Join the discussion @ p2p.wrox.com

Wrox Programmer to Programmer™

大数据应用与技术丛书

Professional Hadoop

Hadoop 大数据解决方案

Benoy Antony　Konstantin Boudnik
[美] Cheryl Adams　Branky Shao　著
Cazen Lee　Kai Sasaki
殷聪贤　杨朋朋　译

清华大学出版社

大数据应用与技术丛书

Hadoop 大数据解决方案

[美] Benoy Antony
Konstantin Boudnik
Cheryl Adams
Branky Shao
Cazen Lee
Kai Sasaki
著

殷聪贤　杨朋朋　译

清华大学出版社

北　京

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, Kai Sasaki
Professional Hadoop
EISBN：978-1-119-26717-1

图书在版编目(CIP)数据

Hadoop 大数据解决方案 / (美) 贝诺・安东尼(Benoy Anthony) 等著；殷聪贤，杨朋朋 译. —北京：清华大学出版社，2017
(大数据应用与技术丛书)
书名原文：Professional Hadoop
ISBN 978-7-302-46645-1

Ⅰ. ①H… Ⅱ. ①贝… ②殷… ③杨… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 032043 号

责任编辑：王　军　于　平
装帧设计：牛静敏
责任校对：牛艳敏
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京泽宇印刷有限公司
经　　销：全国新华书店
开　　本：148mm×210mm　　印　　张：9　　字　　数：242 千字
版　　次：2017 年 2 月第 1 版　　印　　次：2017 年 2 月第 1 次印刷
印　　数：1~3500
定　　价：49.80 元

产品编号：070786-01

第 1 章

Hadoop 概述

本章内容提要

- Hadoop 的组件
- HDFS、MapReduce、YARN、ZooKeeper 和 Hive 的角色
- Hadoop 与其他系统的集成
- 数据集成与 Hadoop

Hadoop 是一种用于管理大数据的基本工具。这种工具满足了企业在大型数据库(在 Hadoop 中亦称为数据湖)管理方面日益增长的需求。当涉及数据时，企业中最大的需求便是可扩展能力。科技和商业促使各种组织收集越来越多的数据，而这也增加了高效管理这些数据的需求。本章探讨 Hadoop Stack，以及所有可与 Hadoop 一起使用的相关组件。

在构建 Hadoop Stack 的过程中，每个组件都在平台中扮演着重要角色。软件栈始于 Hadoop Common 中所包含的基础组件。Hadoop

Common 是常见工具和库的集合，用于支持其他 Hadoop 模块。和其他软件栈一样，这些支持文件是一款成功实现的必要条件。而众所周知的文件系统，Hadoop 分布式文件系统，或者说 HDFS，则是 Hadoop 的核心，然而它并不会威胁到你的预算。如果要分析一组数据，你可以使用 MapReduce 中包含的编程逻辑，它提供了在 Hadoop 群集上横跨多台服务器的可扩展性。为实现资源管理，可考虑将 Hadoop YARN 加入到软件栈中，它是面向大数据应用程序的分布式操作系统。

ZooKeeper 是另一个 Hadoop Stack 组件，它能通过共享层次名称空间的数据寄存器(称为 znode)，使得分布式进程相互协调工作。每个 znode 都由一个路径来标识，路径元素由斜杠(/)分隔。

还有其他一些系统能与 Hadoop 进行集成并从其基础架构中受益。虽然 Hadoop 并不被认为是一种关系型数据库管理系统(RDBMS)，但其仍能与 Oracle、MySQL 和 SQL Server 等系统一起工作。这些系统都已经开发了用于对接 Hadoop 框架的连接组件。我们将在本章介绍这些组件中的一部分，并且展示它们如何与 Hadoop 进行交互。

1.1 商业分析与大数据

商业分析通过统计和业务分析对数据进行研究。Hadoop 允许你在其数据存储中进行业务分析。这些结果使得组织和公司能够做出有利于自身的更好商业决策。

为加深理解，让我们勾勒一下大数据的概况。鉴于所涉及数据的规模，它们会分布于大量存储和计算节点上，而这得益于使用 Hadoop。由于 Hadoop 是分布式的(而非集中式的)，因而不具备关系型数据库管理系统(RDBMS)的特点。这使得你能够使用 Hadoop 所提供的大型数据存储和多种数据类型。

例如，让我们考虑类似 Google、Bing 或者 Twitter 这样的大型数据存储。所有这些数据存储都会随着诸如查询和庞大用户基数等活动事件而呈现出指数增长。Hadoop 的组件可以帮助你处理这些大型数据存储。

