Intel IT Center

2013 年 2 月

规划指南

大数据入门

详实步骤助力 IT 经理充分利用 Apache Hadoop* 软件

为何阅读本文档

本规划指南为 IT 经理提供了重要信息和实施步骤，以帮助他们对大数据分析项目进行规划和实施，并着手开始使用 Apache Hadoop* 软件。指南具体内容包括：

- 大数据 IT 环境，以及与这一颠覆性力量相关的挑战和机遇
- Hadoop* 软件介绍。Hadoop 是用于从大数据中挖掘重要洞察的新兴标准，包括处理和分析工具（Apache Hadoop MapReduce 和 Apache HBase* 软件）
- 有关如何充分利用 Hadoop 软件的指南，重点介绍了英特尔能够在哪些方面提供帮助，包括基础设施技术、优化和调试等
- 五个基础“未来步骤”，以及一份核对清单，以帮助 IT 经理高效规划和实施 Hadoop 项目

Intel IT Center

2013 年 2 月

规划指南

大数据入门

详实步骤助力 IT 经理充分利用 Apache Hadoop* 软件

英特尔®与你共创明天™

目录

大数据分析 IT 环境

有关大数据分析的讨论正在愈演愈烈。

如今，全球所有企业均需要应对前所未有的数据增长这一挑战。设想一下：截至到 2012 年末，数据世界中的数字容量将扩展到 2.72 泽字节（ZB）。这一数量预计每隔两年将会翻一番，到 2015 年将会达到 8 ZB 数据。[1] 如此庞大的信息量很难让人产生一个直观知识，但我们可以通过以下方式让您有一个大致的概念：美国国会图书馆拥有 462 万亿字节（TB）的数据，8 ZB 则相当于近1,800 万个国会图书馆，[2] 这着实是一个非常庞大的数据量。

大数据的价值

那么，什么是“大数据”，它又来自于何处呢？

大数据指规模达到数量级（*容量*），更加多样化，包括结构化、半结构化和非结构化数据（*多样性*），且以比以往更快的速度生成（*生成速度*）的庞大数据集。这一大规模的数据由众多的联网设备生成，包括 PC、智能手机、以及诸如 RFID 阅读器和交通监视摄像头等传感器设备。另外，这些数据为异构数据，具有多种格式，包括文本、文档、影像和视频等。

到 2015 年将达到 8 ZB 的数据意味着什么呢？这一数据海洋将由近 150 亿台联网设备生成，其中包括 30 亿互联网用户和机对机连接。[3]

大数据的真正价值在于它能够在执行分析时提供深入的洞察 — 利用相关的情报信息发现适合的模式、得出具体的含义、制定明智的决策并最终做出有效的应对措施。

使用大数据分析取胜

大数据是一股颠覆性力量，对于 IT 部门而言其中机遇和挑战并存。麦肯锡全球研究所的研究显示，数据对于企业的重要性正变得与劳动力和资本并驾齐驱。[4] 该项研究指出，如果企业能够有效捕捉、分析、可视化、并应用大数据洞察来实现业务目标，他们将能够从激烈的竞争中脱颖而出，在运营效率和收入方面战胜竞争对手。

大数据分析是 IT 部门面临的一项严峻挑战 — 根据英特尔对 200 名 IT 经理所进行的调查，84% 的 IT 经理已经在对非结构化数据进行分析，而在还没有进行这项工作的 IT 经理中，预计有 44% 将在 2014 年以前开始实施。[5] 大数据的潜在优势是非常显著的。

3V 不仅定义了什么是大数据，同时也指出了 IT 需要解决的几大重要问题：

- **容量。**非结构化数据的庞大规模和增长步伐超过了传统存储和分析解决方案的能力范畴。
- **多样性。**传统的数据管理流程无法应对大数据 — 或“影子数据”或“黑暗数据（dark data）”的异构性质，例如访问跟踪和 Web 搜索历史记录等。
- **生成速度。**数据实时生成，并要求按需提供有用信息。

数据高山

大数据以 TB、PB、甚至 EB 表示。注于图中旨在对繁复数据给予解释。

大数据分析揭秘

大数据分析无疑将会改变企业的运营方式，能够使企业从曾发现的新数据源中获得洞察。以下内容介绍了有关大数据的更多信息。

大数据分析是……

- 一种技术导向战略，旨在从客户、合作伙伴与业务中获取更丰富、深入的洞察，以从根本上获得竞争优势。
- 与数据集相互配合，该数据集的容量、多样性均超过传统数据库软件的捕捉、存储、管理和分析能力。
- 用于处理持续的实时稳定的数据流，以便比以前更快地做出重要决策。
- 具有分布式特性。在数据所在之处进行数据分析处理工作，以加快分析速度并提高效率。
- 提供了一种全新模式，支持 IT 与业务用户和“数据科学家”合作来发现和实施分析功能，以提高运营效率并解决新的业务难题。
- 可推动企业将决策权下放，能够支持员工更快做出更明智的决策。

大数据分析不是……

- 仅仅是一种技术。在业务层面，它关系到如何从庞大复杂的数据源中获得重要洞察。
- 仅关乎容量。它还涉及到数据多样性和生成速度。但也许是最重要的，它还在于从数据中获取价值。
- 仅由诸如谷歌和亚马逊等大型在线网络公司生成或被其使用。虽然互联网公司可能在互联网上应用大数据方面处于主导地位，但如今其应用已覆盖到各行各业。
- 一种构建于共享硬盘和内存架构之上的“一刀切式”传统关系型数据库。大数据分析使用庞大计算资源来实现大规模并行处理（MPP）。
- 用于取代关系型数据库或数据仓库。结构化数据对于公司的重要性继续保持不变。然而，针对大数据的全新来源和环境将让传统系统无能为力。

