IBM 软件部
思想领导力白皮书

2013年12月

利用IBM InfoSphere BigInsights 加速实现大数据价值

目录

IBM InfoSphere BigInsights 2.1概述

IBM® InfoSphere® BigInsights™ 2.1是一款基于Apache Hadoop的、与硬件无关的软件平台，它可以提供使用多样化和大规模数据采集的全新方式。本白皮书描述了InfoSphere BigInsights 2.1最为常用的功能，允许组织经济高效地分析类型丰富多样和大量的数据，获得前所未有的洞察。[1]

InfoSphere BigInsights 2.1致力于提供企业所需要的功能，满足关键业务要求，同时维持Hadoop项目的兼容性。InfoSphere BigInsights 2.1包括各种各样的IBM技术，可以增强并扩展开源Hadoop软件的价值，更快获得回报，包括应用加速器、分析设备、开发工具、平台改进和企业软件集成在内。尽管InfoSphere BigInsights提供一系列丰富特性可以拓展Hadoop的各项功能，IBM仍然提供了另一种选择：您可以根据自身需求使用IBM Hadoop的扩展能力，无需强制结合InfoSphere BigInsights 2.1实现扩展。

为了帮助您快速启动大数据项目，InfoSphere BigInsights 2.1会提供大量的增强功能，包括一组通用开源和IBM的技术集合，它们可以分为以下类别：

- 挖掘Hadoop社区创新，加速部署
- 充分利用现有SQL技能和解决方案
- 支持用户驱动的分析和数据预配置
- 支持以人为本的信息发现和主题生成
- 充分利用动态和静态分析
- 集成常用的建模和预测分析解决方案

InfoSphere BigInsights：高度兼容的平台

第三方应用、合作伙伴解决方案和定制开发项目与以下InfoSphere BigInsights支持版本相兼容，可以平稳运行，无需更改更新数据地点。

- Apache Hadoop (1.1.1)，64位Linux版本、面向Java 6和Java的IBM SDK
- Avro (1.7.2)，数据序列化系统
- Chukwa (0.5.0)，数据采集系统，用于监控大型分布式文件系统
- Fair Scheduler，适用于作业提交的基础管理
- Flume (1.3.0)，一种分布式、可靠且高度可用的服务，适合高效移动集群的大量数据
- HBase (0.94.3)，一种非关系式的分布数据库，采用Java语言编写
- HCatalog (0.4.0)，一种适用于Hadoop的表格和存储管理服务
- Hive (0.9.0)，一种数据仓库基础架构，可以促进大数据集的数据提取、转化和加载(ETL)和分析，存储在Hadoop分布式文件系统(HDFS)之中
- IBM InfoSphere BigInsights Jaql，一种查询语言，专门面向JavaScript Object Notation (JSON)，主要用于分析大规模半结构化数据
- Lucene (3.3.0)，一种高性能、全功能的文本搜索引擎库，完全采用Java语言编写
- Oozie (3.2.0)，一种工作流协调管理器
- Orchestrator，一种先进的MapReduce作业控制系统，采JSON格式描述作业图表和它们之间的关系
- Pig (0.10.0)，一种分析大型数据集的平台，由高级语言（可以表达数据分析程序）和评估这些程序的基础架构组成
- Sqoop (1.4.2)，一种将信息由结构化数据库和相关Hadoop系统输入到Hadoop集群的工具
- ZooKeeper (3.4.5)，一种集中式服务，用于维护配置信息，提供分布式同步和群组服务

本白皮书介绍了这些增强功能如何有助于扩展开源Hadoop功能的价值，从而让组织经济高效地支持新兴的大数据工作负载。

挖掘Hadoop社区创新，加速部署

IBM在InfoSphere BigInsights 2.1之中投入的Hadoop开源软件组件承诺有助于提高第三方互操作性，可以支持新特性和新功能的持续开发。无论版本水平是否全部兼容且目录结构是否实现镜像，包含现有MapReduce、Hive、Pig和Sqoop项目的组织可以充分利用InfoSphere BigInsights 2.1的运行优势。

