2013 年 8 月

管理简报

企业大数据部署业务案例

对比 IBM InfoSphere BigInsights 和开源 Apache Hadoop 的使用成本、优势与风险

International Technology Group

609 Pacific Avenue, Suite 102
Santa Cruz, California 95060-4406
电话：831-427-9260
电子邮件：Contact@ITGforInfo.com
网站：ITGforInfo.com

目录

图表列表

执行摘要

挑战和解决方案

对于大数据，还有什么好说的？事实证明有许多话题可以说。

行业争议开始逐步关注大数据分析在变革业务决策制定过程以及企业竞争力方面可能扮演的角色。这些分析的影响显然是革命性的。但还是存在不足之处。一些瓶颈正在逐步出现，可能严重减慢大数据在许多（或许是大部分）企业中发挥潜力的步伐。

具体来讲，围绕 Apache Hadoop 开发的复杂技术就属于此情况。随着基于 Hadoop 的系统的使用范围扩大到社交媒体公司以外，用户发现开发人员的生产力常常很低 - Hadoop 需要大量手动编码工作 – 而且编码技能的短缺减缓了新项目的启动速度，并增加了部署时间和成本。

Hadoop 专家已经进入 IT 领域薪酬最高者的行列。在美国，Hadoop 开发人员的起始工资通常在 100,000 美元以上，经理、数据科学家、架构师和其他高级专家的工资常常高达 200,000 美元。在全球，Hadoop 人员的薪酬正在快速上涨，而且这一趋势预计在可预见的未来仍会持续下去。

较低的开发人员生产力和高薪资水平相结合，导致了低迷的经济形势。通常认为，这一形势受到了大部分 Hadoop 组件都是开源的和可免费下载等事实的制衡。但其总体成本不一定比供应商管理的 Hadoop 发行版低，这些发行版支持更为经济高效的开发和部署工作。

对比为 6 家公司内有代表性的高影响力应用程序使用开源 Hadoop 和 IBM BigInsights Hadoop 发行版的 3 年成本，就可以证明此结论。平均而言，使用 IBM BigInsights 的总成本要低 28%。

这些对比（图 1 中总结了结果）包含软件许可、BigInsights 的使用支持和人员成本，以及仅使用开源 Hadoop 软件的人员成本。

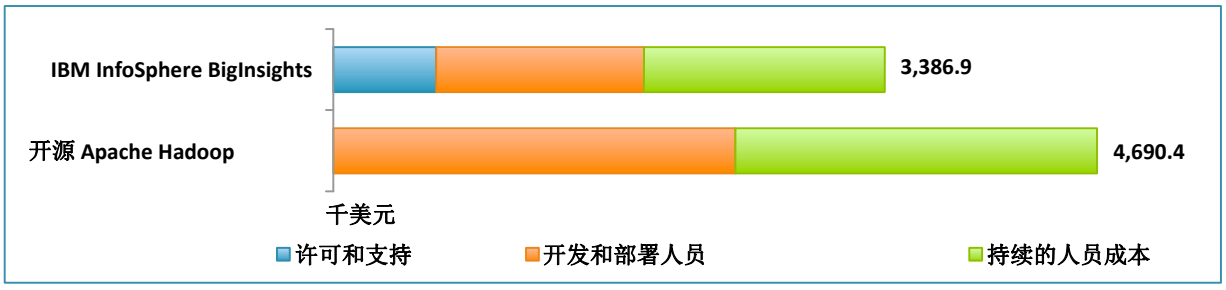

图 1：*为重要应用使用 IBM InfoSphere BigInsights 和开源 Apache Hadoop 的 3 年成本 - 所有安装的平均值*

人员成本是指初始应用程序开发和部署，以及 3 年内的生产后期持续操作的人员成本。

此处的计算中包括了执行初始开发和部署的数据科学家、架构师、项目经理、开发人员及数据和安装专员，以及执行生产后期操作的开发人员、数据专员和系统管理员。BigInsights 成本还包括许可和支持成本。

这些对比用于对金融服务、医疗健康、市场营销服务、媒体、零售和电信公司进行综合剖析。并且根据 29 家采用 BigInsights、开源工具或二者组合的公司所提供的信息来执行这些剖析工作。

有关为计算采用的剖析、方法和假设的进一步信息，以及各个公司的成本分解，可在本报告的“详细数据”部分中找到。

开源

Hadoop 的开源版本主要基于 Google 在 21 世纪前几年所开发的技术。最早且迄今为止最大的 Hadoop 用户是社交媒体和电子商务公司。除了 Google 本身，这些公司包括 Amazon.com、AOL、eBay、Facebook、LinkedIn、Twitter、Yahoo 和类似的国际公司。

