cloudera®

intel®

金融服务行业的企业数据中心

三个客户案例研究

常见的金融服务采用 Apache Hadoop 的周期通常始于运营效率最高、成本节省最多的两个用例之一：借助集中式数据中心实现的数据整合和多租户，或全保真（full-fidelity）分析和法规遵从。但沙丘集团（Sand Hill Group）在 2013 年 10 月开展的一项研究中发现，尽管 62% 的受访者预计未来 12 到 18 个月高级分析将成为主要用例，但只有 11% 的受访者表示已经完成了他们的第一个 Hadoop 项目，仅 9% 的受访者表示正在使用 Hadoop 进行高级分析。[1] 许多企业希望找到可靠、实时、经济的大数据解决方案，但在全面普及和推广过程中存在哪些障碍？

传统数据管理和分析平台通常作为专用系统进行部署，以实现特定目标，相比之下，企业数据中心具有集中、开放和可扩展等特性，更类似于金融服务行业的解决方案引擎。通过将商机成本降至最低，并加强与庞大且不断增长的生态系统合作以提供相关的技术和熟悉的应用，Cloudera 帮助企业应对当前的大数据挑战，并最大限度地提高其数据基础设施的价值，以实现接下来的高级业务目标。通过将计算与数据相融合，帮助企业开始尝试一些简单的入门级应用，让他们有更多的机会充分利用新的信息驱动型业务能力，包括支持高效、自动检测和预防欺诈的机器学习模型，支持打造个性化客户体验以增加向上销售和交叉销售机会的推荐引擎，以及支持专门探索、实验分析和高级风险建模的 360 度全方位业务了解等。而在过去，这种业务能力对于大部分企业来说过于昂贵和复杂。

领先的支付处理公司和欺诈检测

随着金融交易业务从现场到在线流程的转变，全球领先的信用卡公司处理的日常交易量显著增加，导致欺诈率也随之上升。欺诈是一种会导致重大财务或其他损失的意外或罕见事件 — 从企业的角度来看，有效的应对措施可分为检测、预防和降低等不同的类别。金融服务行业经常发生意外事件，因为欺诈者能够事先了解当前系统的运行方式，包括以往的欺诈案例和欺诈检测机制，因此很难创建可靠的检测统计模型。

如果使用大型信用卡处理器，虽然每年用于数据仓库的预算高达 10 亿美元，但统计人员只能简单查询规模相对较小的数据样本，因为扩大查询范围会消耗过多的计算资源。尤其是全球信息安全组的数据科学家，他们希望提高查询响应速度并无限扩大范围，以更好地挖掘和分析关系数据库管理系统（RDBMS）中的数据。

通过在 Cloudera Enterprise 中部署 Hadoop，该公司不仅能够集成通常只有

[1] Graham、Bradley 和 Rangaswami, M.R.，您使用 Hadoop 吗？大数据从业者调查。沙丘集团，2013 年 10 月

© 2015 年 Cloudera，Inc. 版权所有。

独立存储区域网络（SAN）、提取-转换-加载（ETL）网格和数据仓库系统才能完成的工作，从而大幅简化数据处理流程并降低预期成本，还能够立即开始检测时间较长、数据来源各不相同的数据，识别更多不同的潜在意外事件。为了克服延迟，Apache Flume（Hadoop 的服务，包括高效收集、整合和移动大量日志数据）可在几秒钟内将数十亿个事件加载至 HDFS（Hadoop 的分布式文件系统和主存储层），并使用 Cloudera Impala（Hadoop 的大规模并行处理结构化查询语言（SQL）引擎）对这些事件进行分析，甚至还能使用 Apache Spark 的内存功能在流数据上运行模型。Apache Spark 是下一代开源处理引擎，能够在 HDFS 中对所有数据执行批处理、流传输和交互式分析等功能。

如今，信用卡处理器平均每天可将 4TB 数据采集至 Hadoop 集群中，而且能够在数百个小型节点上维护数千 TB 的数据，以进行欺诈建模。部署企业数据中心后不久，一家合作伙伴通知该公司发现了一起小规模欺诈事件，据报告该欺诈事件持续了两周的时间就被检测到。在响应过程中，全球信息安全小组针对 Hadoop 中的长期详细数据运行了专门的描述性分析模型 — 仅靠传统数据基础设施根本不可能完成这一任务。通过搜索更广泛的数据集，该公司发现了该欺诈活动几个月以来的模式。这是该部门有史以来检测到的最大欺诈事件，帮助节省了至少 3,000 万美元。

此外，该公司还使用 Hadoop 集群的数据，为商家创建有助于推动收入增长的报告。过去，有些月度报告需要两天才能完成，而且要求技术团队管理大量的处理资源。现在，信用卡处理器可以将数量更庞大的交易数据和银行的购买数据相结合，进而生成可以出售的报告，仅此一项就可为公司带来 10 亿美元的收入。这些报告可在数小时内运行，并解决了商家在收集客户细分和交叉销售分析数据时所面临的延迟问题。