充分利用现有SQL技能和解决方案

旧有应用取决于SQL访问存储数据的能力，SQL是查询结构化数据的规定语言；因此，大多数组织具备深入和丰富的SQL技能。IBM客户一直在寻找各种通过Hadoop充分发挥其SQL技能优势的各种方式，从而降低运行Hadoop的入门障碍，并且提高与现有面向SQL的工具和应用的互操作性。IBM正在帮助客户完成这一任务，推出IBM Big SQL，这是一款针对Hadoop的数据仓库系统，用于归纳、查询和分析存储在InfoSphere BigInsights 2.1之中的数据。

Big SQL利用JDBC或ODBC驱动器访问存储在InfoSphere BigInsights之中的数据，与用户通过企业应用访问数据库的方式完全相同。您可以利用Big SQL服务器执行标准SQL查询，并且同时执行多个查询（参见图1）。

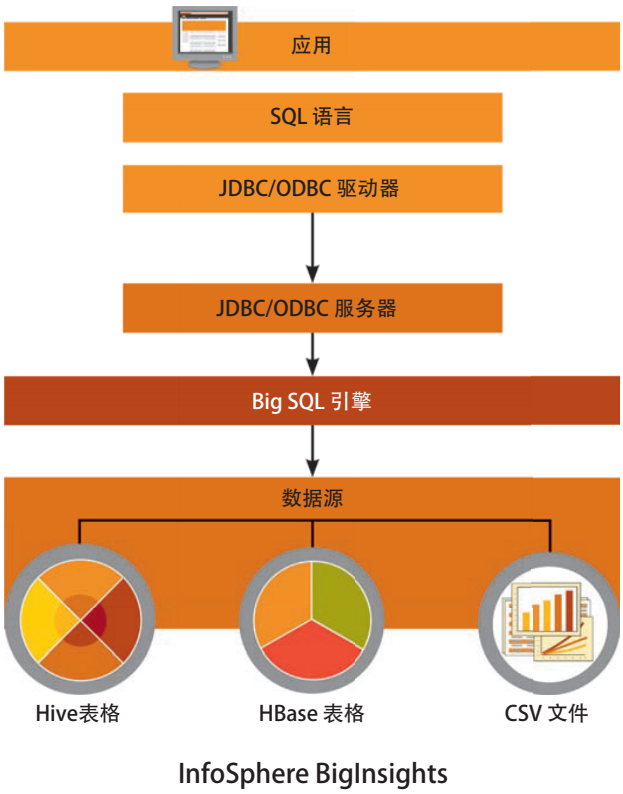

图1. IBM Big SQL概况。

Big SQL采用MapReduce并行和点查询，它们都属于快速返回信息的低延迟查询，可以降低响应时间并提高数据访问速度，能够支持大规模的特别查询。Big SQL服务器为多线程，因此可扩展性仅受运行在服务器内的计算机CPU的性能和数量限制，您可以提高Big SQL运行的服务器计算机的硬件性能，或者将多台Big SQL服务器链接在一起，提高吞吐量。

Big SQL可以改善现有的SQL技能，即时提高生产效率，因此将项目时间降到最小程度，并且降低指定项目的所需资金。凭借Big SQL，所有数据都可由SQL访问，允许您选择适合应用的存储格式。

支持用户驱动的分析和数据预配置

为了获得新洞察并提高业务成果，您需要最为理想的环境来探索并发现数据的关联关系。采用适合的技术，您可以引入全新类型的数据并推动全新类型的分析，从而扩展数据仓库的价值。InfoSphere BigInsights最为常用的部署模式之一便是数据探索区(Exploration Zone)。数据探索区可以为您的环境提供所需的各种功能，从而分析原始格式的信息——无论是结构化数据或非结构化数据——可以使用各种工具，例如文本分析、数据挖掘、实体分析和机器学习。您可以使用这一区域的数据用于探索分析，或者将数据发送到数据仓库，以便深入分析，为您的工作和数据分析提供更高的灵活性。

为了尽可能快地交付正确信息，支持这些系统的数据仓库必须经过优化，实现分析性能和运行查询吞吐量的恰当平衡。InfoSphere BigInsights 2.1可以提供多种用户驱动的功能，通过原始数据集创建新的数据集合，扩展传统关系数据源的现有数据集，并且针对数据执行特别分析，无需IT部门的协助。

