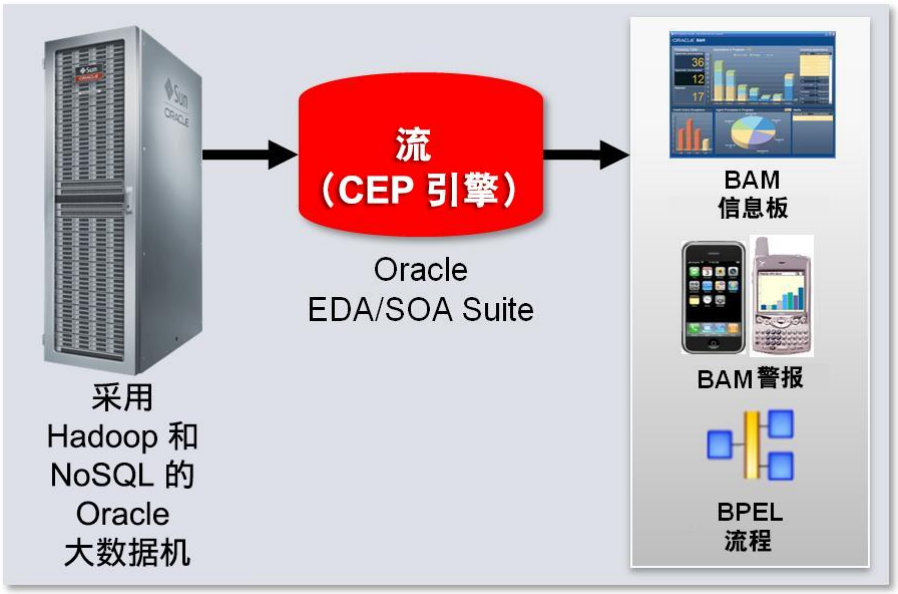

图 7：用例 2：架构决策

上方的逻辑图展示了该架构的主要组件，包括：

- Oracle 大数据机（或其他 Hadoop 解决方案）：
  - 依托包含 Apache Hadoop 的 Cloudera 发行版 (CDH) 的完整版本，用于存储日志、评论和其他相关大数据。
  - NoSQL：专用于捕获数据结构灵活的低延迟数据，并支持快速查询。
  - MapReduce：通过处理大量数据来精简和优化数据集，以便将数据集载入到数据库管理系统中。
- Oracle EDA：
  - Oracle CEP：流式复杂事件引擎，专用于持续处理传入数据、分析和完善模式，并在检测到可疑活动时引发事件。

- Oracle BPEL：业务流程执行语言引擎，可根据引发事件来定义流程和适当的操作。
- Oracle BAM：实时业务活动监视信息板，用于提供即时洞察和生成所需的操作。

总的来说，该架构的关键准则是将大数据与事件驱动架构相集成，以满足复杂的监管要求。尽管此架构中并未涵盖数据库管理系统，但引发的事件、后续处理事务和纪录仍将存储在数据库中，以便于进行事务处理或满足未来分析之需。

用例 3：使用大数据进行联合分析

在第三个用例中，我们将继续讨论前面章节中提到的保险公司。简而言之，该保险巨头需要捕获大量记录客户驾驶习惯的传感器数据，以具有成本效益的方式存储数据并对这些数据进行处理，从而确定趋势和模式，然后再将最终结果与当前捕获到的事务数据、主数据和参考数据集成在一起。