本指南的目标

本指南的其余内容介绍了用于管理和分析大数据的新兴技术，并重点介绍了如何开始使用 Apache Hadoop* 开源软件框架，它为在计算机集群中分布式处理大型数据集提供了框架。我们同时提供了五个实用步骤，以帮助您使用这一技术规划自己的大数据分析项目。

您的新密友：数据科学家

一类全新的专业人员将可以帮助企业充分利用庞大的信息，他们便是数据科学家。

数据科学家负责对复杂业务问题进行建模、捕获业务洞察并发现宝贵商机。他们具备：

- 娴熟的技能，能够帮助集成和准备构建大型的、多样化的数据集
- 高级分析和建模技能，可帮助发现和理解隐藏的关系
- 丰富的业务知识，以应用具体的信息
- 出色的沟通技能，能够准确呈现成果

数据科学是一个新兴领域。目前这一职业在市场中倍受青睐。找到技能娴熟的此类人才正成为大数据分析需要解决的一个艰巨挑战。数据科学家可能位于 IT 部门或业务部门，但不管他们位于何处，他们都将是您的新密友或合作伙伴，可帮助您更有效地规划和实施大数据分析项目。

用于管理大数据的新兴技术

企业要实现大数据的全部潜能，就必须采用新的方法来捕捉、存储和分析数据。传统的工具和基础设施不能高效处理如今快速生成的更大型、更多样的数据集。

一些新技术的出现，使得大数据分析变得可能，同时更加经济高效。其中一种新方法充分利用了分布式计算资源网格的优势，并使用无共享架构、分布式处理框架和非关系型数据库来重新定义管理和分析数据的方式。

面向大规模可扩展系统的无共享架构

全新的无共享架构可以满足大数据的庞大容量、多样性和生成速度要求，它将工作分散在数十、数百、甚至数万台服务器上，以并行方式独立处理数据。无共享架构最先由诸如 SETI@home 等大型商业研究项目、以及谷歌* 和亚马逊* 等在线服务提供商实施，每个节点都是独立的并且无状态，因此该架构可以轻松扩展，只需添加节点即可，从而使得系统能够处理不断增长的负载。

硬件、数据管理和分析应用技术等领域的进步为无共享架构提供了重要基础。
资料来源：“Data rEvolution”。CSC 前沿论坛（2011 年）。

处理工作在数据驻留的节点上进行。这与检索数据在中心位置中进行处理的传统方法截然不同。

最后，数据必须重新集成，以提供重要信息。计算网格基于分布式处理软件框架开展工作。后者能够管理并在机器间分布数据，向联网服务器发送指令以并行开展工作，收集单个结果，以及重新组合结果来获得重要信息。

分布式处理框架和 Apache Hadoop 的出现

Hadoop* 正在发展成为绝佳的大数据分析方法。作为开源 Web 研究项目 Apache Nutch* 的研究成果，[6] Hadoop 是一个软件框架，提供了简单的编程模型来支持在计算机集群上，以分布式方法处理大型数据集。该框架可轻松采用硬件进行扩展，包括基于英特尔® 至强™ 处理器的服务器等。

Hadoop 软件是用于大数据分析的完整开源框架。它包括一个分布式文件系统、一个并行处理框架（Apache Hadoop MapReduce）和多种不同的组件，支持数据获取、工作流协调、任务管理以及集群监控等功能。Hadoop 能够比传统方法更经济高效地处理大型非结构化数据集。

Hadoop 为大数据分析提供给了多项重要优势，其中包括：

- **以原生格式存储任意数据。**由于数据不需要转换为特定方案，因此不会丢失信息。
- **可进行扩展以支持大数据。**Hadoop 已经过诸如 Facebook 和 Yahoo! 等公司的验证，这些企业已经进行了大量部署，能够出色地进行扩展。
- **提供新洞察。**大数据分析能够发现隐藏的关系。这些关系在使用传统的数据挖掘方法时，或者根本不可能实现，或者非常困难，需要花费大量的时间和成本。
- **降低成本。**Hadoop 是一个开源软件，运行于标准服务器之上，在存储和处理方面有着较低的每 TB 成本。在 Hadoop 中，存储可以按需逐步添加，同时硬件可以添加或从集群中移除。
- **更高的可用性。**Hadoop 通过数据复制以及跨计算节点故障切换提供出色的容错功能，因此能够从硬件、软件和系统故障中快速恢复。
- **较低的风险。**Hadoop 社区非常活跃和多样，由来自全球多个行业的开发人员与用户组成。Hadoop 是一项仍在继续演进的出色技术。