数据可以经过清理、转换、整合并且聚合，以便准备进一步利用。一旦准备妥当，数据可以发布到通用架构（例如Hive），或者分阶段为外部解决方案（例如IBM IBM PureData™ System for Analytics）所使用。

IBM BigSheets

IBM BigSheets是一款基于浏览器的分析工具，可以将大量数据分解为可消耗、特定场景的商业背景。可以通过InfoSphere BigInsights控制台轻松访问，BigSheets能够采集多种来源的数据，包括网页爬取、社交媒体数据采集和分析、机器数据处理和分析、特别查询和更多方式（参见图2）。BigSheets还可用于其它工具的数据加载，这些包括Flume或IBM InfoSphere Information Server。

图2. IBM BigSheets概况。

一旦InfoSphere BigInsights采集数据，BigSheets用户便将相关数据加载到一个主工作簿。由此，BigSheets允许您基于主工作簿在工作簿中构建工作表（与电子表格相似），对数据进行格式化和探索操作。您可以将不同工作簿的各列相互组合，运行公式并且过滤数据。这些操作构成了分析的基础任务。

BigSheets可以生成并且执行必要的代码，自动运行所有数据操作工作，允许您在可视模式下运行，无需在低级脚本或Java水平下操作。您还可以将数据与InfoSphere BigInsights的文本分析功能相互组合，过滤并操作数据，并且深入挖掘信息，从原始数据中提炼极具价值的洞察。

在提炼数据并运行分析之后，您可以应用可视化功能，例如标签云、柱状图、地图和饼形图。这些可视化功能可提供可易用的数据输出，凸显出各种关系，从以前互不连贯的数据中提取洞察。

InfoSphere BigInsights Web控制台

InfoSphere BigInsights 2.1中一种基于角色的Welcome页面可以动态配置每一名用户的数据集合与作业。该软件包括可以完成多种数据管理任务的应用。这些预安装应用包含的属性与与您创建的应用相同，可以作为大数据项目的起点。用户可以创建和共享开发的新作业，令Web控制台成为日益强大的入门工具，可与InfoSphere BigInsights共同运行。以下清单代表部分立即可用的应用：

- Ad hoc Hive Query：利用Ad hoc Hive Query应用，可以创建您的自定义Hive查询，分析您的数据。
- Ad hoc Jaql Query：利用Ad hoc Jaql Query应用，可以创建您的自定义Jaql查询，分析您的数据。
- Ad hoc Pig Query：利用Ad hoc Pig Query应用，可以创建您的自定义Pig查询，分析您的数据。
- Ad hoc R Script：该应用适合运行R脚本。因为Oozie可以分配R脚本，运行不常使用的集群节点，R脚本必须安装在您集群的所有节点中。R脚本读取并定局入到本地文件中，而不是HDFS目录。因此，Ad hoc R Script应用可以将输入文件拷贝到所需的本地目录，并且将输出文件移动至HDFS目录。
- BoardReader：BoardReader应用可以搜索、定位并显示多个网络来源的信息，例如在线论坛、消息板、博客、新闻源和视频。
- 数据下载：该应用适合开发者下载IBM developerWorks®来源的数据。当您同意developerWorks的条款和条件之后，您便可以选择数据集下拉清单中的采样数据集或者输入URL访问某些数据集。
- 数据采样：鉴于存在大量的数据集和参数，该应用可以产生具有代表性的数据样本。该应用使用统一的随机样本（不重复抽样），采样输入数据。该应用将结果输出到一个文件，文件格式与输入文件的格式相同。
- 数据子集：数据子集应用适合创建完整数据的子集。然后您可以按照结构、内容和格式分析数据子集，从而提高性能。
- 数据库导出：该应用可以将HDFS中文件的数据写入关系数据库管理系统中的表格，并使用Java程序将存储在HDFS中的数据导出到数据库的一个表格。输入数据存储在DFS的文件中。输出格式可以是CSV或JSON。