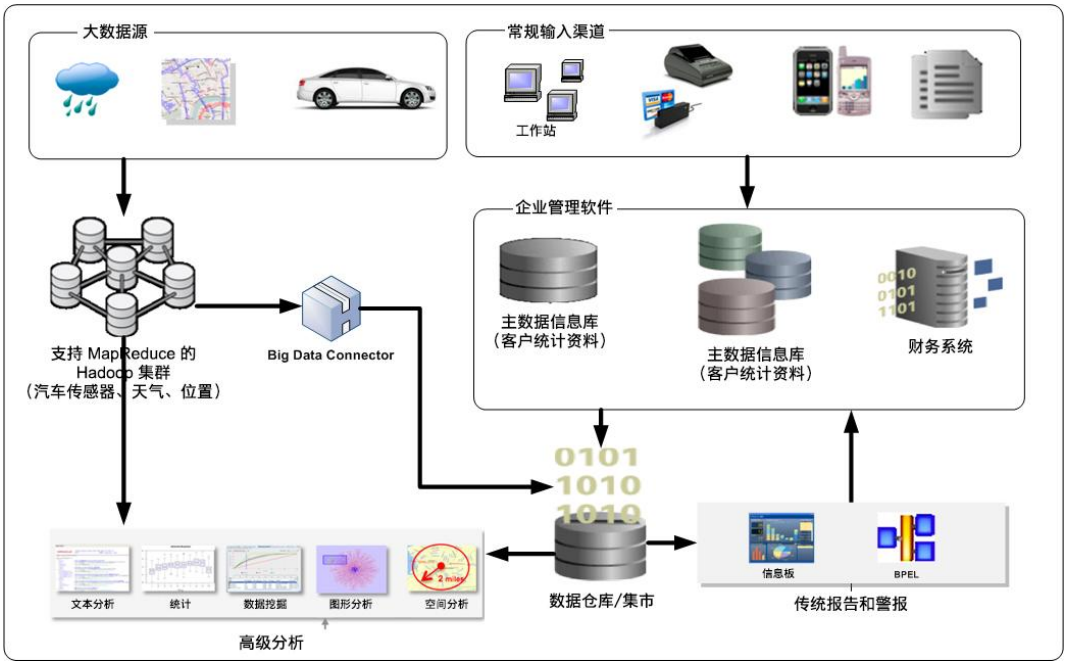

图 8：用例 3：数据流架构图

大数据与结构化数据的集成是该架构面临的最大挑战。

以下高级概念图展示了上述需求。

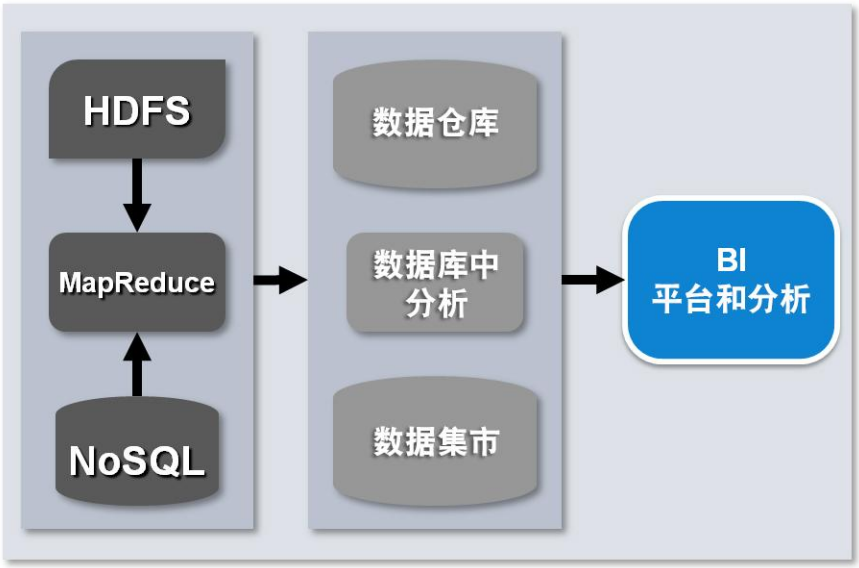

图 9：用例 3：联合分析的概念架构

企业需要将大量传感器数据传输到集中化环境中加以存储，该环境支持灵活的数据结构、快速处理、可伸缩性和并行机制；需要利用 MapReduce 功能来处理低密度数据，从而识别出模式并获取趋势洞察；此外，还需要将最终结果集成到含有结构化数据的数据库管理系统中。