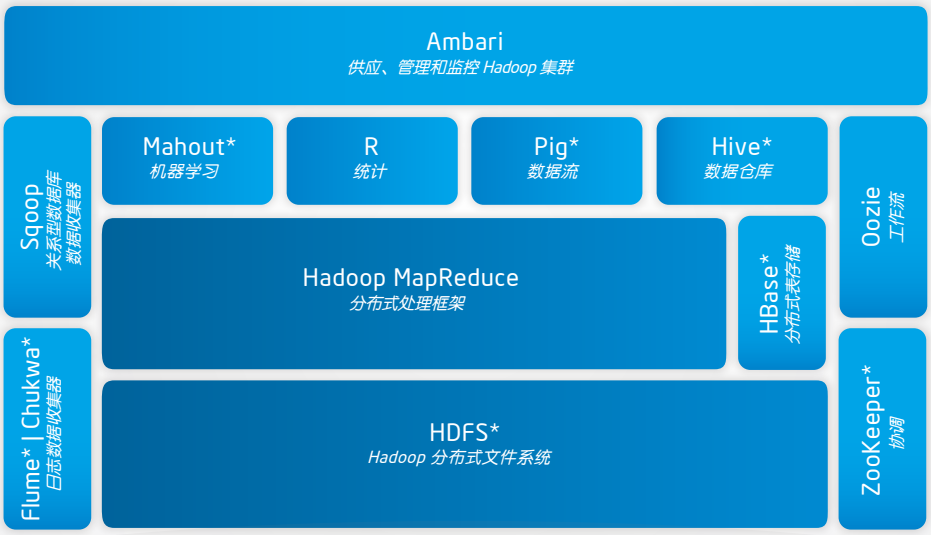
大数据和云

大数据要求通过服务器集群支持相应的工具，以满足大数据的庞大容量、生成速度和多样化的格式要求。云计算已经在服务器池中部署，并且可根据大数据的需求向上或向下扩展。

因此，云计算能够以经济高效的方式为大数据技术和高级分析应用提供支持，进而推动实现出色的业务价值。

了解大数据如何在云计算环境中发挥效力，请参考*《云环境中的大数据：融合技术》*，网址：intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html

Apache Hadoop* Stack



Hadoop* 软件堆栈包括一系列组件。

Hadoop* 软件会取代我现有的数据库系统吗?

Hadoop* 软件是一个大规模可扩展存储和数据处理系统，它不是数据库。事实上，它是您现有系统的重要补充，能够帮助处理超出此类系统能力所限的数据。Hadoop 可以同时吸收和存储来自各种来源的任意类型的数据，以任意方式对其进行整合和处理，并交付给需要的环节。它能够通过您现有的系统支持实时交易数据或提供交互式商业智能。

Apache Hadoop MapReduce 是什么?

MapReduce 是 Hadoop 堆栈中的软件编程框架，能够简化大数据集的处理工作，为编程人员定义和在计算机集群中协调复杂的处理任务提供了一种通用方法。MapReduce 应用的工作原理如下：Map（映射）任务将数据集分拆为独立数据块进行并行处理。之后系统对 Map 任务输出结果进行排序，并提交至 Reduce（化简）任务。这些任务的输入和输出信息均存储在 Apache* Hadoop 分布式文件系统（HDFS*）或诸如 Amazon S3（Amazon Web Services 的一部分）等其它存储中。这一系统通常在相同的节点上处理和存储数据，从而能够更高效地在数据驻留的节点上协调任务，在节点间实现更高的聚合带宽。

MapReduce 通过安排任务、监视活动和重新执行失败的任务来简化应用编程人员的工作。

Apache Hadoop* 概览

Apache Hadoop* 是社区共同努力的结果，包括三个主要的开发子项目，以及其它相关计划。

主要开发子项目	
Apache* Hadoop* 分布式文件系统（HDFS*）	主要的存储系统，使用多个数据块副本，在集群内的节点上进行分配，并为应用数据提供了高吞吐率访问能力
Apache Hadoop MapReduce	一种面向应用的编程模型和软件框架，在计算节点上对大型数据集执行分布式处理
Apache Hadoop Common	支持 Hadoop 框架的实用程序，包括 FileSystem（面向通用文件系统的抽象基类）、远程程序调用（RPC）和序列化库
其它相关 Hadoop 项目	
Apache Avro*	一种数据序列化系统
Apache Cassandra*	一种可扩展的多主数据库，无单点故障
Apache Chukwa*	一种数据收集系统，用于监控在 HDFS 和 MapReduce 上构建的大型分布式系统；包括用于显示、监视和分析结果的工具套件
Apache HBase*	一种可扩展的分布式数据库，支持结构化数据存储，可创建大表；同时支持随机实时读写访问大数据
Apache Hive*	一种数据仓库基础设施，提供了数据汇总和即席查询能力，并支持在 Hadoop 兼容的文件系统中分析大型数据集
Apache Mahout*	一种可扩展的机器学习和数据挖掘库，实施了广泛的算法，包括集群、分类、协作过滤和频繁模式挖掘
Apache Pig*	一种高级数据库语言和执行框架，用于进行并行数据分析
Apache ZooKeeper*	一种高性能中央协调服务，可保持配置信息和命名，为分布式应用提供了分布式同步和群组服务

资料来源：Apache Hadoop，hadoop.apache.org

了解 Apache Hadoop* 专家的观点

与从事开源社区以及相关开发工作的专家直接交流，是了解 Apache Hadoop* 软件及其组件的一种有效方法。收听 Apache Hadoop MapReduce、Apache* HDFS*、Apache Hive*、Apache Pig* 以及 Apache HCatalog 等相关社区负责人的访谈播客，了解每种技术的工作原理、在 Hadoop* 堆栈中的适用范围、后续开发方面的计划等。每次播客同时提供 PDF 文档供您查阅参考。

Hadoop* 采用情况

随着越来越多的企业意识到大数据洞察中蕴含的价值和优势，Hadoop 软件的采用率正在不断增长。Hadoop 开源技术堆栈包括 MapReduce、HDFS 和 Apache HBase* 分布式数据库（支持大型结构化数据表）的开源版本。

经过六年的不断完善，Apache 于 2012 年 1 月推出了首个完整的生产版本 Apache Hadoop 1.0 软件。这一版本支持的认证特性包括 HBase*、Kerberos 安全增强、以及一个用于访问 HDFS 的表述性状态转移（RESTful）API。[7]

Hadoop 软件可以通过 Apache 下载站点进行下载。由于 Hadoop 软件是一个开源志愿者项目，Hadoop wiki 提供了有关从社区中获取帮助的信息，以及教程和最终用户文档链接，以实施和设置集群，并进行故障排除。