类似 Google 这样的商业公司可使用 Hadoop 来操作、管理其数据存储并从中产生出有意义的结果。通常用于商业分析的传统工具并不旨在处理或分析超大规模数据集，但 Hadoop 是一个适用于这些商业模型的解决方案。

1.1.1　Hadoop 的组件

Hadoop Common 是 Hadoop 的基础，因为它包含主要服务和基本进程，例如对底层操作系统及其文件系统的抽象。Hadoop Common 还包含必要的 Java 归档(Java Archive，JAR)文件和用于启动 Hadoop 的脚本。Hadoop Common 包甚至提供了源代码和文档，以及贡献者的相关内容。如果没有 Hadoop Common，你无法运行 Hadoop。

与任何软件栈一样，Apache 对于配置 Hadoop Common 有一定要求。大体了解 Linux 或 Unix 管理员所需的技能将有助于你完成配置。Hadoop Common 也称为 Hadoop Stack，并不是为初学者设计的，因此实现的速度取决于你的经验。事实上，Apache 在其网站上明确指出，如果你还在努力学习如何管理 Linux 环境的话，那么 Hadoop 并不是你能够应付的任务。建议在尝试安装 Hadoop 之前，你需要先熟悉此类环境。

1.1.2　Hadoop 分布式文件系统(HDFS)

在 Hadoop Common 安装完成后，是时候该研究 Hadoop Stack 的其余组件了。HDFS(Hadoop Distributed File System)提供一个分布式文件系统，设计目标是能够运行在基础硬件组件之上。大多数企业被其最小化的系统配置要求所吸引。此环境可以在虚拟机(Virtual

Machine，VM)或笔记本电脑上完成初始配置，而且可以升级到服务器部署。它具有高度的容错性，并且被设计为能够部署在低成本的硬件之上。它提供对应用程序数据的高吞吐量访问，适合于面向大型数据集的应用程序。

在任何环境中，硬件故障都是不可避免的。有了 HDFS，你的数据可以跨越数千台服务器，而每台服务器上均包含一部分基础数据。这就是容错功能发挥作用的地方。现实情况是，这么多服务器总会遇到一台或者多台无法正常工作的风险。HDFS 具备检测故障和快速执行自动恢复的功能。

HDFS 的设计针对批处理做了优化，它提供高吞吐量的数据访问，而非低延迟的数据访问。运行在 HDFS 上的应用程序有着大型数据集。在 HDFS 中一个典型的文件大小可以达到数百 GB 或更大，所以 HDFS 显然支持大文件。它提供高效集成数据带宽，并且单个群集可以扩展至数百节点。

Hadoop 是一个单一功能的分布式系统，为了并行读取数据集并提供更高的吞吐量，它与群集中的机器进行直接交互。可将 Hadoop 想象为一个动力车间，它让单个 CPU 运行在群集中大量低成本的机器上。既然已经介绍了用于读取数据的工具，下一步便是用 MapReduce 来处理它。

1.1.3 MapReduce 是什么

MapReduce 是 Hadoop 的一个编程组件，用于处理和读取大型数据集。MapReduce 算法赋予了 Hadoop 并行化处理数据的能力。简而言之，MapReduce 用于将大量数据浓缩为有意义的统计分析结果。MapReduce 可以执行批处理作业，即能在处理过程中多次读取大量数据来产生所需的结果。

对于拥有大型数据存储或者数据湖的企业和组织来说，这是一种重要的组件，它将数据限定到可控的大小范围内，以便用于分析

或查询。

如图 1-1 所示，MapReduce 的工作流程就像一个有着大量齿轮的古老时钟。在移动到下一个之前，每一个齿轮执行一项特定任务。它展现了数据被切分为更小尺寸以供处理的过渡状态。

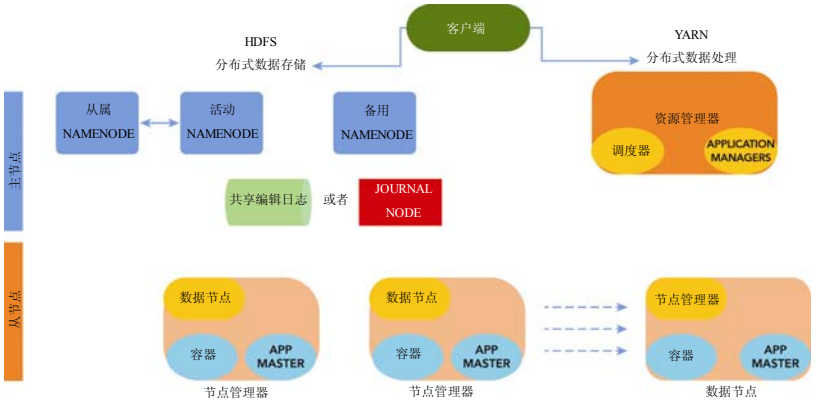


图 1-1

MapReduce 的功能使得它成为最常用的批处理工具之一。该处理器的灵活性使其能利用自身的影响力来挑战现有系统。通过将数据处理的工作负载分为多个并行执行的任务，MapReduce 允许其用户处理存储于 HDFS 上不限数量的任意类型的数据。因此，MapReduce 让 Hadoop 成为了一款强大工具。