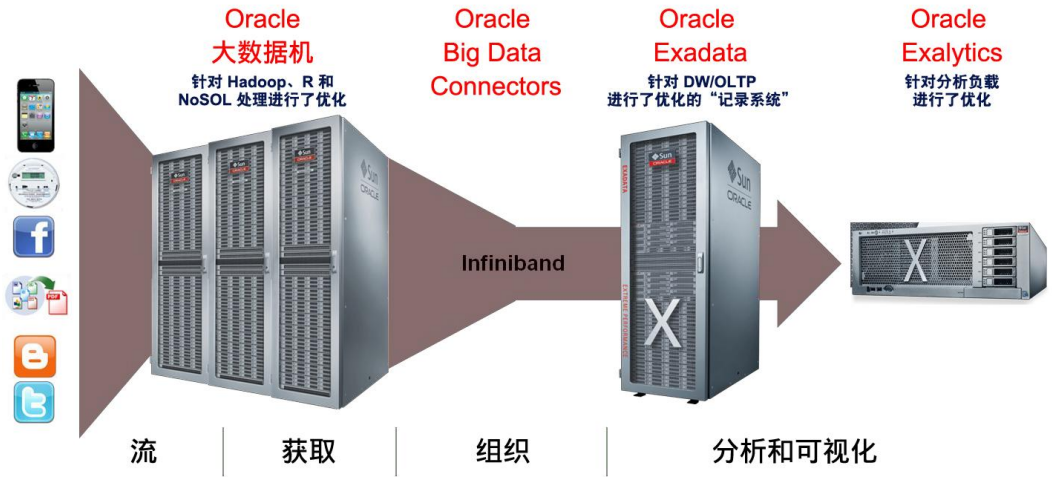

图 10：用例 3：联合分析的物理架构

借助 Oracle 集成式系统（包括 Oracle 大数据机、Oracle Exadata 和 Oracle Exalytics），企业可以降低实施风险、快速实现价值并获得卓越的性能和可伸缩性，从而能轻松应对复杂的业务和 IT 挑战。

该架构的主要组件包括：

- Oracle 大数据机（或其他 Hadoop 解决方案）：
  - 依托包含 Apache Hadoop 的 Cloudera 发行版（CDH）的完整版本，用于存储日志、评论和其他相关大数据。
    - NoSQL：专用于捕获数据结构灵活的低延迟数据，并支持快速查询。
    - MapReduce：通过处理大量数据来精简和优化数据集，以便将数据集载入到数据库管理系统中。
- Oracle Big Data Connectors：提供了一个 Hadoop 适配器，可通过方便易用的图形用户界面轻松将 Hadoop 与 Oracle 数据库集成在一起。
- Oracle Exadata：集成式数据库系统，支持混合负载并可为事务处理和/或数据仓库环境赋予卓越的性能，从而为进一步的联合分析奠定基础。
- Oracle Exalytics：集成式 BI 系统，可为最终用户提供快如闪念的分析功能。
- Infiniband：Oracle 大数据机、Oracle Exadata 和 Oracle Exalytics 之间，这将为批处理或查询负载提供高速数据传输。

大数据最佳实践

以下列出了打造成功大数据架构基础的一些通用指导准则：

1：确保大数据与特定业务目标相一致

大数据处理的主要宗旨在于通过对低密度的大容量数据进行智能筛选来发现隐藏的价值。作为一名架构师，我将就如何运用大数据技术来为您的企业提供相应的建议。举例来说，您应掌握如何通过过滤网志来理解电子商务行为、掌握如何从社交媒体和客户支持交互中获取舆情、并应掌握统计关联方法以及它们与客户、产品、制造或工程数据的相关性。尽管大数据是一个新的 IT 前沿领域，而且大家都在满怀激情地学习相关的新技术，但务必要保证新的投资以技能、组织或基础设施为本，从而在强大的业务驱动的环境下确保持续的项目投资和资金支持。如需确定自己是否处于正确的轨道上，请思考自己当前支持业务架构的方式以及首要 IT 事项。