尽管使用该技术的公司范围自那时起已经扩大，包含了数百家由风险资本资助的创业企业，以及既有的系统和服务供应商和大型最终用户，但社交媒体企业继续控制着 Hadoop。在 Apache Hadoop 体系中，目前其十几亿行代码中的大部分（根据一些估算数据，超过 90%）是由这些公司贡献的。

这个群体所关注的优先事项不可避免地会影响 Hadoop 的演变。人们倾向于这样来臆断，Hadoop 开发人员的技能非常出众，能够随着需求的改变而逐个案例地处理“原始”开源代码并配置软件组件。手动编码就是目前的标准。

数十年的经验表明，无论采用哪些技术，与更为全面的技巧相比，手动编码只能实现较低的开发人员生产力并导致更高的出错可能性。

随着业务需求和数据源的改变，可能需要实现持续更新。随着时间的推移，这可能会生成复杂的、未经充分备案的大量“无头绪代码”，对其进行维护和增强需要很高的成本。这些问题通常会影响那些采用遗留的大型机应用程序的企业。在新一代软件技术中重蹈此类覆辙毫无意义。

当然，还出现了其他问题，包括：

- ***稳定性。***目前人们已通过至少 25 个不同的 Apache Foundation 项目、子项目和孵化器来定义和增强 Hadoop 及其开源软件体系。随着 Hadoop 环境的扩大，新工具和技术的出现，预计会看到更多的此类项目。

  各个计划的进展速度可能不同，发布日期充其量只经过了简单的协调。开发人员暴露在持续的变更流中。

  不稳定性可能是一项巨大的挑战。对技术战略进行规划变得更加困难，项目时间表和成本变得更无法预测，而项目失败的风险则变得更高。未来出现互操作性问题的可能性也在增加。

  而且，Apache 体系可能以一种无法预测的方式进行演变。可能企业对各个组件进行标准化后，才发现这些组件获得的关注在不断下降。而社交媒体公司中的技术变更步伐，比其他大部分行业中的用户所习惯的变更步伐都要快得多。

- ***互操作性。***在不需要与其他应用程序和数据库进行互操作方面来说，很少（如果有）有基于 Hadoop 的系统是“独立的”。

举例而言，此报告中调查的所有企业都已采用或计划采用关系型数据库、数据仓库、传统分析工具、查询和报告内部网，以及/或 CRM 和后端系统的接口。甚至基于 Hadoop 的服务的“单一业务”提供商也是如此。

互操作性需求在金融服务、医疗保健、保险、零售和电信公司中特别重要。例如，一家大型银行机构报告称，它预计要实施 40 到 50 个不同的接口，其基于 Hadoop 的系统才能完全正常运行。

- ***弹性。***开源 Hadoop 产品体系包含各种为保持可用性而设计的机制，而且支持在发生计划外的（如意外的）断电，以及为软件修改、计划维护和其他任务而制定的计划内宕机时，能够进行故障转移和恢复。

  但是，这些机制远没有传统业务关键型系统那么成熟。如果环境中包含大量手动配置的组件，则会更为复杂且更容易出错。系统经历频繁的更改时，各种漏洞会被放大。

  主要社交媒体公司常常实现了较高的可用性水平。但是，这通常需要投入大量资本来加强软件，确保冗余性，以及提供深入操作监视和响应人员，还有各种流程。

- ***可管理性。***开源 Hadoop 限制已在一些领域中显现出来，如配置和安装、监视、作业调度、工作负载管理、调优和可用性，以及安全管理。尽管一些开源组件解决了这些问题，但它们在大多数 Apache 贡献企业中的优先级相对较低。

  在一定程度上，用户可能通过“劳动密集型”的管理实践来弥补这些限制造成的不足。此方法不仅导致更高的人员成本，而且并不太可靠。

  开源的可管理性限制可能在应用程序开发和部署期间不太明显。但是，它们将在系统管理人员更高的持续全职当量 (FTE) 中反映出来，可能会影响生产后期的服务质量。

- ***支持。***开源软件仅通过社区支持来提供，也就是说，用户要依赖在线同行论坛来获取增强、技术建议并解决问题。此方法可能适合平常所遇到的问题，但要取决于其他人是否愿意分享其时间和经验。事实证明，此方法在处理特定于企业的配置问题方面非常不可靠。

  实际限制可能非常显著。问题解决方面的延迟可能会降低开发人员的生产力，可能导致应用程序错误、性能瓶颈、运行中断、数据丢失和其他负面影响。

随着 Hadoop 的部署量不断增加，这些问题催生了供应商管理的付费发行版，它们包含增强的工具和功能，而且提供了更有效的客户支持。

目前的示例包括 Amazon Elastic MapReduce (Amazon EMR) Web 服务、Cloudera 的 Distribution Including Apache Hadoop (CDH)、EMC 的 Pivotal HD、Hortonworks Data Platform、IBM InfoSphere BigInsights、Intel Distribution for Apache Hadoop (Intel Distribution) 和 MapR M 系列。