- 数据库导入：该应用可将关系数据库管理系统的数据加载到HDFS文件中。它使用Java程序将数据库的数据导入，将数据写入到HDFS的文件中。您可以指定SQL（选择）查询，定义由数据库导入的数据。通过数据库恢复的数据然后会写出到HDFS的文件中，文件格式为CSV或JSON。
- 分布式拷贝：利用MapReduce作业，您可以将远程来源的数据拷贝到HDFS或从HDFS拷贝到远程来源。利用分布式拷贝应用可以实现来源之间的文件和目录拷贝。
- HBase：HBase应用可以帮助您将HBase表格的各行数据导出到InfoSphere BigInsights控制台。您可以将HBase表格的数据导出为JSON文件。应用需要导出数据的各项参数；它不接受用户的HBase查询。
- Web Crawler：Web Crawler应用是一种自动程序，可以系统地跟踪互联网页面并采集数据。它还可以针对存储在InfoSphere BigInsights中的文件版本，对比文件的大小和内容。
- Web REST导入：该应用可以获取特定URL位置的内容，并且将内容存储在指定的HDFS目录中。

这些应用可以经过修改然后根据安全权限发布给特定用户或一类用户，为他们提供启动项目的多种方案。

分析加速器

IBM提供多种分析加速器，可以大幅降低大数据应用的回报时间。这些加速器为特定的使用案例提供商业逻辑、数据处理和可视化功能。通过使用加速器，您可以应用先进的分析，有助于集成并管理持续不断进入组织的各种速度和容量的数据。

加速器还可创建一种开发环境，构建新的自定义分析应用，为组织的具体需求量身定制。

以下两种加速器均配备了InfoSphere BigInsights：IBM Accelerator for Machine Data Analytics和IBM Accelerator for Social Data Analytics。以下两种加速器配备了InfoSphere Streams：IBM Accelerator for Social Data Analytics和IBM Accelerator for Telecommunications Event Data Analytics。这些加速器涵盖众多的通用使用案例，可以针对企业特定需求实现轻松扩展。

IBM Accelerator for Social Data Analytics

社交媒体论坛的数据包含有关用户偏好的高价值信息。然而，这种信息的访问和运行需要大规模的导入配置和分析功能。IBM Accelerator for Social Data Analytics可以深入理解社交媒体来源的运行方式、提取推特网站、论坛和博客的相关信息，然后根据和探索使用案例和行业打造用户的社交档案。典型的工作流由以下部分组成：导入数据文件、然后配置、索引和分析数据。

利用IBM Accelerator for Social Data Analytics，您可以：

- 导入并分析社交媒体数据，识别用户特征，例如性别、位置、名称和爱好等
- 通过消息和来源开发综合全面的用户档案
- 将信息档案与关于品牌、产品和公司的情绪、抱怨、意图和所有权表达联系在一起

IBM Accelerator for Social Data Analytics通常用于采集数据，增强客户分析能力——与众多社交监听工具不同，它们只能针对已知客户识别社交活动。

IBM Accelerator for Machine Data Analytics

IBM Accelerator for Machine Data Analytics可以吸收、解析并且提取众多来源的各种大量机器数据（例如机器数据文件、日志文件、智能设备和遥测技术），并且有助于在数分钟（无需数天或数周）内处理这些数据。它帮助组织获得关于运营、客户体验、事务处理和行为的洞察，可以识别基础架构问题和客户偏好的变化。或者驱动系统交互的陷阱事件。众多IBM客户使用IBM Accelerator for Machine Data Analytics主动提升运行效率，排除故障问题，调查安全事件，并且监控端到端基础架构，以避免服务降级或中断。

一个典型的引导工作流由以下部分组成：组织并导入成批数量，然后提取、索引、搜索、转换并分析数据。利用IBM Accelerator for Machine Data Analytics，您可以：

- 基于文本搜索、分面搜索或时间线搜索，在多个机器数据条目内部和之间进行搜索，寻找事件
- 添加或提取日志类型到现有储存库，从而丰富机器数据的背景
- 链接并关联系统之间的事件
- 揭示各种模式