在 Hadoop 最近的发展中，另有一款称为 YARN 的组件已经可用于进一步管理 Hadoop 生态系统。

1.1.4　YARN 是什么

YARN 基础设施(另一个资源协调器)是一项用于提供执行应用程序所需的计算资源(内存、CPU 等)的框架。

YARN 有什么诱人的特点或是性质？其中两个重要的部分是资源管理器和节点管理器。让我们来勾勒 YARN 的框架。首先考虑一个两层的群集，其中资源管理器在顶层(每个群集中只有一个)。资

源管理器是主节点。它了解从节点所在的位置(较底层)以及它们拥有多少资源。它运行了多种服务，其中最重要的是用于决定如何分配资源的资源调度器。节点管理器(每个群集中有多个)是此基础设施的从节点。当开始运行时，它向资源管理器声明自己。此类节点有能力向群集提供资源，它的资源容量即内存和其他资源的数量。在运行时，资源调度器将决定如何使用该容量。Hadoop 2 中的 YARN 框架允许工作负载在各种处理框架之间动态共享群集资源，这些框架包括 MapReduce、Impala 和 Spark。YARN 目前用于处理内存和 CPU，并将在未来用于协调其他资源，例如磁盘和网络 I/O。

1.2 ZooKeeper 是什么

ZooKeeper 是另一项 Hadoop 服务——分布式系统环境下的信息保管员。ZooKeeper 的集中管理解决方案用于维护分布式系统的配置。由于 ZooKeeper 用于维护信息，因此任何新节点一旦加入系统，将从 ZooKeeper 中获取最新的集中式配置。这也使得你只需要通过 ZooKeeper 的一个客户端改变集中式配置，便能改变分布式系统的状态。

名称服务是将某个名称映射为与该名称相关信息的服务。它类似于活动目录，作为一项名称服务，活动目录的作用是将某人的用户 ID(用户名)映射为环境中的特定访问或权限。同样，DNS 服务作为名称服务，将域名映射为 IP 地址。通过在分布式系统中使用 ZooKeeper，你能记录哪些服务器或服务正处于运行状态，并且能够通过名称查看它们的状态。

如果有节点出现问题导致宕机，ZooKeeper 会采用一种通过选举 leader 来完成自动故障切换的策略，这是它自身已经支持的解决方案(见图 1-2)。选举 leader 是一项服务，可安装在多台机器上作为冗余备用，但在任何时刻只有一台处于活跃状态。如果这个活跃的

服务因为某些原因发生了故障，另一个服务则会起来继续它的工作。

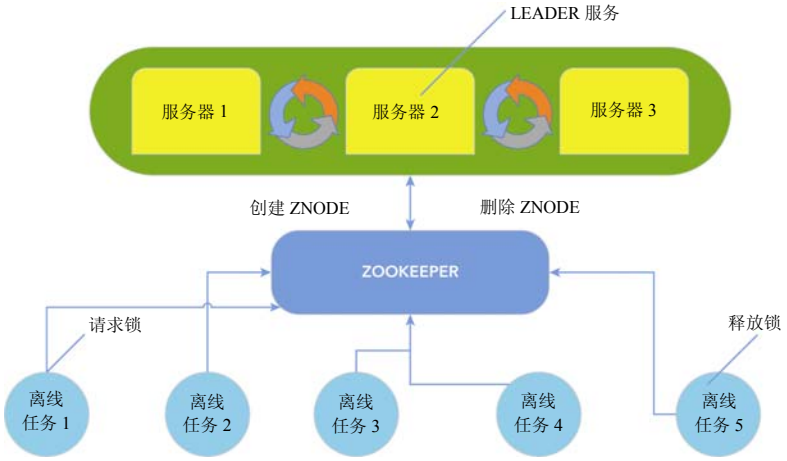


图 1-2

ZooKeeper 允许你处理更多的数据，并且更加可靠省时。ZooKeeper 能够帮助你建立更可靠的系统。托管的数据库群集能从集中式管理的服务中受益，这些服务包括名称服务、组服务、*leader* 选举、配置管理以及其他。所有这些协调服务都可以通过 ZooKeeper 进行管理。

