IBM DataWorks

简便、强大、集成式的云端数据准备与迁移

亮点

- 让所有技术用户和非技术用户快速、轻松地从数据中攫取价值。
- 在云端提供简便的数据准备与迁移服务，确保数据质量。
- 与领先的云数据服务相集成，打造无缝的数据管理平台。

随着云技术、大数据、物联网时代的到来，企业面临着信息过载带来的挑战。目前，数据的创建量和收集量与日俱增，这让商业智能团队和数据科学团队并没有充足时间或资源来分析这些数据。事实上，Forrester 通过研究发现，68% 的简单 BI 请求会耗费 IT 部门数周、数月，甚至更长时间来进行处理。[1]

为确保满足新数据需求，保持自身的竞争地位，企业必须寻求多种方法将业务线专业人士转变为技能娴熟的数据工作者，以便减轻 IT 部门的工作负担。不过，期间也会出现一些问题：这意味着需要为业务用户授予多种方法，使其能够对多个来源（不论是本地还是云端）的数据进行快速地排序、准备和分析，而不需要数据库管理员或数据科学家来提供深层面的技术专业指导。

然而，借助一些新推出的云服务，比如 IBM® DataWorks，技术业务用户和非技术业务用户只需点击访问便可从数据中攫取有益的洞察力，不论对象是存储在本地的 Excel 表，还是托管于云端的大型数据库。

IBM DataWorks 简介

DataWorks 是一种全运维管理数据准备及迁移服务，能够让分析师、开发人员、数据科学家和数据工程师通过简单而强大的云界面来使用数据。作为 IBM Cloud Data Services 组合的关键组成部分，DataWorks 让业务分析师或“高级 Excel”用户能够借助应用开发及分析用例的支持，对数据进行挖掘、清理、标准化、转换及迁移等操作。

DataWorks 是一款无缝数据处理工具，与多种云数据服务相集成，包括 IBM dashDB™ 云数据仓库、IBM Cloudant® NoSQL 数据库、IBM Watson™ Analytics 等，可用于准备本地及外部的数据，并将其迁移到分析云生态系统，进而对数据进行快速分析及可视化操作。此外，DataWorks 也受持续交付支持，这在常规基础上为此产品新增了更多更强大的功能。DataWorks 的处理引擎基于 Apache Spark™ 构建，Spark 作为领先的开源分析项目，拥有不断发展的大型开发社区。通过结合二者的优势，我们最终得到了一款最佳组合型解决方案，能够紧跟大数据及云计算迅速迈进的创新步伐。