InfoSphere BigInsights Text Analytics

InfoSphere BigInsights Text Analytics是一款功能强大和说明式信息提取系统，十分善于通过文本输入创建结构化信息，允许用户从底层文本数据中获得切实有效的洞察。InfoSphere BigInsights Text Analytics模块可以经过自定义设计，充分发挥面向Hadoop的处理模型优势。与传统的文本分析方法相比，它的速度极快，能够迅速处理大量的非结构化信息。InfoSphere BigInsights Text Analytics模块还具有说明性，这意味着它可以采用类似SQL的方法，随时适应您特定的分析需求，这对传统的文本工具而言无法实现。这有助于降低成本，提供舒适度，这对Apache Hadoop领域是独一无二的优势。

Text Analytics包括在Eclipse开发环境和InfoSphere BigInsights Text Analytics Workf的低级部分之中。Text Analytics Eclipse开发工具可以用于开发和测试Eclipse的提取器。一旦您已经识别并选择了所用的提取器，便可将它发布InfoSphere BigInsights控制台中作为一种应用，管理员可以部署并且由InfoSphere BigInsights的用户消耗使用。

InfoSphere BigInsights控制台的Welcome页面包括如何支持Eclipse环境、从而使用InfoSphere BigInsights开发应用的信息。一旦发布了提取器，应用部署到了InfoSphere BigInsights控制台，然后它便可作为BigSheets功能或工作流的一部分运行起来。

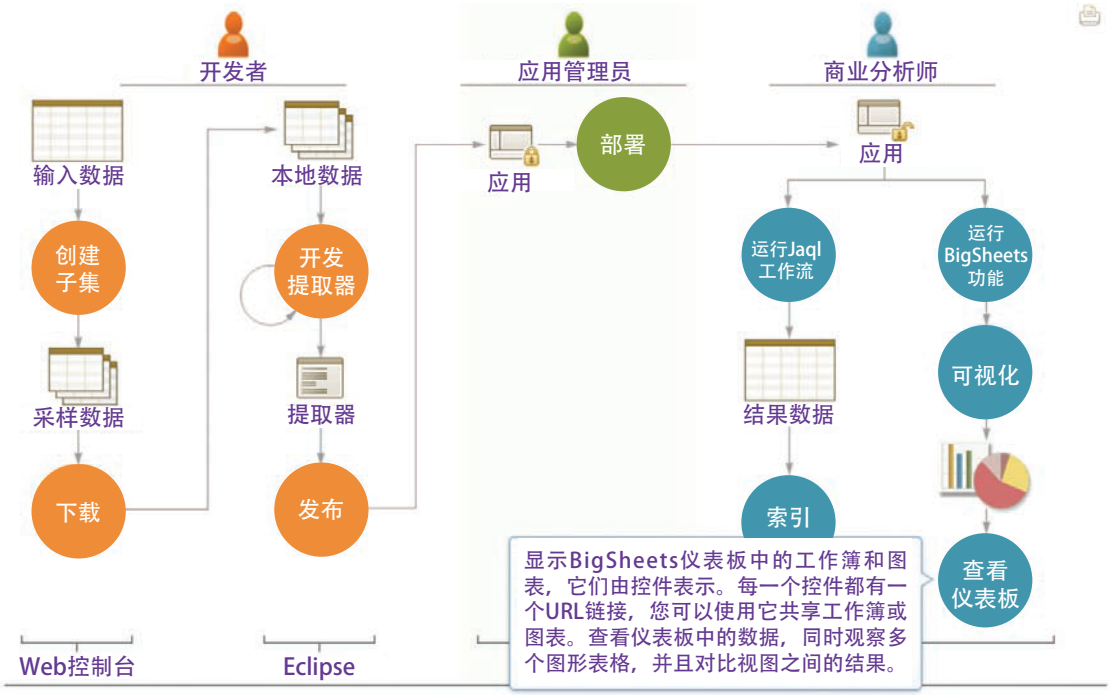

图3. InfoSphere BigInsights Text Analytics工作流概述。

由Text Analytics应用得到的结果可以导出到BigSheets、仪表板和其它InfoSphere BigInsights组件以供进一步分析（参见图3）。

IBM InfoSphere Data Explorer

数据探索任务的关键部分，支持以人为本的探索并且迅速理解掌握的信息。它可以支持最终用户驱动创建相关主题并自动发现相关信息，并且让用户快速构建和部署交互网络应用，它们通常用于客户洞察和客户服务环境。InfoSphere Data Explorer Engine服务器可以实时接收InfoSphere BigInsights服务器集群或InfoSphere Streams服务器集群的数据。InfoSphere Data Explorer还可将相关数据推关给信息应用用户，支持联合访问其它IBM产品。