IBM InfoSphere BigInsights 的独特优势

BigInsights 环境目前包含图 2 中所示的组件。

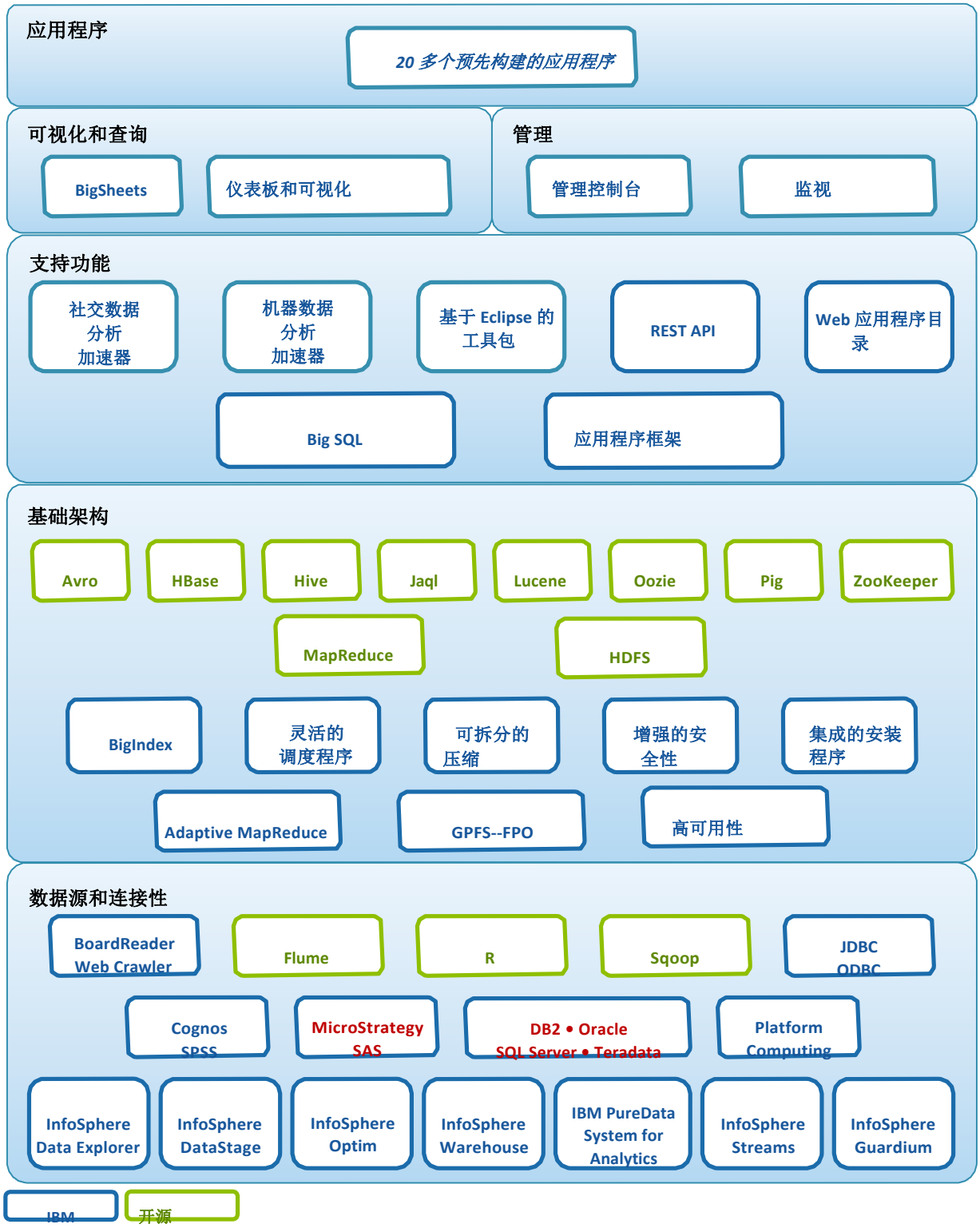

图 2：*IBM InfoSphere BigInsights 环境*

尽管此环境包含完整的 Apache Hadoop 产品体系，但其不同之处在于包含了众多 IBM 组件，这些组件解决了上述问题。在 BigInsights Version 2.1 中（已于 2013 年 6 月面世），这些组件可总结如下：