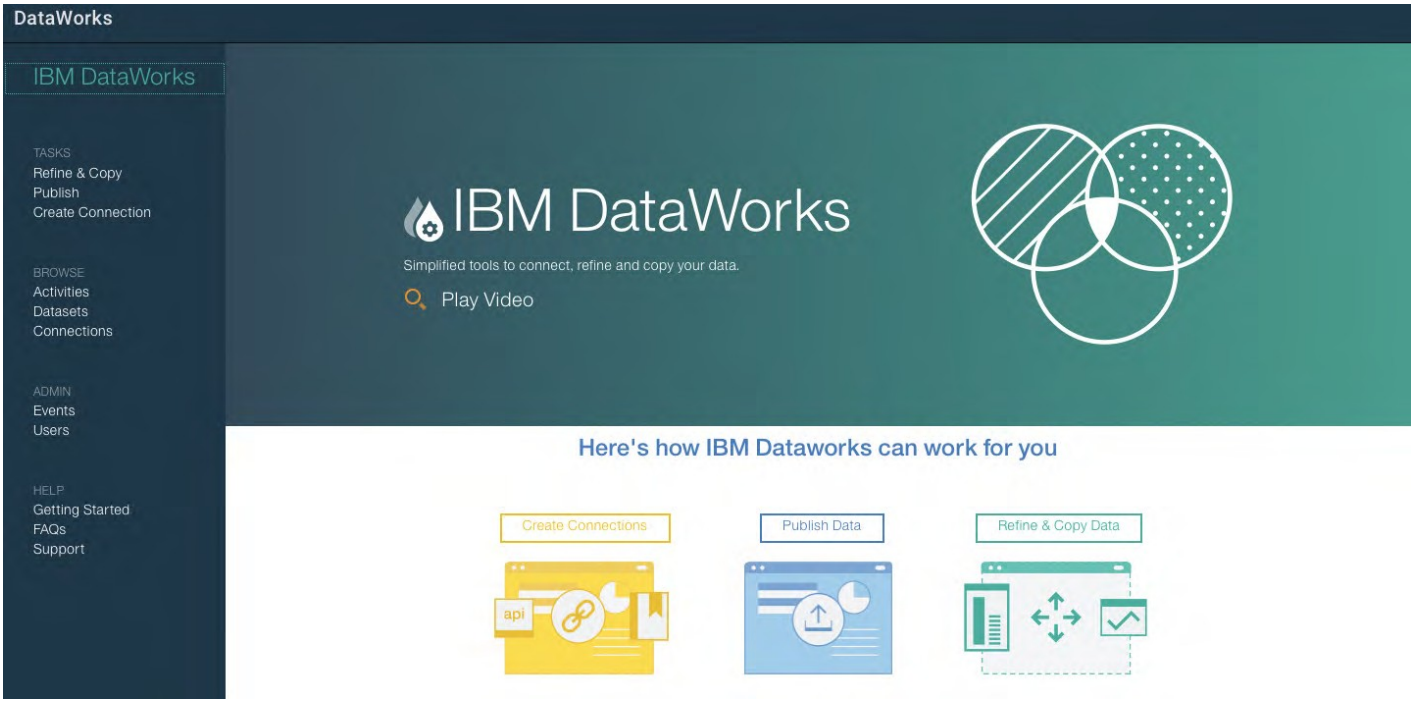

图 1：IBM DataWorks：全托管于云端的点击型数据准备即服务

实现整个企业内数据访问的普及化

要说人人都可以成为数据科学家可能为时尚早，但 DataWorks 等工具的出现却在不断推进数据访问及高级分析的普及化。企业运用 DataWorks 的方式多种多样，但主要用例包括：

1. 融合多个来源的数据：访问任何受支持的来源提供的数据，然后整合这些数据，以创建与目标分析任务相关的文件/数据表。
   - 例如：数字版权公司的数据科学家希望分析客户的媒体资产组合，以及 Nielsen、Rovi、Twitter、Rotten Tomatoes、EIDR 等来源提供的第三方收视率数据，并根据分析结果开发一些适于广告投放的算法。他们可以将 dashDB 用作媒体资产仓库，将多结构化内容存储于 Cloudant，同时利用 DataWorks 将综合数据整理成型，以便即刻用于制定报表。
2. 访问混合云环境中的数据：不论数据的存储位置如何，均可通过连接最常用的行业数据源对其进行访问，同时也能够轻松、安全地访问防火墙下游的数据。
   - 例如：用户需要访问存储于云端的客户情感数据和本地数据库的营销活动数据，以便评估营销活动的实效性。借助 DataWorks，用户可以构建一个安全的渠道，用以检索防火墙下游的数据。
3. 整理原始数据，以供分析：筛选源数据的值和列，对数据进行排序，删除重复数据，并通过标准化评分了解数据质量。
   - 例如：业务分析师需要根据近一年的历史销售数据来制定销售情况预测报告。制定报告之前，他访问了本地销售数据库，但无法确定数据的质量和相关性。DataWorks 可为用户提供数据质量评分，并让用户通过数据预览目测判定所获得的数据是否合适。此外，DataWorks 还可提供筛选功能，用以筛除不符合要求的值。
4. 加载数据，以供分析：不受位置限制地访问已准备好的数据，并将其加载到云端的数据服务。
   - 例如：数据科学家需要将一些文件从本地数据仓库加载到 dashDB 云实例，以便为客户保留项目构建统计模型。DataWorks 让用户可以通过轻松的点击式访问，选择要移动的数据表和文件，然后再选择目标数据源。
5. 控制来自 Web 应用的数据工作流：使用 DataWorks API 可创建并控制来自应用的工作流活动。
   - 例如：根据物联网传感器及手机、社交平台等互动系统发出的事件，应用开发人员可以触发由业务分析师、数据科学家或 IT 管理人员创建的活动，进而使用 DataWorks API 对数据进行迁移、整理和转换操作。
6. 将关系数据/结构化数据映射到半结构化数据：将标准化的表格数据加载至 Cloudant NoSQL 存储空间。
   - 例如：开发人员需要将关系数据加载至 Cloudant，以便在 Web 应用中使用，同时也需要将标准化数据转换为按等级划分的 JSON 文档结构。DataWorks 能够以无缝方式指向关系数据源及相应的 NoSQL 目标 Cloudant，因而可将关系数据转换为 JSON 文档。