2：通过标准和治理弥补技能匮乏

McKinsey Global Institute[1] 撰文指出大数据发展的最大障碍之一就是技能匮乏。随着对深度分析技术的采用不断升温，预计这一缺额在 2018 年将达到 60%。为了降低此风险，用户必须确保将大数据技术、考量因素和决策纳入到 IT 治理计划中。实现标准化的方法可让用户更加有效地管理成本和利用资源。另一项需要考虑的策略就是采用相应的软件设备，这将帮助企业在发展和培养内部专家的同时迅速上手、在更短的时间内实现价值。

3：通过卓越中心优化知识传输

使用卓越中心 (CoE) 可以共享解决方案知识、计划组件、监管和对项目的管理通信。无论大数据技术是新投资还是扩充性投资，软成本和硬成本都是可在整个企业内部共享的投资。CoE 方法的另一项优势在于可以通过一种更加结构化和系统性的方式来持续推进大数据和整个信息架构的成熟度。

4：首要目标是确保非结构化数据与结构化数据协调一致

单独对大数据加以分析是一项极具价值的任务。但是，通过将高密度的大数据与结构化数据相关联，用户可以获得更加深刻的洞察力。举例来说，所有客户的舆情与最佳客户的舆情是不同的。无论是捕获客户、产品、设备还是环境大数据，其目标都是在核心主数据和分析概要中添加相关性更高的数据点，从而改善用户从结果中获取的洞察。鉴于此，许多人都将大数据视为现有商务智能和数据仓库平台的不可或缺的扩展。

请谨记，大数据分析流程和模型既可以采用人工方式也可以由机器处理。大数据分析功能包括统计、空间数据、语义、交互式发现和可视化。这些功能可让企业的知识工作者和新的分析模型关联不同类型和不同来源的数据，建立关联并揭示有意义的结果。但总而言之，请将大数据技术视作相关事务数据的预处理器和后处理器，并充分利用在基础设施、平台、BI 和 DW 方面的前期投资。

[1] McKinsey Global Institute，2011 年 5 月，“大数据的挑战和机遇”，
https://www.mckinseyquarterly.com/The_challenge_and_opportunity_of_big_data_2806

5：通过计划沙盒来确保性能

企业并非总是能直观地洞察数据的含义。有时，我们甚至不知道应从何处入手。这完全在预料之中。管理层和 IT 部门需要为这种“方向感缺失”或“明确需求的缺失”提供支持。因此，为了适应交互式数据探索和统计算法实验，我们需要建立一些高性能工作区。但请确保“沙盒”环境具备所需功能并且得到正确的治理。

6：与云运营模式相协调

无论是进行迭代试验还是运行生产作业，大数据流程和用户都需要能够访问各种各样的资源。跨数据种类的数据（事务、主数据、参考、概要）是大数据解决方案不可或缺的一个要素。企业应按需创建分析沙盒，并且资源管理人员必须能够控制整个数据流各个阶段，包括预处理、集成、数据库中汇总、后处理和分析建模。合理规划私有云和公有云供应及安全策略是支持这些不断变化的需求的一个不可或缺的关键环节。

总结

大数据时代已经来临。分析人员和研究组织已经揭示，挖掘机器生成的数据对于企业未来的成功至关重要。采用新技术始终是一项极具挑战性的任务，但只要架构师合理规划，就可以建立一条快速、可靠的采用路径。

与其盲目地探索大数据功能中的“新特性”，我们建议企业考虑将大数据集成到现有基础架构和 BI 投资中。举例来说，企业应确保新的运营和管理功能与标准 IT 相一致，构建企业级伸缩性和弹性，通过开源项目统一数据库和开发范例，以及通过共享元数据来进行集成和分析。

最后，但同样重要的一点是，请扩大当前的 IT 治理范围，在其中涵盖大数据卓越中心，从而确保业务一致性、发展内部技能、管理开源工具和技术、共享知识、建立标准并管理最佳实践。

有关 Oracle 和大数据的更多信息，请访问 _____

您也可以在网播中收听 Helen Sun 关于本白皮书中话题的讨论。请选择第 6 课“征服大数据”。