1.3　Hive 是什么

Hive 在设计之初是 Hadoop 的一部分，但现在它是一个独立的组件。之所以在这里简单提及，是因为有些用户发现在标准的 Hadoop Stack 之外，它还是很有用处。

我们可以这样简单总结 Hive：它是建立在 Hadoop 顶层之上的数据仓库基础设施，用于提供对数据的汇总、查询以及分析。如果你在使用 Hadoop 工作时期望数据库的体验并且怀念关系型环境中的结构(见图 1-3)，那么它或许是你的解决方案。记住，这不是与传统的数据库或数据结构进行对比。它也不能取代现有的 RDBMS 环

境。Hive 提供了一种为数据赋予结构的渠道，并且通过一种名为 HiveQL 的类 SQL 语言进行数据查询。

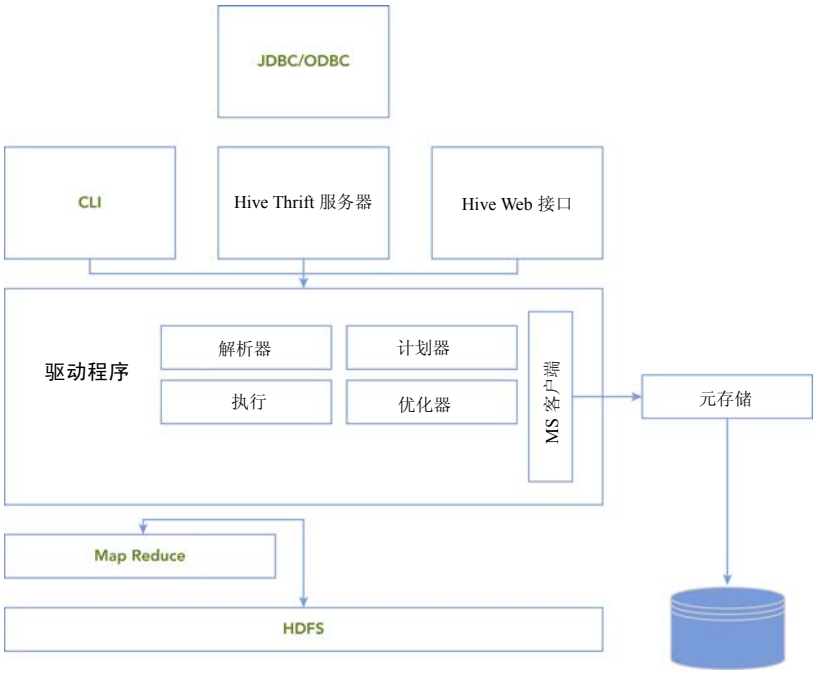


图 1-3

1.4 与其他系统集成

如果在科技领域工作，你一定清楚地知道集成是任何成功实现中必不可少的部分。一般来说，通过一些发现流程或计划会议，组织可以更高效地确定管理大数据的需求。后续步骤包括做出关于如何将 Hadoop 落实到现有环境的决定。

正在实现或考虑 Hadoop 的组织有可能将其引入到现有环境中。为获取最大的利益，了解如何能让 Hadoop 和现有环境一起工作以及该如何利用现有环境是非常重要的。

为说明这一点，考虑一种著名的积木玩具，它允许你通过相互连接创建新的玩具积木。仅通过将积木块简单连接在一起，你便可以创造出无限可能。关键原因在于每块积木上的连接点。类似于积木玩具，厂商开发了连接器以允许其他企业的系统连接到 Hadoop。通过使用连接器，你能够引入 Hadoop 来利用现有环境。

让我们介绍一些已经开发完成、用于将 Hadoop 与其他系统集成的组件。你应该思考在自己的环境中使用这些连接器所能够带来的优势。显然当集成时，你必须根据现有的系统环境，成为自己的 SME(Subject Matter Expert，领域专家)。

这些 Hadoop 的连接器将有可能适用于环境中系统的最新版本。如果想与 Hadoop 一起使用的系统不是应用程序或数据库引擎的最新版本，那么你需要将升级的因素考虑在内，以便使用增强版完整功能。我们建议全面检查你的系统需求，以避免沮丧和失望。Hadoop 生态系统会将所有新技术带入到你的系统中。

1.4.1　Hadoop 生态系统

Apache 将他们的集成称作生态系统。字典中将生态系统定义为：生物与它们所处环境的非生物组成部分(如空气、水、土壤和矿产)作为一个系统进行交互的共同体。基于技术的生态系统也有类似的属性。它是产品平台的结合，由平台拥有者所开发的核心组件所定义，辅之以自动化(机器脱离人类自主运转)企业在其周边(围绕着一个空间)所开发的应用程序。