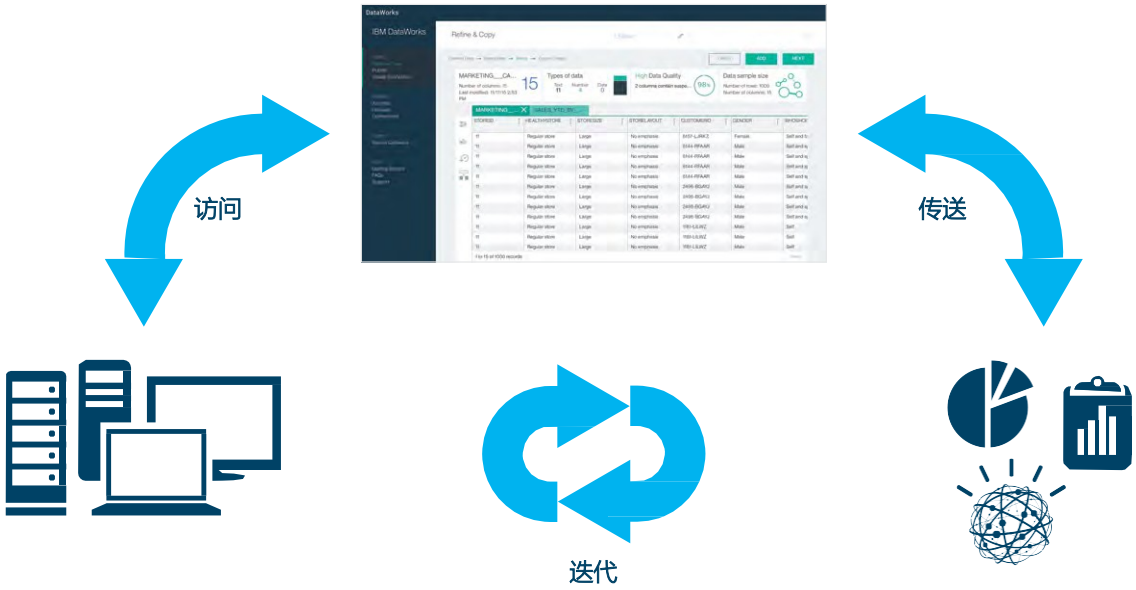

图 2：访问、整理数据并传送至分析云服务。然后，持续迭代。

在复杂的混合云环境中，实现简单的数据访问

随着 IT 环境所具备的"混合性"日趋复杂，当今企业在数据访问和数据迁移方面需要应对更为严峻的挑战。"混合"的定义可以很宽泛：对某些情况来说，它指的是本地基础架构与云服务之间的无缝式全同步；而对于其他情况，它也可以指不受数据存储位置的限制，为数据访问提供支持。尽管目前对混合云的定义不一，但不论是何种混合部署方案，不可避免地都会产生明显的业务挑战，包括如何快速、安全地实现数据迁移和访问等。