- ***可视化和查询工具***包含 IBM BigSheets，一个用于识别、集成和探索非结构化和/或结构化数据模式的，可高度自定义的最终用户分析解决方案。它采用了一种类似电子表格的界面，但比传统的电子表格更复杂，在可处理的数据量方面没有限制。

- ***管理工具***包含一个基于 Web 的管理控制台，为所有应用程序和基础架构组件的监视、健康检查和管理提供了一个通用、高效的界面。集成的安装程序可自动完成所有组件的配置和安装任务。

- ***开发工具***包含***基于 Eclipse 的工具包和一个 Web 应用程序目录***，前者支持主要的 Hadoop 开发工具和语言，后者包含即席查询、数据导入和导出，以及专为快速原型制定而设计的测试应用程序。

- 针对社交媒体和机器数据分析的***加速器***包含了多个预先构建的模板和组件，用于实现众多特定于行业和应用程序的功能。人们根据客户的体验开发了加速器，这显著改善了基于 Hadoop 的应用程序开发和部署工作的“价值实现速度”。未来有望实现更多的功能。

  文本分析功能已整合为 BigInsights 的一个标准特性。社交媒体和机器加速器包含了多种文本提取器，可用于各种应用程序领域。

- ***Big SQL***（BigInsights 2.1 中引入）是一种原生的 SQL 查询引擎。它允许开发人员利用现有的 SQL 技能和工具来查询 Hive、HBase 或分布式文件系统数据。开发人员可使用标准 SQL 语法来工作，并且在某些情况下，可使用为配合 Hadoop 工作而优化的且由 IBM 提供的 SQL 扩展。

  Big SQL 为类似 SQL 的 HiveQL（Facebook 开发的一个 Hive 扩展）提供了一个替代方案。Big SQL 更易于使用，而且与主流 SQL 开发工具和技术的一致性更高。它还合并了一些可改善某些应用程序和工作负载运行时性能的特性，这些是原生 HiveQL 中所没有的。

  此方法有望得到广泛的采用。尽管熟练的 Hadoop 专家仍然相对较少，但 SQL 自上世纪 80 年代以来已得到了广泛的使用。据信全球有超过 400 万开发人员熟悉此语言。大部分大型企业长期在 SQL 技能集，以及基于 SQL 的应用程序和工具方面进行投资。

- 在一些领域提供了***基础架构增强***，如大规模索引 (BigIndex)、作业调度 (BigInsights Scheduler)、管理和监视工具、可拆分的文本压缩和安全性。

  BigInsights 支持 ***Adaptive MapReduce***，后者利用了 Platform Symphony 中的 IBM 工作负载管理技术。Adaptive MapReduce 允许更高效地执行更小的 MapReduce 作业，支持比开源 MapReduce 更有效、开销更低的方式来管理混合工作负载。

- ***Platform Symphony*** 是一个高性能的网格中间件解决方案，最初由 Platform Computing（于 2012 年被 IBM 收购）开发。在 BigInsights 中，它可用于取代开源 MapReduce 层，同时允许用相同的方式创建 MapReduce 作业。客户可自行选择安装哪个产品。

- ***IBM General Parallel File System – File Placement Optimizer (GPFS-FPO)*** 是 IBM GPFS 分布式文件系统的一种 Hadoop 优化实现，提供了 HDFS 的替代方案。10 多年来，GPFS 已被广泛部署在科学和技术计算，以及大量的商业应用中。

  除了提供更高的性能，GPFS 还实现了更高的集群可用性，并且具有比 HDFS 更有效的系统管理、快照复制、故障转移和恢复，以及安全性功能。（IBM 并不是唯一采用此方法的公司。目前使用 HDFS 替代方案的 Hadoop 用户越来越多，如 MapR 文件系统、Cassandra 和 Lustre。）

- ***高可用性***特性包括增强的 HDFS NameNode 故障转移。IBM 实现支持无缝且透明的故障转移。该流程是自动的，无需管理员进行干预，而且比传统开源环境中的执行速度更快且更为可靠。Platform Symphony 还提供了更多的复杂特性。

- ***互操作性工具***符合广泛的行业标准要求，并且/或者其设计宗旨是与关键的 IBM 和第三方数据库和应用程序解决方案相集成。

  目前提供了常用开源软件的接口；兼容 JDBC 和 ODBC 的工具；IBM DB2、Oracle、Microsoft SQL Server 和 Teradata 数据库；以及构成该公司大数据平台的重要 IBM 解决方案。

  这些解决方案包括 InfoSphere Information Warehouse 数据仓库框架；Cognos 商业智能；SPSS 统计建模和分析；InfoSphere DataStage 提取、转换和加载 (ETL) 工具；用于企业安全管理的 InfoSphere Guardium；Platform Symphony 高性能网格中间件；以及 IBM PureData System for Analytics 设备。