以 Apache 的多种可用产品和大量供应商提供的将 Hadoop 与企业工具相集成的解决方案为基础，Hadoop 的开放源码和企业生态系统还在不断成长。HDFS 是该生态系统的主要组成部分。由于 Hadoop 有着低廉的商业成本，因此很容易去探索 Hadoop 的特性，无论是通过虚拟机，还是在现有环境建立混合生态系统。使用 Hadoop 解决方案来审查当前的数据方法以及日渐增长的供应商阵营是一种非

常好的方法。借助这些服务和工具，Hadoop 生态系统将继续发展，并清除分析处理和管理大数据湖中的一些障碍。通过使用本章中讨论的一些工具和服务，Hadoop 即可集成到数据生态系统的层次结构中。

Horton 数据平台(Horton Data Platform，HDP)是一个生态系统。HDP 能够帮助你通过使用虚拟机上的单节点群集来开始 Hadoop 之旅，如图 1-4 所示。由于 Hadoop 是一个商用(几乎没有额外成本)的解决方案，因此 HDP 使得你能够将其部署到云端或者自己的数据中心。

HDP 为你提供数据平台基础以供搭建自己的 Hadoop 基础设施，这包括一长串商业智能(BI)及其他相关供应商的列表。平台的设计目标是支持处理多种来源及格式的数据，并且允许设计自定义解决方案。资源列表过大，以至于无法在这里展示，强烈推荐直接从供应商处获取此信息。选择像 HDP 这样产品的美妙之处在于他们是 Hadoop 的主要贡献者之一。这便开启了在多种数据库资源上使用 Hadoop 的大门。

*请向供应商确认。资源可能会有所不同。

图 1-4

HDP 被视为一个生态系统，因为它创造了一个数据社区，将

Hadoop 和其他工具汇集在一起。

Cloudera(CDH)为其数据平台创建了一个类似的生态系统。Cloudera 为集成结构化和非结构化的数据创造了条件。通过使用平台交付的统一服务，Cloudera 开启了处理和分析多种不同数据类型的大门(见图 1-5)。

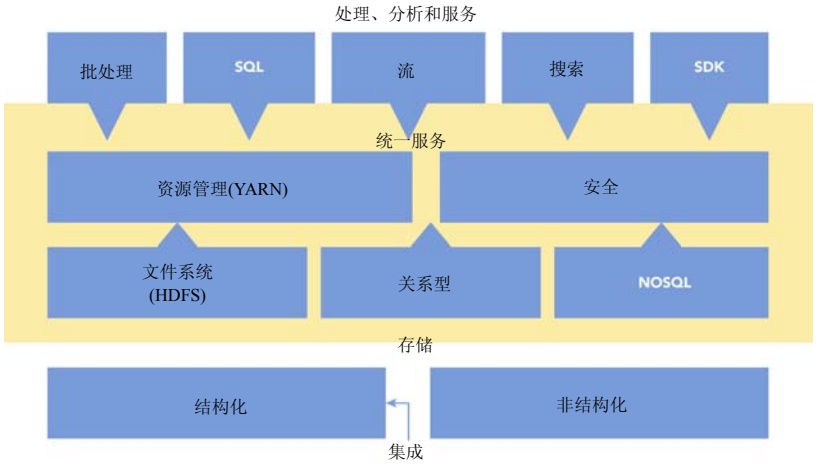


图 1-5

1.4.2　数据集成与 Hadoop

数据集成是 Hadoop 解决方案架构的关键步骤。许多供应商利用开源的集成工具在无须编写代码的情况下即可轻松地将 Apache Hadoop 连接到数百种数据系统。如果你的职业不是程序员或开发人员，那么这点对你来说无疑是使用 Hadoop 的加分项。大多数供应商使用各种开放源码解决方案用于数据集成，这些解决方案原生支持 Apache Hadoop，包括为 HDFS、HBase、Pig、Sqoop 和 Hive 提供连接器(见图 1-6)。

基于 Hadoop 的应用程序具有良好的平衡性，能够支持 Windows 平台并与微软的 BI 工具(例如 Excel、Power View 和 PowerPivot)良

