IBM SPSS Analytic Server V3.1.1

概述

注释

在使用本信息及其支持的产品之前，请先阅读第 5 页的『声明』中的信息。

产品信息

此版本是用于 IBM SPSS Analytic Server 的 V3.1.1 以及后续发行版和修订版，直至在新版本中另有说明为止。

目录

概述

IBM® SPSS® Analytic Server 是一套用于大型数据分析的解决方案，融合了带有大型数据系统的 IBM SPSS 技术，允许您使用熟悉的 IBM SPSS 用户界面来解决问题，而这是以前无法做到的。

为什么大型数据分析至关重要

各组织所收集的数据量正成倍增长；例如，金融和零售业保留着一年（或两年、甚至十年）内的所有客户交易数据，电信服务提供商保留着呼叫数据记录 （CDR) 和设备传感器读数，互联网公司保留着网络检索结果。

在以下情况下都需要大型数据分析：

- 存在大量数据（TB、PB、EB），尤其是结构化和非结构化数据混合时
- 存在迅速变化/累积的数据

在以下情况下同样存在大型数据分析：

- 构建大量（数以千计）模型
- 频繁构建/刷新模型

挑战

收集大量数据的相同组织在利用这些数据时实际上经常都会遇到困难，有以下原因：

- 传统分析产品的体系结构不适用于分布式计算，并且
- 现有的统计算法并不是设计来处理大型数据的（这些算法期望数据向它们移动，但是移动大型数据成本过高），因此
- 执行大型数据分析的当前发展状况亟需大型数据系统的新技能和深入的认识。 很少有分析人员具有这些技能。
- 内存解决方案能够解决中型问题，但对于真正的大型数据，其扩展性不是很好。

解决方案

分析服务器 提供以下功能：

- 可重用大型数据系统的以数据为中心的体系结构，例如，数据位于 HDFS 的 Hadoop Map/Reduce。
- 定义的接口可合并新的统计算法（设计为流向数据）。
- 熟悉的 IBM SPSS 用户接口隐藏了大型数据环境的详细信息，以便分析人员能够集中于分析数据。
- 可随意扩展以解决任何大小的问题的解决方案。