加快大数据分析标准

开放数据中心联盟（ODCA）是一个由来自全球 300 多家公司的多名 IT 领导者组成的独立 IT 协会。联盟最近宣布成立数据服务工作组，以应对 IT 在数据管理领域面临的最紧急要求。这一工作组在初期将重点关注创建使用模式要求，使用传统数据管理和数据仓库解决方案来满足新兴大数据框架的安全性、可管理性和互操作性需求。基于使用模式，工作组成员将为商业版本开发参考架构与概念证明，以供独立软件厂商和 OEM 合作伙伴测试部署，并为企业市场推出解决方案。同时联盟还将与开源社区合作，推出性能指标评测套件。作为 ODCA 的技术顾问，英特尔在为大数据分析制定标准和最佳实践方面发挥着重要作用。

Hadoop 价值链

Hadoop 价值链是一个由厂商和解决方案组成的复杂社区，包括多家知名企业和新创公司。几个厂商提供了其自己的 Hadoop 发行版，将基本堆栈与诸如 Apache Hive*、Apache Pig* 和 Apache Chukwa* 等其它 Hadoop 项目结合在一起。部分这些发行版本可与数据仓库、数据库、以及其它数据管理产品相集成，因此数据可在 Hadoop 集群和其它环境之间移动，以扩展要处理或查询的数据池。

其他厂商提供了 Hadoop 管理软件，简化了管理和故障排除工作。同时还有一些厂商提供了产品来帮助开发人员编写 Hadoop 应用，提供搜索功能，以及在不使用 MapReduce 的情况下分析数据。这些产品位于平台软件之上，包括将结构化查询语言（SQL）数据仓库与 Hadoop 集群相连接的抽象层，以及实时处理和分析功能。最后，这些厂商均对通过云提供订阅服务表示出了极大的兴趣。

Apache Hadoop* 推出商业版本

Apache Hadoop* 相关的产品分为多个类别。以下厂商代表了不断壮大的 Hadoop* 价值链的一部分。

类别	厂商/产品
集成 Hadoop 系统	▪ EMC* Greenplum* ▪ 惠普* 大数据解决方案 ▪ IBM* InfoSphere* ▪ Microsoft* 大数据解决方案 ▪ Oracle* Big Data Appliance
采用 Hadoop 连接的 Hadoop 应用和分析数据库	▪ Datameer* 分析解决方案 ▪ Hadapt* Adaptive Analytic Platform* ▪ HP Vertica* 分析平台 ▪ Karmasphere* Analyst ▪ ParAccel* 分析平台 ▪ Pentaho* Data Integration ▪ Splunk* Enterprise* ▪ Teradata* Aster* Solution
Hadoop 发行版	▪ Cloudera 的发行版，包括 Apache Hadoop（CDH） ▪ EMC Greenplum HD ▪ Hortonworks ▪ IBM InfoSphere BigInsights ▪ 英特尔® Apache Hadoop 发行版软件 ▪ MapR* M5 Edition ▪ Microsoft 大数据解决方案 ▪ Platform Computing* MapReduce
基于云的解决方案	▪ Amazon* Web Services ▪ Google* BigQuery

请参阅大数据厂商精选了解提供大数据解决方案的英特尔合作伙伴。

注：Hadoop 价值链正在快速成长。这一列表摘自两个来源：Dumbill, Edd，“Big Data Market Survey: Hadoop Solutions.”，*O'Reilly Radar*（2012 年 1 月 19 日）。http://radar.oreilly.com/2012/01/big-data-ecosystem.html；以及“*Data rEvolution: CSC Leading Edge Forum*”。CSC（2011 年）。http://assets1.csc.com/lef/downloads/ LEF_2011DatarEvolution.pdf

使用 Hadoop 软件进行大数据分析的两种方法

企业正在采用两种基本方法实施 Hadoop。

仅 Hadoop 的部署。Hadoop 部署可作为开源软件从 Apache 免费下载，相关厂商也可以提供发行版软件，即将 Hadoop 框架与特定的软件组件预打包，以便支持系统管理。

仅 Hadoop 的部署为构建大数据管理平台提供了一个理想选择，以分析非结构化数据和从中发现洞察。此外，开源工具也使得使用 MapReduce 应用、HBase 或 Hive* 查询结构化数据成为可能。

集成传统数据库的 Hadoop。建立了传统数据仓库和分析能力的企业可以扩展其现有的平台，以包括集成的 Hadoop 版本。将现有数据管理资源与 Hadoop 连接为分析结构化和非结构化数据以获得洞察提供了重要机会。例如，复杂的呼叫中心脚本分析结果可以与有关购买行为的结构化数据进行关联，包括特定 SKU、零售店面和地理位置等。在这种情况下，专门的连接器可在 Hadoop 和传统环境之间来回移动数据。

英特尔® Apache Hadoop* 发行版软件

英特尔® Apache Hadoop* 发行版软件（英特尔® Hadoop 发行版）包括 Apache Hadoop 和其它软件组件，这些组件经过英特尔专门的优化，具备硬件增强的性能和出色的安全功能。英特尔发行版经过专门的设计，能够在 Apache Hadoop 上支持大量数据分析功能。它针对 Apache Hive* 查询进行优化、为 R* 提供了连接器以进行统计处理，并支持通过 Intel Graph Builder for Apache Hadoop 软件进行图形分析。该软件可将大型数据集汇集到图形中，以便实现数据关系的可视化。英特尔发行版中附带的 Intel Manager for Apache Hadoop 提供了一款管理控制台，可简化 Hadoop* 项目的部署、配置和监控。