好地集成，创造出轻松分析这些大规模商业信息的独特方式。

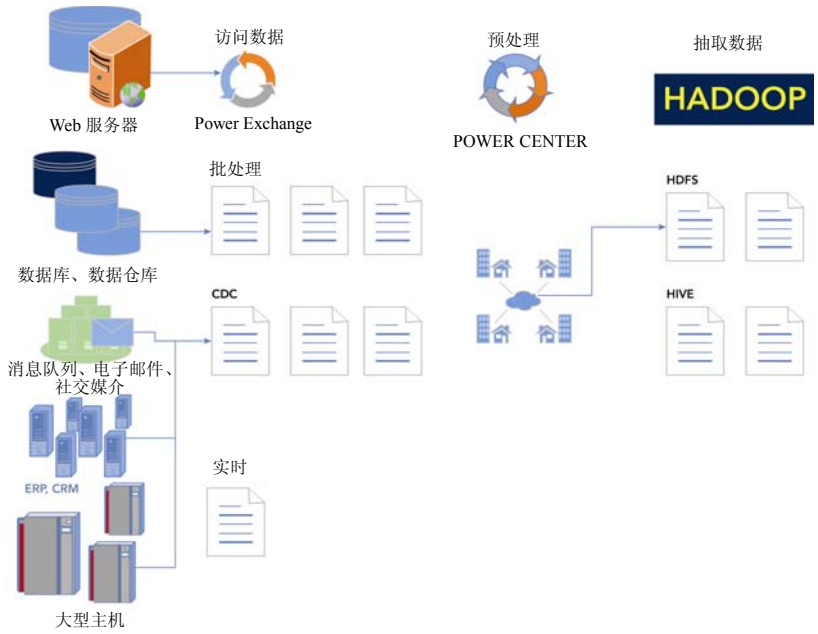


图 1-6

这并不意味着 Hadoop 或者其他数据平台的解决方案无法在非 Windows 环境下运行。你应该细心检查现有的或者计划使用的环境以决定最优解决方案。数据平台或者数据管理平台正如其名。它是一个集中式计算系统，用于收集、集成和管理大型结构化和非结构化数据集。

从理论上讲，无论 HortonWorks，还是 Cloudera，均是可供选择的平台，包括用于与现有数据环境和 Hadoop 一起工作的 RDBMS 连接器。大多数供应商均有关于系统需求的详细信息。一般来说，大量工具都会提到 Windows 操作系统或者基于 Windows 的组件，这是因为基于 Windows 的 BI 工具得到了广泛使用。微软的 SQL Server 是用于数据库服务的首要 Windows 工具。使用该商业工具的

组织将不再受大数据的约束。微软有能力通过提供灵活性以及增强 Hadoop、Windows Server 和 Windows Azure 的连通性来更好地操作和集成 Hadoop。Informatica 软件，使用 Power Exchange 连接器协同 Hortonworks，优化了 Hadoop 上的整条大数据供应链，将数据转换为具有可操作性的信息来驱动商业价值。

例如，现代的数据架构正在越来越多地用于建造大型数据湖。通过将数据管理服务集成为更大的数据湖，企业可以利用各种各样的渠道来存储和处理大量数据，这些渠道包括社交媒体、点击流数据、服务器日志、客户交易与交互、视频以及来自现场设备的传感器数据。

Hortonworks 或者 Cloudera 数据平台，以及 Informatica，使得企业能够优化 ETL(抽取、转换、加载)工作流程，以便在 Hadoop 中长期存储和处理大规模数据。

Hadoop 与企业工具的集成使得组织能够将内部和外部的所有数据用于获得完整的分析能力，并以此推动现代数据驱动业务的成功。

另一个例子，Hadoop Applier 提供了 MySQL 和 Hadoop 分布式文件系统之间的实时连接，可以用于大数据分析——例如情绪分析、营销活动分析、客户流失建模、欺诈检测、风险建模以及其他多种分析。许多得到广泛使用的系统，例如 Apache Hive，也将 HDFS 用于数据存储(见图 1-7)。