体系结构

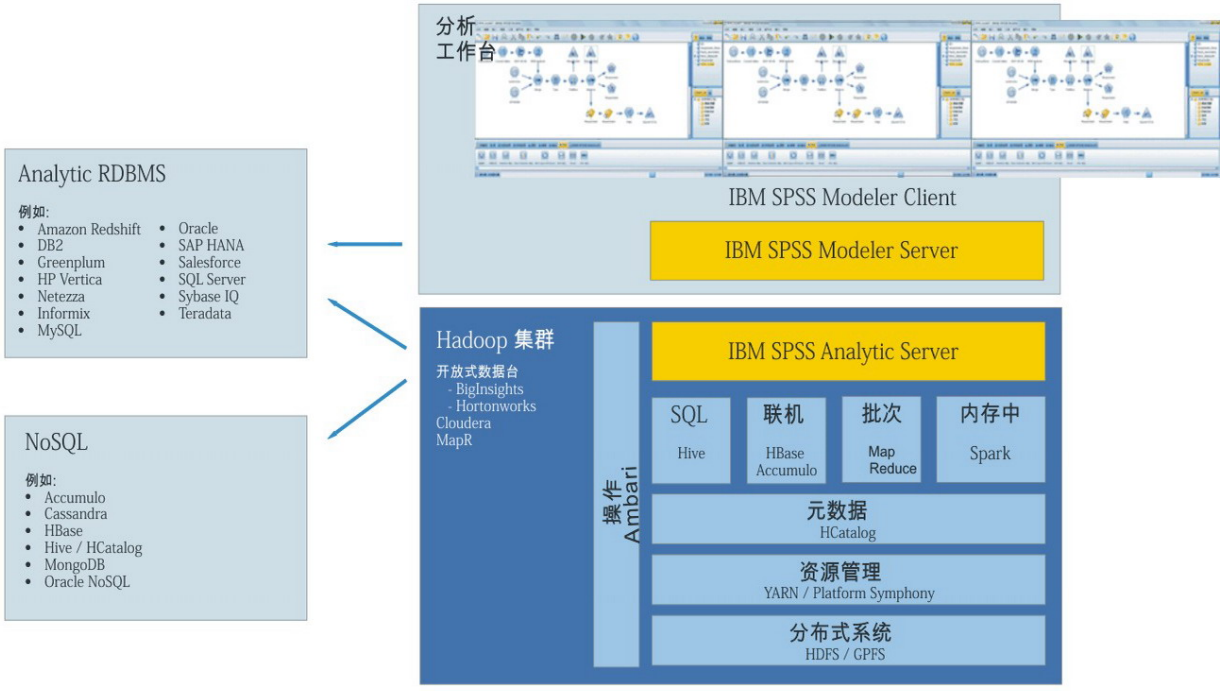

图 *1.* 体系结构

分析服务器 位于客户机应用程序和 Hadoop 云之间。 假设数据存在于云中，那么使用 分析服务器 的方法概述如下：

1. 在云中的数据上定义 分析服务器 数据源。
2. 定义要在客户机应用程序中执行的分析。 对于当前发行版，客户机应用程序是 IBM SPSS Modeler。
3. 当您运行分析时，客户机应用程序将提交一个 分析服务器 执行请求。
4. 分析服务器 会对该作业进行编排以在 Hadoop 云中运行，然后将结果报告给客户机应用程序。
5. 您可以使用这些结果来定义更进一步的分析，如此重复循环。

Spark 和 分析服务器

分析服务器 与 Apache Spark 集成以提高性能。

何时使用和不使用 Spark

如果在 Hadoop 集群中将 Spark 作为 Ambari 服务进行安装，那么 分析服务器 使用它来处理大数据作业。以下准则适用于确定何时不使用 Spark。

1. 如果数据集小于 128 MB，那么 分析服务器 在 分析服务器 JVM 中使用嵌入式 MapReduce 函数并且不使用 Spark 或 Hadoop 集群。
2. 如果未在集群上安装 Spark，那么 分析服务器 使用 MapReduce v2。

3. 分析服务器 使用 MapReduce v2 来构建 PSM 模型。 作业以 PSM 模型构建结束时，分析服务器 使用 Spark 通过导致模型构建的所有步骤来处理作业，并写入磁盘，然后使用 MapReduce 来构建 PSM 模型。例如，如果作业包含后面跟着 PSM 模型构建的连接，那么在 Spark 中运行连接并在 MapReduce 中的连接数据上运行 PSM。

如何使用 Spark

分析服务器 服务启动并发现 Spark 可用后，它将对"Spark Hadoop 作业"进行初始化，该作业允许在集群中的分布式任务之间进行通信。 只要 分析服务器 服务运行，此作业都会运行，并用于所有 分析服务器 执行。此方法可提高与编排多个 MapReduce Hadoop 作业有关的性能，因为它将除去重新装入每个 Hadoop 作业的所有 分析服务器 组件所产生的开销。

Spark 可运行 MapReduce 作业。 这将使 分析服务器 尽可能使用"本机"Spark 算法，如连接、排序和并集。同时，分析服务器 可运行 Spark 中的现有 SPSS Map and Reduce 算法，而不直接使用 Hadoop API。

V3.1.1 中的新增内容

V3.1.1

平台

- 支持 Cloudera 5.11 和 5.12
- 支持 Ubuntu Linux 16.04（含 Hortonworks 数据平台 2.6 和 Cloudera 5.11）
- 不再支持 Cloudera 5.8 和 5.9
- 不再支持 Big Insights 4.1、4.2 和 4.2.5
- 不再支持 MapR 5.0

数据源

- 支持 Apache Hive 2.1
- 不再支持 MongoDB 2.6
- 不再支持 MySQL 5.1

性能增强

- HDP 现在可以使用更加自动化的脱机安装过程。有关更多信息，请参阅脱机安装
- 多集群支持。 多集群功能是 IBM SPSS Analytic Server 的高可用性功能的增强功能，它可在多租户环境中提供更强的隔离。 缺省情况下，安装 分析服务器 服务（在 Ambari 或 ClouderaManager 中）会导致系统定义一个分析服务器集群。
- 现在可以为每个 分析服务器 租户配置单独的 YARN 队列。有关更多信息，请参阅为每个分析服务器租户配置单独的 YARN 队列 - HDP 或为每个分析服务器租户配置单独的 YARN 队列 - Cloudera。
- 现在支持对 SQL 回推采样。
- 现在提供了以下 Spark ML 增强功能：
  - 支持二等分 K 均值
  - 支持 XGBoost
  - 支持保序回归
- Spark RDD 数据源交叉现在可以在不同的 ASL 作业中共享。