英特尔发行版已在全球范围内发布以供用户评估。目前仅在美国、中国和新加坡提供技术支持，其它国家和地区的支持服务预计将在今年晚些时候提供。

了解关于英特尔® Hadoop 发行版的更多信息。

在您的数据中心内部署 Hadoop

大数据分析是一项基于技术的战略，而不仅限于硬件和软件。然而，作为 IT 经理，在数据中心内实施大数据计划的责任将需要由您来负责。Hadoop 部署有着广泛的基础设施要求，同时在设计时选择的硬件和软件将会对性能和总体拥有成本产生重要影响。数据中心可通过确保建立正确的基础设施，并优化和调试 Hadoop 软件以实现最佳性能，来充分利用其 Hadoop 部署。

建立绝佳基础设施

Hadoop 框架在距离数据驻留的地方最近的位置进行计算，通常运行于采用标准硬件构建的大型服务器集群之上。同时数据也在这些集群上进行存储和处理。Hadoop 基础设施和标准服务器平台的组合为经济高效的高性能分析平台提供了重要基础，以支持并行应用。

建立 Hadoop 系统架构

每一个集群都包含一个具有多个从属节点的“主节点”。主节点使用 NameNode 和 JobTracker 功能，并负责协调从属节点来确保完成任务。从属节点使用 TaskTracker 功能来管理通过 JobTracker 安排的任务，另外还使用 HDFS 来存储数据，以及使用 Map 和 Reduce 功能进行数据计算。基本软件堆栈包括面向语言和编译器的 Hive 与 Pig*、面向 NoSQL 数据库管理的 HBase、以及面向日志收集的 Apache Sqoop 和 Apache Flume*。Apache ZooKeeper* 为堆栈提供了中央协调功能。

大数据分析成本

*《InformationWeek》*最近的调查分析了有关大数据经济性的问题，并发现预算限制和其它成本相关问题是 IT 经理面临的最大难题。构建您自己的 Apache Hadoop* 部署项目，以及投资购买存储和开发资源或实施专有厂商解决方案会花费大量的成本。虽然云提供了部分解决方案，但公有云提供商的定价模式可能无法满足全部需求。随着存储和计算成本的持续下降，部署和管理您自己的 Hadoop 集群可能会提供比公共云和厂商系统更佳的经济性。虽然部署自己的集群可能需要雇佣一名技能娴熟的人员来管理硬件，但相比之下其经济性仍然要更为出色。

资料来源：Biddick, Michael，“The Big Data Management Challenge”，*《InformationWeek》*（2012 年 4 月）。http://reports.informationweek.com/abstract/81/8766/business-intelligence-and-information-management/research-the-big-data-management-challenge.html

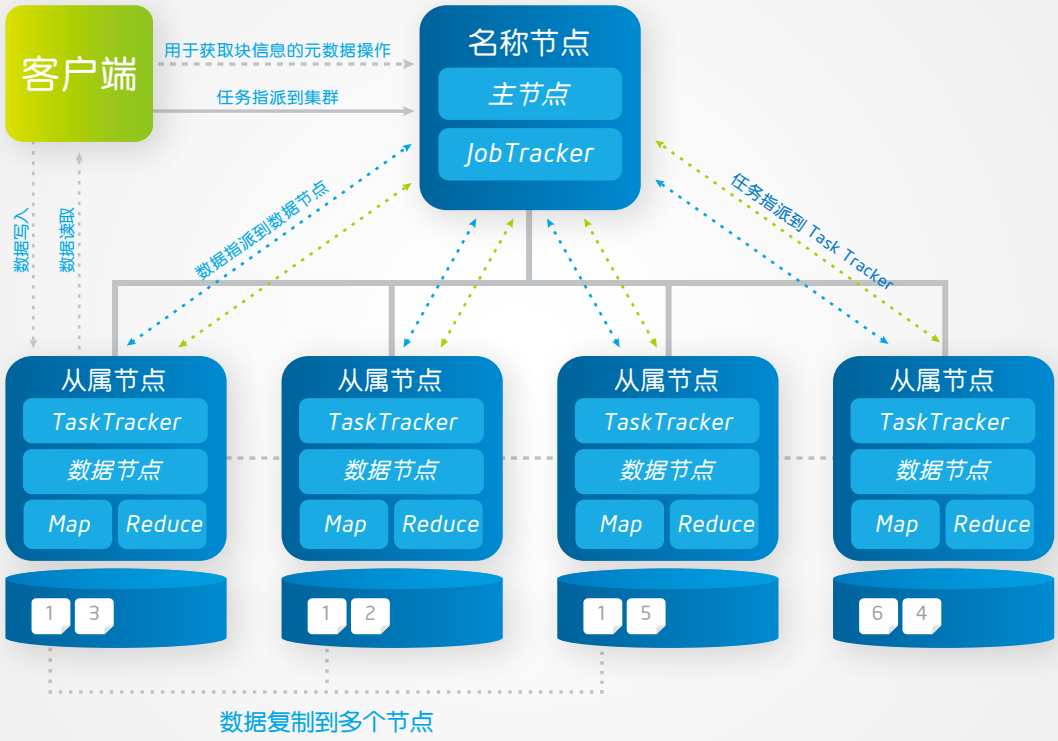
基于标准服务器节点集群的 Apache Hadoop* 部署

资料来源：*在英特尔® 平台上进行云设计与部署的英特尔® 云构建计划指南：Apache* Hadoop**。英特尔（2012 年 2 月）。
intelcloudbuilders.com/docs/Intel Cloud Builders Hadoop.pdf