  Web Crawler 应用程序可自动完成互联网搜索工作，并可根据用户定义的条件来收集数据。可将数据导入 BigSheets。

BigInsights 兼容 IBM ***InfoSphere Streams***，而且常常与之同时使用来执行实时大数据分析。此解决方案在架构上与开源的 Storm 相似，但包含针对开发生产力、可管理性、恢复能力和互操作性的众多 IBM 增强。BigInsights 包含使用首先的 InfoSphere Streams 许可。

BigInsights 功能正在快速演变。IBM 致力于集成新出现的开源组件，而且该公司因研究各种其他的功能增强而闻名。

总结

Hadoop 的使用仍处在早期阶段。除了少量的主要社交媒体公司，大部分 Hadoop 部署都出现在最近的两年。行业调查表明，许多产品仍未部署到生产环境中。

但是，Hadoop 的采用范围正在快速扩大，而且显然大数据将成为大多数企业 IT 环境中的一个首要特性。在此情况下，技术体系和部署模式将不可避免地发生更改。可以预期，与以前的开源技术浪潮一样，Hadoop 市场的划分将变得更加详细，并且解决方案产品也将变得更加多样化。

企业用户（该类别可能包含许多中型企业以及创业公司）将不可避免地倾向于采用更加高效、弹性更高且供应商支持的发行版。

许多企业还会倾向于融合的 Hadoop 和 SQL 环境，应用 SQL 技能集和应用程序组合来应对新的大数据挑战。人们还倾向于使用 Hadoop 数据子集或聚合来增强基于 SQL 的数据仓库。

有这些趋势的企业将越来越多地充分利用更广泛的 IBM 独特优势。这包括***软件工程***（BigInsights 组件不仅是预集成的，而且已经过广泛测试，具有最优的性能和功能透明性）、***自定义***（IBM 服务组织提供行业和企业特定解决方案的能力已成为 BigInsights 号召力的主要源泉）和***客户支持***方面的长期公司优势。

但是，IBM 在关系技术和数据仓库方面拥有数十年经验。该公司的***SQL 优势***超越了任何其他 Hadoop 的经销商（与其拉开了显著的差距），而且它的***系统集成***功能是全球最出色的。

与其软件业务的其他领域一样，该公司正在雄心勃勃地招募和支持业务合作伙伴。这些合作伙伴目前包括 300 多家独立软件供应商 (ISV) 和服务公司，包括大量补充性工具和行业特定解决方案提供商。该数字目前正在快速增加。

毋庸置疑，Hadoop 开源社区将保持繁荣，免费下载产品的使用范围将继续扩大。但一个完全不同的企业解决方案类别显然将会出现，而且这些解决方案将更多地关注提高生产力、稳定性、弹性、可管理性、系统集成和深入的客户支持。

对于期望采用企业范型的组织，早部署 IBM BigInsights 比晚部署更有意义。

解决方案集

概述

在目前的版本中，BigInsights 包含 Apache Hadoop 和相关项目的主要组件，以及之前介绍的 IBM 增强。BigInsights 由 IBM 作为一款授权软件产品提供，也可通过 IBM SmartCloud Enterprise 和第三方云服务提供商提供。

除了旗舰级 ***Enterprise Edition***（它目前包含图 3 中总结的组件），IBM 还提供两个免费的 BigInsights 版本。

Basic Edition 包含主要的 BigInsights 开源组件，以及数据库和 Web 服务器接口，还有一个简单的管理控制台。***Quick Start Edition*** 是一个接近全功能的产品，仅限于非生产用途。其设计宗旨是让用户评估 BigInsights 企业特性并获得其使用经验，以及设计应用程序原型并开发概念验证。

部署选项

服务器和存储

IBM 提供了围绕 IBM System x3550 M4 和 x3630 双插槽 x86 服务器构建的 BigInsights 集群，这两种服务器分别用作管理和数据节点。数据节点可配置 Near Line SAS (NL-SAS) 或 SATA 驱动器。可以用低于 20、20 到 50 和 50 个以上的节点数对不同配置进行增量封装。支持 Red Hat Enterprise Linux (RHEL) 和 SUSE Linux Enterprise Server (SLES)。

也可将 BigInsights 部署在非 IBM x86 服务器和带 RHEL 或 SLES 的 IBM Power Systems 上。可部署 IBM 或第三方阵列作为外部存储。

Platform Symphony

IBM 提供了将 BigInsights 部署在 ***Platform Symphony*** 上的选项。借助此方法，Platform Symphony 作业调度和管理机制取代了 MapReduce 的相关机制，而且可以充分利用额外的高可用性特性。