顶级投资银行和 360 度全方位了解业务

随着用于投资组合分析的数据量和数据种类的增多，许多投资银行正设法找到处理更多数据的最佳方式，并从这些数据中获得出色的洞察和价值。大多数投资银行依赖于数据采样的方法，不仅降低了模型的准确性，还在一定程度上限制了数据探查。

360 度全方位视野的概念通常与零售银行联系在一起，此类银行希望使用来自多个业务部门的更多数据，并与线上和线下购买行为趋势相结合，以便了解如何有效、高效地与客户进行互动，进而提高客户忠诚度并带来更多新的销售机会。但是，广泛、充分、实时地了解业务与客户满意度和营销指标并无必然的联系。将相关或完全不同的数据集结合在一起，有助于发现其中的模式、关联或因果关系，如果能够将这些转化为机会或风险，便可以在帮助投资银行领先其他公司上跨出重要的一步。 ▶

在一家领先的批发银行，它所具备的竞争优势不仅与数据的数量和质量有着直接的关系，更重要的是，它可以灵活地研究各种洞察之间的相互关系，并将其转化为业务成效。根据报告，该公司在2013 年管理的客户资产达到了数万亿美元，这不仅与其自身的市场和投资数据相吻合，还可使用自定义算法从公开信息及政策信息、宏观经济数据、客户资料及交易记录，甚至网络数据中提取到切实可行的洞察 — 基本上单击任何一条数据都能获得有用的信息。

投资银行的数据科学家希望使用大量数据来进行投资组合分析，但他们内部的传统数据库和网格计算技术无法实现扩展。过去，IT 部门会创建自定义数据结构，查找数据，并与数据表保持一致，然后支持分析人员编写 SQL 查询。这个过程极其精确，但非常耗时。通常情况下，应用移交给业务部门后，如果分析人员指出该项目达不到最初的要求，那么该应用将不会得到使用。

作为首例基于 Cloudera 的大数据概念验证，银行 IT 部门组装了 15 台使用寿命已结束的服务器，安装了 CDH（Cloudera 的开源发行版 Apache Hadoop），并加载了所有公司日志，包括各种根据时间关联性设置的网络和数据库日志。凭借大量的在线数据以及 Hadoop 中的数据，该银行首次能够从各个角度了解其规模达到 PB 级的投资业务。由于 Hadoop 以无模式结构存储所有数据，因此 IT 部门能够根据业务部门的需求，灵活地通过他们想要的任意输入组合来划分记录或输出，并准确交付最终结果。

作为 Cloudera Enterprise 的客户，该投资银行不再依赖于采样，而是以更大的规模运行投资组合分析，从而实现更好的结果。Hadoop 可搜索庞大的数据量，并对每一个可能的属性运行模式匹配。用户不必知道自己在查找什么 — 让软件和模型检测相应的模式，然后跟进后续调查即可。

企业数据中心可根据日志数据获得时间关联性，其完整性和清晰度都达到了空前的水平，有利于银行了解市场活动，以及它们与 Web 问题和数据库读写问题之间的关联。例如，公司可实时追溯活动的整个过程，包括活动参与人员、活动内容、时间、方式、问题起因，以及交易数据类型。银行可以将前台的活动与后台办公室的活动，以及导致出现意外结果的数据联系起来。过去，企业需要花费几个月的时间才能查出系统异常的原因，而且这期间可能会导致出现重大损失。

现在借助 Cloudera，公司可以找出并解决所发生的问题，甚至还能预先阻止问题的发生。此外，作为企业数据中心一部分而部署的高级分析工具还可为银行的财务顾问提供定制化建议，帮助他们根据实时收集的当前位置和市场条件信息，建议客户出售或购买股票 — 实质上是将 Cloudera Enterprise 数据中心版提供和支持的 Hadoop 功能转化为收入。

大型保险公司与金融产品个性化

随着传感器、移动设备、纳米技术和社交应用数量的激增，人们比以往任何时候都更加倾向于监控，以及被动或主动地分享自己日常的行为数据。过去，保险公司一直打价格战，或通过广泛、昂贵的市场营销活动开展竞争。但现在，他们希望收集人们的生活方式、健康模式、习惯和喜好等相关信息，以此为依据定制保险计划，进而为客户提供差异化的产品和服务。但是，传统数据库在规模和速度上的扩展能力有限，无法提供制定个性化保险产品所需的实时、多结构化数据。企业数据中心支持进行实时存储和流传输数据，有助于打造一种极具竞争力的按需付费保险模式。

美国一家最大的个人保险公司成立于 20 世纪 30 年代初，最初只是国家连锁百货公司的一部分。在 80 多年的经营过程中，该公司收集了大量数据，其中许多数据从未实现数字化，而且其中大部分为非结构化文档内容。这家保险公司开始将过去和当前的策略数据迁移至线上保存，并尝试运行与交通模式、社会经济研究、气象信息等外部数据相关的程序，IT 部门发现，原有系统扩展能力不足，不能满足公司对不同的数据格式和数据源的需求。