- IBM SPSS Analytic Server最佳实践和建议部分已更新为包含特定于 Cloudera 的信息。

有关最新的系统需求信息，请使用 IBM 技术支持站点上的详细系统需求报告：http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html。 在此页面上：

1. 输入 SPSS 分析服务器 作为产品名称并单击**搜索**。
2. 选择想要的版本和报告范围，然后单击**提交**。

声明

本信息是为在美国国内供应的产品和服务而编写的。可以从 IBM 获取本资料的其他语言版本。 但是，您必须拥有该语言的产品或产品版本副本，才能访问对应语言的资料。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。 有关您所在区域当前可获得的产品和服务的信息，请向您当地的 IBM 代表咨询。 任何对 IBM 产品、程序或服务的引用并非意在明示或默示只能使用 IBM 的产品、程序或服务。 只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。 但是，评估和验证任何非 IBM 产品、程序或服务的操作，由用户自行负责。

IBM 可能已拥有或正在申请与本文档内容有关的各项专利。 提供本文档并不意味着授予用户使用这些专利的任何许可。 您可以用书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

International Business Machines Corporation"按现状"提供本出版物，不附有任何种类的（无论是明示的还是默示的）保证，包括但不限于默示的有关非侵权、适销和适用于某种特定用途的保证。 某些管辖区域在某些交易中不允许免除明示或默示的保证。因此本条款可能不适用于您。

本信息可能包含技术方面不够准确的地方或印刷错误。本信息将定期更改；这些更改将编入本信息的新版本中。 IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。 那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的：(i) 使其能够在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及 (ii) 使其能够对已经交换的信息进行相互使用，请与下列地址联系：

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本文档中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可协议或任何同等协议中的条款提供。

此处引用的性能数据和客户示例仅用于描述目的。 实际性能可能因特定配置和操作条件而异。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。 IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。 有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

所有 IBM 的价格均是 IBM 当前的建议零售价，可随时更改而不另行通知。 经销商的价格可与此不同。

本信息仅用于规划的目的。 在所描述的产品上市之前，此处的信息会有更改。

本信息包含在日常业务操作中使用的数据和报告的示例。 为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。 所有这些名称都是虚构的，若实际人员或企业与此相似，纯属巧合。

版权许可证：

本信息包含在日常业务操作中使用的数据和报告的示例。 为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。 所有这些名称都是虚构的，若实际人员或企业与此相似，纯属巧合。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明：

©（贵公司的名称）（年）。 此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

© Copyright IBM Corp. （输入年份）。 All rights reserved.

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域的商标或注册商标。 其他产品和服务名称可能是 IBM 或其他公司的商标。 最新的 IBM 商标列表可以在 Web 上的 "Copyright and trademark information"中获取，地址为：www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 徽标、PostScript 以及 PostScript 徽标是 Adobe Systems Incorporated 在美国和／或其他国家或地区的注册商标或商标。

IT Infrastructure Library 是 Central Computer and Telecommunications Agency 的注册商标，该企业现已成为 Office of Government Commerce 的一部分。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、 Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和@3B72其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 以及 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

ITIL 是一个注册商标，是 Minister for the Cabinet Office 的共同体注册商标，并且已在 U.S. Patent and Trademark Office 进行注册。

UNIX 是 The Open Group 在美国和／或其他国家或地区的注册商标。

Cell Broadband Engine 是 of Sony Computer Entertainment, Inc. 在美国和/或其他国家或地区的商标并且在当地许可证下使用。

Linear Tape-Open、LTO、LTO 徽标、Ultrium 和 Ultrium 徽标是 HP、IBM Corp 和 Quantum 在美国和其他国家或地区的商标。

Printed in China