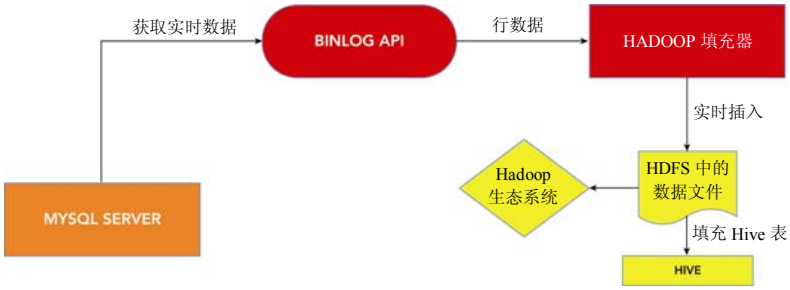


图 1-7

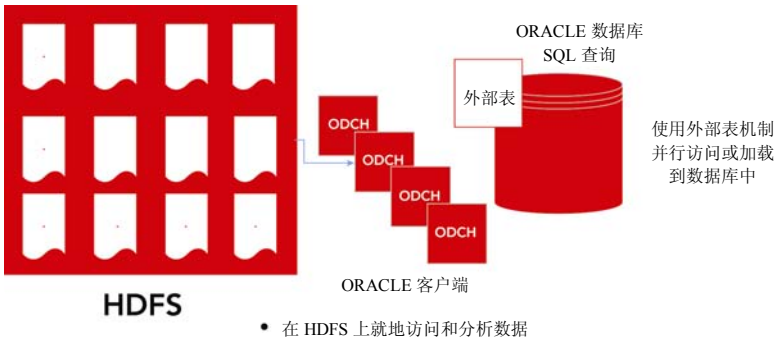
Oracle 公司为其旗舰数据库引擎和 Hadoop 开发了一款软件。这是一个实用工具的集合，协助集成 Oracle 的服务与 Hadoop Stack。大数据连接器套件是一个工具集，提供深入分析和发现信息的能力，并能快速集成基础设施中存储的所有数据。所有工具均是可扩展的，如果你已经是或者未来将会成为 Oracle 的客户，那么这将很好地适配于你的环境。Oracle 公司的套件中有很多工具，但我们在本章中只会讲述其中的一部分。

Oracle XQuery for Hadoop 运行一个处理流程，它基于 XQuery 语言中表达的转换，将其转化成一系列 MapReduce 作业，这些作业在 Apache Hadoop 群集上并行执行。输入数据可以位于文件系统上，通过 Hadoop 分布式文件系统(HDFS)访问，或者存储在 Oracle 的 NoSQL 数据库中。Oracle XQuery for Hadoop 能够将转换结果写入 Hadoop 文件、Oracle NoSQL 数据库或者 Oracle 数据库。

适用于 Hadoop 分布式文件系统(HDFS)的 Oracle SQL Connector 是一款高速的连接器，用于通过 Oracle 数据库(见图 1-8)加载或查询 Hadoop 中的数据。Oracle SQL Connector for HDFS 将数据放入数据库，数据移动是由 Oracle 数据库中的 SQL 进行数据选择所发起。用户可将数据加载到数据库，或者通过外部表使用 Oracle SQL 在 Hadoop 中就地查询数据。Oracle SQL Connector for HDFS 能够查询或者加载数据到文本文件或者基于文本文件的 Hive 表中。分区也可以在从 Hive 分区表中查询或加载时被删减。

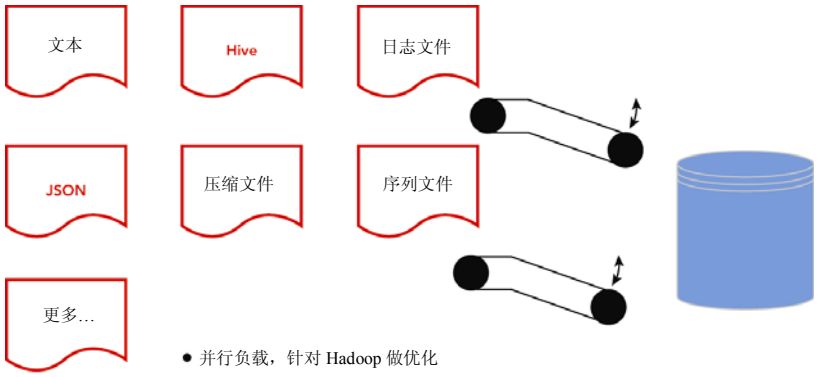
另一种 Oracle 解决方案 Oracle Loader for Hadoop 是一种高性能且高效率的连接器，用于从 Hadoop 中加载数据到 Oracle 数据库。当 Hadoop 发起数据传送时，Oracle Loader for Hadoop 将数据推送到数据库中。如图 1-9 所示。Oracle Loader for Hadoop 利用 Hadoop 计算资源进行排序、分区并在加载之前将数据转换成适配于 Oracle 的数据类型。当加载数据时，在 Hadoop 上进行的数据预处理降低了数据库 CPU 的使用率。这样就减少了对数据库应用程序的影响，减

轻了对资源的竞争，而这正是插入大量数据时的一个常见问题。它使得此连接器在连续且频繁地加载时尤其有用。



- 在 HDFS 上就地访问和分析数据
- 查询和连接 HDFS 数据库中的常驻数据
- 在需要时使用 SQL 加载到数据库中
- 自动负载均衡，从而最大限度地提高性能