运行服务器集群

客户端提交任务到主站点，并由后者协调集群上的从属节点。JobTracker 控制 MapReduce 任务，向 TaskTracker 报告。发生故障时，JobTracker 在相同或不同的从属节点中选择最高效的节点，重新安排任务。HDFS 能够识别位置或机架，负责管理集群内的数据，并在不同节点上复制数据，以确保数据可靠性。如果 HDFS 上的一个数据副本发生损坏，知晓其它副本所存储位置的 JobTracker 会重新在副本所驻留的位置上重新安排任务，从而无需在节点间移动数据。此举可以节省带宽，保持卓越的性能和可用性。当任务完成映射后，系统将对其输出结果进行排序并分成几个组，然后分发给 Reducer。Reducer 可能与 Mapper 位于相同的节点上，也可能位于不同的节点上。

Apache Hadoop* 集群的运行



主节点协调任务，并安排在从属节点上处理。

Hadoop 基础设施：大数据存储和网络

得益于主流计算和存储资源的大幅改进，以及万兆位以太网（10 GbE）解决方案的推出，Hadoop 集群得到了显著的增强。万兆位以太网实现的更高带宽对于在服务器间导入和复制大型数据集至关重要。英特尔® 万兆位以太网融合网络适配器提供了高吞吐率连接，同时英特尔固态硬盘为原始存储提供了高性能、高吞吐率硬盘。要提升效率，存储需要能够支持高级功能，包括压缩、加密、自动数据分层、重复数据删除、删除码和精简配置等。最新的英特尔® 至强™ 处理器 E5 家族支持所有这些功能。

获得相关的指南在 10 GbE 上构建平衡、经济高效的 Hadoop 集群。

英特尔® 架构：高性能集群

使用基于英特尔® 至强™ 处理器的服务器作为集群的基准服务器平台，一支由英特尔大数据、网络和存储专家组成的团队针对不同的网络和存储组件组合，对 Apache Hadoop* 的性能结果进行了测量。下表针对您的大数据环境的英特尔基础设施的性能定义了三种不同的选择，它们分别是“好”、“更好”和“最好”（请注意，特定变量可能会影响数据中心的结果）。

性能	服务器	网络	存储
好	英特尔® 至强™ 处理器 E5 家族	千兆位以太网或万兆位以太网	传统硬盘
更好	英特尔® 至强™ 处理器 E5 家族	万兆位以太网	传统硬盘和支持分层存储功能的固态硬盘
最好	英特尔® 至强™ 处理器 E5 家族	万兆位以太网	固态硬盘

了解关于每种平台组合性能的详细信息。

优化和调试以实现最佳性能

英特尔是诸如 Linux*、OpenStack*、KVM 和 Xen* 软件等开源计划的主要贡献者。此外，英特尔还投入了大量资源进行 Hadoop 分析、测试和性能特征化工作，包括在内部开展工作和与惠普、美超微以及 Cloudera 等合作伙伴一同进行。通过这些技术努力，英特尔已注意到了硬件、软件和系统设置方面的许多重要权衡因素，此类因素在数据中心中产生了重要影响。设计解决方案堆栈来提高工作效率、限制能耗和降低总体拥有成本可帮助您优化资源利用率，同时降低运营成本。

Hadoop 环境的设置对于能否充分利用其余硬件和软件解决方案的优势至关重要。通过使用基于英特尔处理器的架构在实验室以及客户站点所进行的广泛的性能指标评测，英特尔可提供面向 Hadoop 系统的优化与调试建议，这些建议在充分考虑性能和成本的前提下对您的 Hadoop 环境进行配置和管理。

为确保正确的设置，您需要提前进行大量的准备工作，因为根据具体的任务和工作负载，每个企业 Hadoop 系统的要求会有很大的差异。当然，您在工作负载优化方面所投入的时间会得到丰厚的回报，这不仅体现在系统性能的提升，而且整个 Hadoop 环境的总体拥有成本也会显著下降。

性能指标评测

性能指标评测为衡量任意计算机系统的效率提供了一个量化基础。英特尔开发了 HiBench 套件作为 Hadoop 环境的完整性能指标评测集合。[8] 其中的单项衡量结果代表了重要的 Hadoop 工作负载，这些工作负载具有一系列不同的硬件使用特征。HiBench 包括微性能指标评测，以及实际 Hadoop 应用，代表广泛的数据分析任务（例如，搜索索引和机器学习）。HiBench 2.1 目前作为开源软件提供，使用需遵守 Apache License 2.0，具体请见：https://github.com/hibench/HiBench-2.1