充分利用动态和静态分析

越来越多的组织在部署分析和应用，它们横跨动态（或实时）和静止的用例。这些横跨动态和静止用例的新一代洞察需要数据和分析在两个类型的环境之间来回流动。InfoSphere BigInsights 2.1可以充分利用IBM InfoSphere Streams，加速动态和静止信息之间的流动。

InfoSphere Streams是一款高性能计算平台，支持用户开发的应用迅速吸收、分析并关联数千个实时来源的信息。对于流数据，InfoSphere Streams可以持续分析海量数据，具有极低的延迟，支持您快速应对各种发生的趋势和事件。程序员可以利用InfoSphere Streams按照InfoSphere BigInsights的需要写入数据，从而不断进行深入的趋势分析。这一分析的结果可以采集并且反馈到InfoSphere Streams，从而细微调整应用逻辑和操作。为了降低部署时间和成本，InfoSphere Streams应用可以通过本机映射到分布式InfoSphere BigInsights存储。

集成常用的建模和预测分析解决方案

IBM客户一直在寻找各种便捷的方式，以便利用丰富多样的预测建模解决方案与基于Hadoop的发现区相互结合，快速获得回报价值，并且提高现有解决方案和技能的利用率。InfoSphere BigInsights可以支持最为常用的建模和分析包（包括SAS、IBM SPSS®和R. Users类似工具），这些建模环境可以使用InfoSphere BigInsights的数据进行分析，从而满足这一要求。这支持用户在熟悉的环境中持续运行，同时可以访问前所未有、更加丰富的新环境。

每一个分析包在InfoSphere BigInsights环境中都有不同等级的Hadoop支持。SAS和其它环境主要使用BigInsights作为数据源环境，允许您利用Hive等这种架构中的信息。有些分析包（例如SPSS Catalyst for InfoSphere BigInsights,）支持模型开发和执行，可以直接在InfoSphere BigInsights平台中完成任务。

SPSS Catalyst有助于实现数据准备自动化，采用交互可视的方式自动阐释结果并且展示分析过程，提供清晰简明的概要，可以提高分析生产效率，并且缩短回报时间。SPSS Analytic Catalyst结合InfoSphere BigInsights运行，可以采用复杂算法支持自动关键驱动要素识别、自动测试和基于回归的技术。SPSS Analytic Catalyst还可以提供预测分析结果的交互可视和通俗归纳，清晰显示洞察，包括各种解释和统计详细信息。

结论

InfoSphere BigInsights 2.1提供一组独有的功能，它们将Apache Hadoop生态系统的创新功能、传统技能的强大支持以及安装完毕的各种工具相互结合于一体。通过开源功能充分利用现有技能和工具，有助于降低总拥有成本，并且加速获得回报价值。

备注

欲了解更多详细信息

欲了解关于InfoSphere BigInsights和InfoSphere BigInsights for Hadoop Quick Start Edition的更多详细信息，敬请联系您的IBM代表处或IBM商业合作伙伴，或者访问以下网站：

- **ibm.com**/software/data/infosphere/biginsights
- **ibm.com**/infosphere/quickstart

作者简介

Tom Deutsch (@thomasdeutsch)是一名IBM大数据团队的项目总监。他在将Hadoop技术由IBM研究院过渡到IBM软件部的过程中起到了重要作用，他继续从事IBM研究院大数据活动和IBM研究院向商业产品的过渡工作。Deutsch创建了InfoSphere BigInsights基于Hadoop的产品，历时数年帮助客户研发Hadoop、InfoSphere BigInsights和InfoSphere Streams等技术，包括与200多名客户共同合作，确认架构适宜性、开发商业战略和管理初期阶段的项目。在业界拥有20多年的丰富经验以及两家创业公司的资深人士，Deutsch如今已经成为一位技术、战略和商业信息管理领域的企业专家。

国际商业机器中国有限公司
北京市朝阳区北四环中路27号
盘古大观写字楼
邮编：100101

在中国印刷
2014年9月

IMW14684-USEN-01