Platform Symphony 采用基于 x86 的集群来支持应用程序，提供了极高的性能和可伸缩性水平。大体上来讲，目前支持配置 10,000 或更多个核心。依据 IBM 的数据，已证明基准测试在大规模社交媒体分析工作负载方面，其性能是开源 MapReduce 的 7 倍以上。

Platform Computing 已长时间用于科学和技术计算 HPC，以及针对金融服务、制造、数字媒体、石油和天然气、生命科学和其他行业中各种商业应用 HPC。

GPFS-FPO

BigInsights beta 模式的许多早期用户已部署了 ***GPFS-FPO***，而且已正式发布了 2.1 版。

在 HPC 应用程序中，GPFS 证明了在极大型配置（包含超过 1,000 个节点的安装很常见，最大超过 5,000 个节点）中可实现接近线性的可伸缩性。存储量常常达到数百 TB，而且已经有能正常工作的 PB 级系统。

用户体验，以及使用各种 HPC 基准测试工具进行的测试运行都证明，GPFS 的性能比 HDFC 高得多 - 在某些情况下达到 20 倍以上。

GPFS 还合并了一种分布式元数据结构、策略驱动的自动存储分层、托管的高速复制，以及信息生命周期管理 (ILM) 工具。

应用程序开发	
Social Data Analytics Accelerator	该应用程序套件用于提取社交媒体数据，构造用户概要，并与情绪、舆论、意图和所有权建立关联。其中包含了用于品牌管理、线索生成和其他常用功能的可自定义工具。预先集成了用于 IBM（Unica 以前）Campaign 和 CCI 解决方案的选项
Machine Data Analytics Accelerator	该应用程序套件可导入和聚合来自日志文件、仪表、传感器、读卡器和其他机器来源的结构化、半结构化和/或非结构化数据。在文本、基于分面和基于时间线的搜索、模式识别、根源分析、连锁分析和其他功能方面提供了帮助
BigSheets	可识别、集成和分析大量非结构化和/或结构化数据且类似于电子表格的工具。它合并了 IBM 开发的分析宏和模式识别技术。可针对各种用户需求进行高度自定义
Big SQL	该原生 SQL 查询引擎允许开发人员使用标准 SQL 语法和 Hadoop 优化的 SQL 扩展来查询 Hive、HBase 或分布式文件系统数据。允许管理员使用多个来源的数据填充 Big SQL 表。JDBC 和 ODBC 驱动程序支持许多现有的 SQL 查询工具
Web 应用程序目录	包含示例查询、数据导入和导出，以及专为“概念验证”应用程序部署而设计的测试工具
基础架构	
Avro	该数据序列化和远程过程调用 (RPC) 框架定义了 JSON 数据模式
HBase	该 NoSQL（非关系）数据库合并了基于行和列的表结构。基于 Google BigTable 技术
Hive	简化了大型 HDFS 数据集的数据提取、转换和加载 (ETL) 以及分析工作
Jaql	这种高级声明性查询和脚本语言具有基于 JSON 的数据模型和类似 SQL 的接口，用于处理结构化和非结构化数据。最初由 IBM 开发
Lucene	该文本搜索引擎库描述了作业图和这些图之间的关系
Oozie	针对 Hadoop 作业管理的工作流调度程序
Pig	这个分析大数据集的平台包含高级表达语言和评估程序的基础架构
MapReduce	用于 Hadoop 集群的并行编程模型
Hadoop Distributed File System (HDFS)	Hadoop 分布式文件系统支持基于 x86 NameNode (master) 和 DataNodes 构建的集群。与 MapReduce 紧密集成
BigIndex	实施基于 Hadoop 的索引作为原生 InfoSphere BigInsights 功能，支持更多复杂的功能，包括分布式索引和分面搜索
BigInsights Scheduler	这个 Hadoop Fair Scheduler 扩展实现了对 MapReduce 作业进行基于策略的调度
可拆分的压缩	这个 Apache Lempel--Ziv--Oberhumer (LZO) 算法的扩展实现允许使用压缩的数据在多个映射器上运行作业
增强的安全性	包含增强的身份验证、授权（角色）和审计功能。是 IBM InfoSphere Guardium 解决方案的接口
集成的安装程序	这个由 GUI 驱动的工具允许快速、自动地进行 BigInsights 集群配置、安装和维护工作。引导式的安装特性简化了管理员任务
自适应 MapReduce	该 Platform Symphony 技术加快了小型 MapReduce 作业的处理速度，支持更有效地执行混合的 Hadoop 工作负载
GPFS File Placement Optimizer (GPFS--FPO)	针对在 Hadoop 集群中的使用而优化的 IBM General Parallel File System 高性能分布式文件系统的扩展