HiBench：详细信息

英特尔 HiBench 套件将 10 种工作负载分成四个类别。

类别	工作负载	简介
微性能指标评测	排序	• 该工作负载负责对由 Apache Hadoop* RandomTextWriter 等程序生成的二进制输入数据进行排序。 • 代表了负责将数据从一种格式转换为另一种的实际 MapReduce 任务。
	字数统计	• 该工作负载会统计使用 Hadoop* RandomTextWriter 等程序生成的输入数据中每个单词的出现次数。 • 代表了用于从大型数据集中提取出少量感兴趣的数据的实际 MapReduce 任务。
	TeraSort	• 一种针对由 TeraGen 程序生成的大型数据排序任务的标准性能指标评测。
	增强的 DFSIO	• 测试 Hadoop 集群的 Apache* HDFS* 系统吞吐率。 • 通过按固定时间间隔从每个 Map 任务中对读取/写入的字节量进行采样，计算聚合带宽。
Web 搜索	Apache Nutch* 索引	• 这一工作负载测试了常见的 Apache 开源搜索引擎 Nutch* 的索引子系统。Nutch 中的 Crawler 子系统用于分析内部 Wikipedia* 镜像，会生成 8.4 GB 压缩数据（约 240 万网页）作为工作负载输入。 • 大型索引系统是 MapReduce 最重要的一种使用情况（例如谷歌* 和 Facebook* 平台）。
	页面排名	• 这一工作负载是页面排名算法的开源版本。该算法是一种链路分析算法，在 Web 搜索引擎中被广泛应用。
机器学习	K-Means 集群	• MapReduce 在大型数据挖掘和机器学习方面的主要应用（例如谷歌和 Facebook 平台）。 • K-Means 是一种非常有名的集群算法。
	贝叶斯分类	• MapReduce 在大型数据挖掘和机器学习方面的主要应用（例如谷歌和 Facebook 平台）。 • 该工作负载会测试 Apache Mahout* 开源机器学习库中的原生贝叶斯（一种知名的知识发现和数据挖掘分类算法）训练器。
分析查询	Apache Hive* Join	• 该工作负载通过计算单个只读表中每个小组的总和，模拟了结构化（关系型）数据表的复杂分析查询工作。
	Hive* Aggregation	• 该工作负载通过合并两个不同的数据表来计算每个小组的平均值和总和，模拟了结构化（关系型）数据表的复杂分析查询工作。

五大步骤及核对清单：开始您的大数据分析项目

通过上文的介绍，您现在应该已经出色地了解了大数据 IT 环境，其对于企业的潜在价值，以及可帮助您从这些非结构化数据源中获取洞察的新兴技术。此外，您也基本了解了如何建立正确的基础设施，以平稳支持您的 Hadoop 计划。

通过以下五大步骤，现在您可以开始自己的大数据分析项目。

第 1 步：与您的业务用户合作确定巨大商机。

- ☐ 与业务用户（包括分析师、数据科学家和营销专业人员等）合作，从而发现在您企业中开展大数据分析的最佳商机。例如，考虑当前面临的业务难题，尤其是当使用目前的数据源和分析系统时需要花费大量成本、且难以、甚至是不可能完成任务时。或者在考虑全新的非结构化数据源带来的全新难题时。
- ☐ 对您的机会列表进行优先顺序排序，选择有着明确投资回报的项目。
- ☐ 确定成功完成计划所需要的技能。

第 2 步：开展研究以加快技术采用速度。

- ☐ 与 IT 同仁交流。
- ☐ 充分利用英特尔 IT 中心提供的大数据资源。
- ☐ 了解厂商产品。
- ☐ 阅读 Apache 提供的教程和用户文档。

第 3 步：为您的项目制定使用案例。

- ☐ 制定使用案例以开展项目。
- ☐ 绘制数据流，定义解决业务难题需要哪些技术和大数据功能。
- ☐ 确定要包含和去除哪些数据，从而发现能够生成重要洞察的战略数据。
- ☐ 确定如何关联数据，以及业务规则的复杂性。
- ☐ 确定生成目标结果需要使用的分析查询功能及算法。

第 4 步：发现现状和未来功能之间的差距。

- ☐ 您具有哪些额外的数据质量要求，以收集、整理并将数据整合为可用的格式？
- ☐ 您需要建立哪些数据治理策略，以对数据进行分类、定义相关性、并对数据进行存储、分析和访问？
- ☐ 您需要提供哪些基础设施功能，以确保出色的可扩展性、低延迟和高性能？
- ☐ 数据如何呈现给用户？研究结果需要以易于理解的方式呈现给从高管到信息专家在内的所有业务用户。

第 5 步：为生产版本建立测试环境。

- ☐ 根据您企业的情况调整参考架构。英特尔正在与领先的合作伙伴合力制定参考架构。这些架构是英特尔云构建计划的一部分，可为建立大数据使用案例提供重要支持。
- ☐ 定义展示层、分析应用层、数据仓库、以及适用的私有或公共云数据管理要求。
- ☐ 通过有意义的方式确定用户成果展示所需的有效工具，用户所采用的工具对于项目的成功有着重要影响。

来自英特尔的更多资源

关于大数据

大数据分析（“汇总”页面）

该页面汇聚了英特尔提供的重要资源，可帮助您有效实施您的大数据计划。您可以通过“英特尔 IT 中心”访问该页面，以便获得规划指南、同行研究、厂商解决方案信息以及真实案例等丰富资源。

intel.com/bigdata

挖掘企业大数据，成就更出色商业智能

这份来自英特尔 IT 部门的白皮书介绍了英特尔如何部署系统和开发技能，以深入分析大数据，进而推动实现更出色的运营效率和竞争优势。英特尔 IT 部门与相关的业务部门密切合作，联手针对大数据平台部署多个概念验证计划，其中包括恶意软件检测、芯片设计验证、市场情报和推荐系统。

http://www.intel.cn/content/www/cn/zh/big-data/i08025-zho-mining-big-data.html

IT 揭秘：大数据

在这一播客中，英特尔大数据商业智能战略团队的领导者 Monty Fania 介绍了如何提高大数据分析所需的必要技能以及合适的平台。

http://connectedsocialmedia.com/intel/5773/inside-it-big-data/

同行研究：大数据分析

英特尔对 200 名 IT 经理进行了一次调查，相关的调查结果可帮助您更好地了解企业当前如何利用大数据，其中包括企业向前发展所要采取的措施，以及整个研究对于 IT 行业意味着什么。请收看视频“IT 经理畅谈大数据分析”，了解相关的重点内容。

intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html

大数据思想家

我们就大数据采访了多位业内知名的思想领袖，其中包括 LiveRamp 首席执行官 Auren Hoffman（访谈内容：大数据发展促进业务竞争）；Forrester 首席分析师 Mike Gualtieri（访谈内容：未来发展趋势）；以及 Cognito 首席执行官 Joshua Feast（访谈内容：大数据、人类行为以及业务成果）。

intel.com/content/www/us/en/big-data/big-thinkers-on-big-data.html

关于 Hadoop 软件

Apache Hadoop 热点聚焦

访问此页面收听 Apache Hadoop 开源社区专家介绍 Hadoop 堆栈中相关软件组件的工作原理以及未来的发展方向。此页面汇集了多位专家就不同组件的访谈播客，其中包括 Hortonworks 的 Alan Gates（HCatalog 和 Pig）、AltoScale 的 AltoScale（HDFS）、Hortonworks 的 Deveraj Das（MapReduce）以及 Cloudera 的 Carl Steinbach（Hive）。

intel.com/content/www/us/en/big-data/big-data-aDache-hadooD-framework-SDOtliahts-landing.html