图 1-8



- 并行负载，针对 Hadoop 做优化
- 自动负载均衡
- 在 Hadoop 上转换成 Oracle 格式
  ——节省数据库的 CPU
- 加载特定的 Hive 分区
- Kerberos 认证
- 直接加载到 In-Memory 表

图 1-9

Oracle R Connector for Hadoop 能够快速开发，并通过模拟并行的支持，在用户桌面对并行 R 代码使用 R 语言风格的调试功能(见图 1-10)。此连接器允许分析师将来自多种环境(客户桌面、HDFS、Hive、Oracle 数据库和内存中的 R 语言数据结构)的数据组合到单个分析任务执行的上下文中，从而简化数据的组装和准备。Oracle R Connector for Hadoop 也提供了一个通用的计算框架，用于并行执行 R 代码。

如本章所述，如果 Oracle 是贵组织所选用的工具，那么你便有一组工具套件可供选择。它们与 Hadoop 有合作关系，Oracle 网站上有说明文档，并且允许下载前面所提到的所有连接器。此外，还有配置它们以便与 Hadoop 生态系统协同工作的方法。

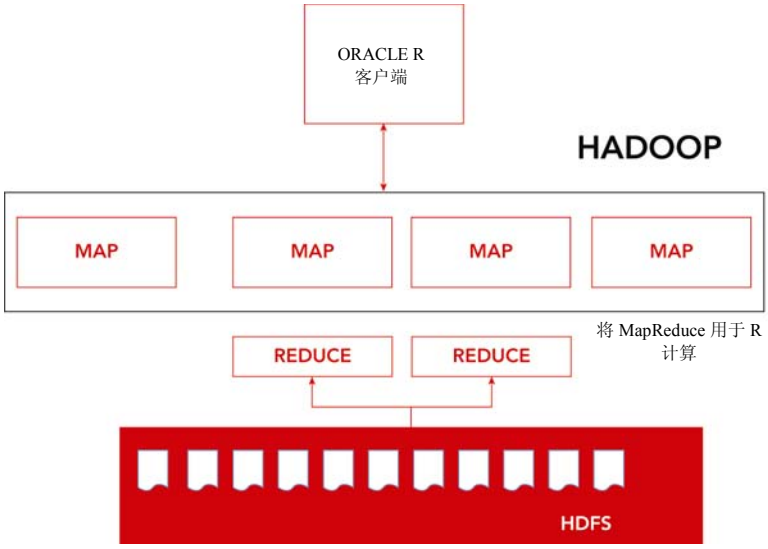


图 1-10

1.5 小结

通过使用 Hadoop Stack，你利用 Hadoop 在企业中实现最优方

案，并且与混合编程和高级工具相结合。如今大多数群集都在你的本地，但服务提供商给予了更多选择，使得数据也可以存储在云端。目前，SQL、关系型和非关系型数据存储均可使用 Hadoop 的功能。

当涉及数据时，Hadoop 已经从长远角度考虑了自身的设计。它非常适用，因为数据会随着时间持续增长。它使用已存在的企业系统，而这些系统可扩展为 Hadoop 数据平台。公司和开源社区中的开发人员正在设计和定义基于 Hadoop 的大规模企业数据的最佳实践。企业以及 IT 社区都非常关注各种数据类型的可扩展性。使用 Hadoop，公司便不再局限于昂贵的企业级解决方案或者价格不菲的数据仓库设备。

Hadoop 并不是大多数组织现有富数据环境的替代品。在考虑使用 Hadoop 时，也要同样重视其他方面，例如 MapReduce 或 YARN，它们在做深度数据分析和高级分析方面取得了重大进步。Hadoop 提供对大数据的实时处理，它能对你的决策结果产生实时影响。不同的产业，从金融业到医疗业，通过使用 Hadoop Stack 或者任何与之相关的组件，均能得到直接收益。它推翻了以前认为只有依靠数据挖掘工具才能实现的界限，使你能够以一种截然不同的方式来查看数据。Hadoop 并不能替代组织查看数据的方式，却能显著提高其查看数据的效率。Hadoop 排除了各种局限性，并且正在各个新领域中继续发展。

理解 Hadoop 的存储系统将使你能够利用数据集成和业务分析来汇总大型数据湖并分析各种数据类型，而且不依赖于它们的当前来源。充分理解 Hadoop 平台能够使其用户实时处理大量可扩展的数据，并提供最优分析。Hadoop 存储流程的突出优点在于没有额外的存储或计算开销，而是存在收益，比如提高数据的准确性并且能够对其进行分析。第 2 章将详细讨论 Hadoop 存储的各个方面。