图 3：*IBM InfoSphere BigInsights 组件*

数据源和连接性	
BoardReader	这个 BoardReader 搜索引擎接口支持查询访问、数据下载和导入 BigInsights 文件系统
Web Crawler	这个 IBM Web Crawler 应用程序的接口用于互联网数据收集和组织
Flume	方便了跨 Hadoop 集群的大量数据的聚合和集成
R	支持对使用 R 统计语言编写的应用程序进行集成
Sqoop	支持在 SQL 和 Hadoop 数据库之间导入和导出数据
JDBC	DBMS 的标准 Java 数据库连接接口
ODBC	DBMS 的标准开放数据库连接接口
MicroStrategy、SAS	广泛使用的第三方分析工具的接口
数据库接口	IBM DB2、Oracle Database、Microsoft SQL Server 和 Teradata Database 的接口
IBM 数据交换	支持与 IBM Cognos Business Intelligence、InfoSphere DataStage ETL 工具、InfoSphere Warehouse 数据仓库框架、Platform Symphony 网格中间件、IBM PureData System for Analytics、SPSS 统计建模和分析及 InfoSphere Streams 实时中间件解决方案交换 BigInsights 数据

图例： IBM
开源

图 3*（续）：IBM InfoSphere BigInsights 组件*

详细数据

综合剖析

本报告中使用的计算基于图 4 中给出的 6 种综合剖析。FTE 指全职当量数据库管理员数量。

医疗保健公司	金融服务公司	零售公司
应用程序		
医疗保险提供商 - 执行索赔分析，以提供高质量的治疗建议和成本/可盈利变量 80 TB 磁盘存储	多样化的零售银行 - 对社交媒体、通信和交易记录执行客户情绪分析，以提供忠诚度计划选项。数据仓库接口 130 TB 磁盘存储	对客户的在线和店内购买行为执行对比分析。来源包括 Web 日志、销售点终端和其他数据。对推销应用程序执行预测分析。数据仓库和决策支持接口 200 TB 磁盘存储
IBM INFOSPHERE BIGINSIGHTS FTE		
开发和部署（6 个月）：5.25 生产后期操作：2.95	开发和部署（8 个月）：7.5 生产后期操作：3.15	开发和部署（12 个月）：11.3 生产后期操作：4.75
开源 FTE		
开发和部署（6 个月）：8.25 生产后期操作：4.3	开发和部署（10 个月）：13.15 生产后期操作：6.0	开发和部署（15 个月）：17.0 生产后期操作：8.5
媒体公司	**市场营销服务公司**	**电信公司**
应用程序		
分析 Web 日志中的多个属性，以确定使用模式、客户概况，跟踪广告事件活动并识别新的市场机会 300 TB 磁盘存储	分析客户电子邮件流量，以执行人口统计和情绪跟踪、营销活动管理和其他应用程序 350 TB 磁盘存储	分析呼叫细节记录 (CDR)、互联网和社交媒体活动，以识别交叉销售机会并提高忠诚度计划有效性。可连接到 CIS、数据仓库和操作系统 500 TB 磁盘存储
IBM INFOSPHERE BIGINSIGHTS FTE		
开发和部署（7 个月）：8.4 生产后期操作：3.25	开发和部署（6 个月）：7.55 生产后期操作：2.6	开发和部署（9 个月）：8.85 生产后期操作：3.0
开源 FTE		
开发和部署（9 个月）：12.35 生产后期操作：5.0	开发和部署（8 个月）：10.95 生产后期操作：4.5	开发和部署（12 个月）：14.05 生产后期操作：5.5

图 4：*综合剖析*

执行这些剖析时，利用了 14 家使用开源 Hadoop 的公司、使用 BigInsights 的相同数量公司和一家同时使用二者的公司所提供的数据。对于上述每个行业，都是根据规模大体相同，具有大体相似的业务概况和应用程序的公司来进行对比。这些公司位于美国（26 家）和欧洲（3 家）。

这些公司提供了应用程序的信息；这些应用程序的开发和部署时间；以及应用程序开发和部署 (1) 所需的 FTE 人员数以及 (2) 生产后期持续操作所需的 FTE 人员数。因为工作描述和职位因公司的不同而有所不同，所以等效职业的 FTE 数量在一些情况下是 International Technology Group 估算得到的。

成本计算

人员成本基于图 5 中所示的年薪假设，并针对 FTE 数量计算而来。为 BigInsights 和开源 Hadoop 工具的使用采用了相同的假设。

职业	薪资	职业	薪资
数据科学家 [1]	20 万美元	**开发人员** [1] [2]	13.2 万美
架构师/等同职位 [1]	18.9 万美	**数据专家** [1] [2]	14.7 万美
项目经理 [1]	15.4 万美	**安装专家** [1]	13.5 万美
首席开发人员 [1]	14 万美元	**系统管理员** [2]	10.4 万美