*英特尔® 云构建计划指南 — 在英特尔® 平台上进行云设计与部署：Apache*Hadoop**

这一参考架构面向希望构建自己的云计算基础设施的企业，包括使用 Apache Hadoop 集群来管理大数据。它包含了在数据中心实验室环境中设置部署的步骤，并详细介绍了 Hadoop 拓扑结构、硬件、软件、安装、配置和测试信息。实施该参考架构将可以帮助您构建并运行自己的 Hadoop 基础设施。

http://www.intel.cn/content/www/cn/zh/cloud-computing/cloudbuilding-guide.html

优化 Hadoop 部署*

这一白皮书为计划部署 Hadoop 的企业提供了指南。它基于英特尔使用 Hadoop 开展的广泛的实验室测试，描述了用于建立服务器硬件规格的最佳实践，讨论了服务器软件配置环境，并提供了配置和调试建议以改进性能。

intel.com/content/www/us/en/cloud-computing/cloud-computing-optimizing-hadoop-deployments-paper.html

其它资源

Big Data: Harnessing a Game-Changing Asset（大数据：充分利用颠覆性资产）

该报告由 Economist Intelligence Unit 在 SAS 的赞助下发布，介绍了大数据及其对于公司的影响。调查分析了已开始从数据中挖掘价值的公司的组织特征，并发现了有效的数据管理与出色业绩之间存在着的紧密联系。这些公司可为其他希望有效管理并从大数据中获得价值的企业树立榜样。

sas.com/resources/asset/SAS_BigData_final.pdf

The Forrester™ Wave: Enterprise Hadoop Solutions, Q1 2012（Forrester™ Wave：企业 Hadoop 解决方案，2012 年第一季度）

本报告由 Forrester 公司的 James Kobielus 撰写，介绍了 13 个企业级 Hadoop 解决方案提供商，应用 15 项标准对每个提供商进行了评估。具有代表性的包括 Amazon Web Services、IBM、EMC Greenplum、MapR、Cloudera 和 Hortonworks 等。

forrester.com/The+Forrester+Wave+Enterprise+Hadoop+Solutions+Q1+2012/quickscan/-/E-RES60755

尾注

1. Gens, Frank。“*IDC Predictions 2012: Competing for 2020*”。IDC（2011 年 12 月）。http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf
2. “Big Data Infographic and Gartner 2012 Top 10 Strategic TechTrends”。*Business Analytics 3.0（博客）*（2011 年 11 月 11 日）。http://practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and- gartner-2012-top-10-strategic-tech-trends/
3. “Global Internet Traffic Projected to Quadruple by 2015”。*The Network（新闻稿）*（2011 年 6 月 1 日）。http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=324003
4. “*Big Data: The Next Frontier for Innovation, Competition, and Productivity*”。麦肯锡全球研究所（2011 年 5 月）。mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.pdf
5. “*Peer Research on Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data*”。英特尔（2012 年 8 月）intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html
6. Nutch* 软件最初是一个由 Doug Cutting 和 Mike Cafarella 创立的独立开源项目。2005 年，Nutch 接受 Apache 软件基金会的管理，成为 Apache Lucene（搜索软件）的一个子项目，并于 2010 年成为基金会的高级项目。资料来源：“Nutch 加入 Apache Incubator”（新闻稿）。Apache 软件基金会（2005 年 1 月）。nutch.apache.org/#January+2005%3A+Nutch+Joins+Apache+Incubator
7. “Hadoop Hits Primetime with Production Release”。Datanami（2012 年 1 月 6 日）。datanami.com/datanami/2012-01-06/hadoop_hits_primetime_with_production_release.html
8. Huang, Shengsheng、Jie Huang、Jinquan Dai、Tao Xie 和 Bo Huang。“*The HiBench Benchmark Suite: Characterization of the MapReduce-Based Data Analysis*”。IEEE（2010 年 3 月）。

来自英特尔® IT 中心的更多资源

“规划指南：大数据 入门”*由英特尔® IT 中心（英特尔的 IT 专业人员计划）提供。该英特尔 IT 中心致力于提供直观易懂、正确、且无偏见的信息，帮助 IT 专业人员按需实施其战略项目，包括虚拟化、数据中心设计、云、客户端和基础设施安全性等。英特尔 IT 中心提供以下资源：

- 规划指南、同行研究和解决方案亮点，以帮助您实施重要项目
- 真实案例研究，显示您的同行如何解决与您面临的相同挑战
- 有关英特尔 IT 部门如何实施云、虚拟化、安全性和其它战略计划的信息
- 相关活动信息。在这些活动上，您可以聆听到英特尔产品专家以及英特尔内部 IT 专业人员的真知灼见

如欲了解更多信息，请访问：intel.com/ITCenter

与同事分享

3213/RF/ME/PDF-CN 328687-001

英特尔®与你共创明天™