业务分析师也面临许多挑战，图表链接分析就是其中之一。例如，他们可一次查看某一个州的数据（每个州的分析需耗时一天），但不能一次同时分析多个州（总共 50 个州）。尽管新部署的数据系统可采集和准备用于报告和商业智能的数据，但这些新系统主要是对之前的数据管理方法做了略微改进，而且将各种数据和工作负载分隔成了不同的数据孤岛。

为了提高分析的可扩展性，首先必须加快处理速度，并整合不同的数据集，为此，这家领先的保险公司使用 Cloudera Enterprise 构建了企业数据中心。集中式 Hadoop 实施涵盖了整个公司的每一个系统，有 助于消除数据孤岛，并

提供整体数据视图。从技术上来说，Hadoop 主要有三种应用情形，分别为灵活、主动的数据存储；集成、高效的 ETL，以及应用统计与计算。

这家保险公司将客户账户信息、公共经济和社会研究以及遥测传感器数据集中在其初始 Hadoop 集群中。其中一些数据源之前从未进行过汇总，而且在登录 Hadoop 之前，许多刚刚实现数字化的历史数据无法与外部数据源一同分析。如今，该公司的企业数据中心与现有的大型机和数据仓库相集成，经过精心设计后可有效补充（而非取代）现有的基础设施。

现在，这家保险公司可使用 Apache Hive 运行描述性模型分析所有 50 个州的历史数据，分析速度平均提高了 7500%，并借助 Impala 显著增强了业务成效。Apache Hive 是一款开源软件，支持在 Hadoop 中以可扩展的方式转换和分析复杂的半结构化数据。消除了数据孤岛之后，公司的分析师和数据科学家正在建立预测模型，以帮助企业为客户定制更符合个人行为并有助于降低风险的产品；更精确地调整保险计划的定价，以最大限度地提高其生命周期内的价值；开发独具特色的差异化营销活动，以宣传产品的价值，创造最合适的交叉销售和向上销售机会，同时不降低利润。

大数据和企业数据中心

当信息从数据孤岛中释放出来、加以保护，并提供给数据分析师、工程师和科学家来回答有关市场的关键问题（他们使用熟悉的工具根据需求访问原始数据），每位高层管理人员都能全面地了解业务，这在整个业界可能尚属首次。对金融服务公司来说，要想获得先进的大数据流程，就必须符合多租户方面的要求并确保系统安全性，包括机器学习、推荐引擎、安全信息和事件管理、图表分析和其他功能，从而将数据转化为收入，并且无需购买专用的工具，可实现显著的成本节约。

当前，在信息架构的核心位置引进基于 Apache Hadoop 构建的企业数据中心，有助于集中所有格式的数据供业务用户使用，还可实现全保真度和安全性。相比于传统数据管理技术，每 TB 数据的资本支出可降低高达 99%。

企业数据中心可作为灵活的存储库，获取企业所有价值未知的数据来满足不同的需求，包括合规、改进核心业务流程（比如客户细分和投资建模），以及运行更加复杂的应用（比如实时异常检测）。它可加快商业智能报告和分析的速度，以便在关键服务级别协议方面显著提高吞吐量。此外，它还可增强活动数据的可用性和可访问性，支持实现业务增长并全面展示金融服务公司的整体运营，以实现流程创新 — 所有这一切都能与现有基础设施和应用完全集成，进而大幅扩展（而非取代）过去投资的价值。

然而，信息驱动型企业的最大价值取决于与业务相关的问题，但是金融服务公司历来都不能或不敢提出这些问题，因为数据缺乏一致性，或专用工具的成本令人望而却步。企业数据中心鼓励更多的探索和发现，着眼于帮助决策者制定前瞻性的行业发展计划:

在不构建专用系统或不局限于小规模样本的情况下，我们如何使用几十年积累的客户数据来检测欺诈行为？

跨不同业务领域 360 度全方位了解客户可为我们提供哪些有关下游机会和风险的信息？

我们能否从一个中心点保存所有当前客户及潜在客户的海量数据，以符合法规要求、保护客户隐私，并提供给不同业务用户使用？

关于 Cloudera

Cloudera 提供业界首个大数据统一平台（基于 Apache Hadoop 构建的企业数据中心），推动整个企业数据管理的变革。Cloudera 为企业提供一个统一的地方来存储、访问、处理、保护和分析所有企业数据，帮助企业扩大其现有投资的价值，同时支持以全新的方法从企业数据中获得价值。Cloudera 的开源大数据平台是全球范围内采用最广泛的平台，而且 Hadoop 是对开源 Hadoop 生态系统贡献最大的供应商。作为领先的 Hadoop 专家培训机构，Cloudera 在全球累计培训了超过 40,000 名学员。超过 1,600 个 Cloudera 合作伙伴和 Cloudera 资深专业服务团队帮助客户更快地获取价值。最后，唯有 Cloudera 能够提供前瞻性、预测性的支持，以确保企业数据中心无忧地运行。各行各业的许多领先企业和顶尖公共组织都正在全球范围内采用 Cloudera 作为实际生产平台。

CLOUDERA
www.cloudera.com