[1] 开发和部署
[2] 生产后期操作

图 5：*FTE 薪资假设*

这些计算基于适用时间段的 FTE 数量。举例而言，对于医疗保健公司，成本根据 6 个月的开发和部署 FTE 数量计算而来，而生产后期人员成本为 36 – 6 = 30 个月的计算结果。薪资增加了 55.48%，以包含福利、奖金和其他人均成本。

计算使用 BigInsights 的***软件成本***时，要根据 IBM 对图 6 中所示容量的磁盘存储的每 TB 定价。BigInsights 许可费用包含一年的免费软件维护 (SWMA)，而支持成本是两年的。计算考虑了用户报告的折扣。

成本分解

图 6 中给出了各种剖析的细分。

	公司类型					
	医疗保健	金融服务	零售	媒体	市场营销服务	电信
IBM INFOSPHERE BIGINSIGHTS						
许可和支持	**327.49**	**439.62**	**676.34**	**800.93**	**747.53**	**800.93**
人员						
开发和部署	596.46	1,125.68	2,526.16	1,097.07	838.97	1,477.74
持续操作	1,389.21	1,442.26	1,887.53	1,561.21	1,272.80	1,313.61
个人合计	**1,985.67**	**2,567.93**	**4,413.69**	**2,658.28**	**2,111.77**	**2,791.35**
总计（千美元）	**2,313.16**	**3,007.55**	**5,090.03**	**3,459.21**	**2,859.30**	**3,592.28**
开源 APACHE HADOOP						
许可和支持	**0**	**0**	**0**	**0**	**0**	**0**
人员						
开发和部署	914.57	2,471.03	4,725.62	2,068.84	1,611.65	3,024.47
持续操作	2,006.08	2,566.53	2,954.22	2,225.79	1,399.97	2,173.61
个人合计	**2,920.65**	**5,037.56**	**7,679.84**	**4,294.63**	**3,011.62**	**5,198.08**
总计（千美元）	**2,920.65**	**5,037.56**	**7,679.84**	**4,294.63**	**3,011.62**	**5,198.08**

图 6：*3 年成本分解*

关于 INTERNATIONAL TECHNOLOGY GROUP

ITG 让您深入知晓正在发生的情况和您的竞争优势

……这会影响您的未来发展和盈利前景

International Technology Group (ITG)（成立于 1983 年）是一家独立研究和管理咨询公司，专门研究信息技术 (IT) 投资战略、成本/收益指标、基础架构研究、部署战术、业务一致性和财务分析。

ITG 是开发总体拥有成本 (TCO) 和投资回报 (ROI) 流程与方法的早期创新者和先驱。2004 年，该公司获得了 Information Technology Financial Management Association (ITFMA) 颁发的 Decade of Education Award 大奖，ITFMA 是一家领先的专业协会，致力于最终用户 IT 企业中的金融管理实践教育和发展。

该公司承担过 120 多个主要咨询项目，向全球各个客户、用户组、行业会议和研讨会发布了 250 多篇管理报告和白皮书，以及 1,800 多份简报和演示文稿。

客户服务旨在提供真实数据和可靠的文档，从而在决策制定过程中提供帮助。所提供的信息为开发战术和战略计划打下了基础。该公司以最有效的方式分析重要的开发工作并提供实用指南，对可能影响复杂 IT 部署日程的变化进行响应。

该公司提供了广泛的服务，为客户提供必要的信息来为其内部能力和资源提供补充。自定义的客户计划涉及以下可交付成果的各种组合：

状态报告	重要问题的深入研究
管理简报	重大开发工作的详细分析
管理情况介绍	与管理层的定期互动会议
高层演示	为决策制定者安排的战略演示
电子邮件通信	及时回复信息请求
电话咨询	立即响应信息需求

该公司的客户包括各种私营和公共领域中典型的 IT 最终用户，他们代表着跨国企业、工业公司、金融机构、服务企业、教育机构、联邦和州政府机构，以及 IT 系统提供商、软件供应商和服务公司。联邦政府客户包括国防部内的机构（如 DISA - 国防信息系统局）、运输部内的机构（如 FAA -美国联邦航空局）以及财政部（如 US Mint - 美国铸币局）

International Technology Group
609 Pacific Avenue, Suite 102
Santa Cruz, California 95060-4406
电话：831-427-9260
电子邮件：Contact@ITGforInfo.com
网站：ITGforInfo.com