IBM DataWorks 可提供多种工具，帮助您安全、迅速地在混合云环境中实现数据访问和迁移。DataWorks 具备两大混合云支持功能：

1. 安全网关，为客户提供简便的解决方案，供其访问云端的企业数据。具体来说，我们只需要通过易于安装的 SSL 隧道，支持用户访问防火墙下游的数据即可。相比广义 VPN 访问，安全网关要简单许多，用户只需要满足一个要求：打开出站端口，安装本地代理。
2. 分析用户创建的数据准备流程并将尽可能多的操作推向源数据库，以便逐层优化，从而减小待传输的数据量。此功能可确保仅传输目标要求的数据，从而借用数据源的计算能力来分配工作负载，将更小的数据集迁移到云端。

DataWorks 贯穿混合云环境中的各个流程，通过提供本地和云端的安全集成点，确保更高的安全级别。

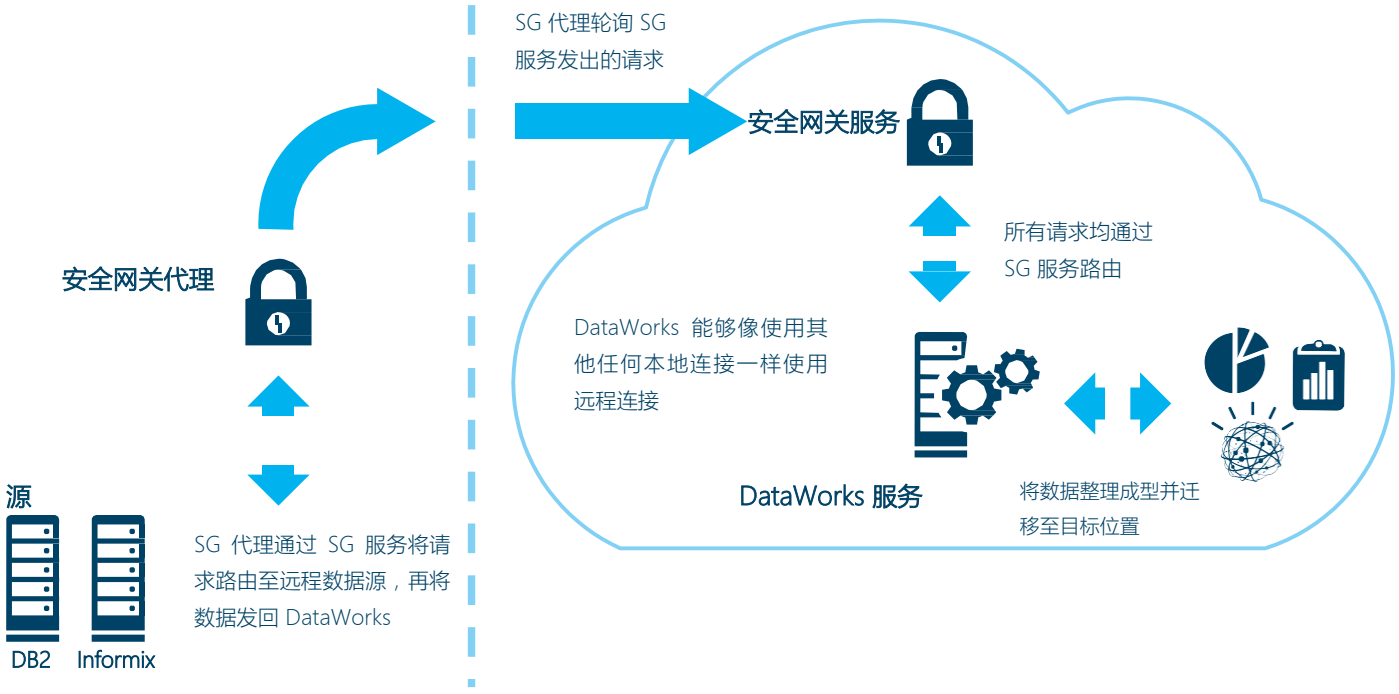

图 3：使用 DataWorks 安全网关安全地访问防火墙下游的数据

数据质量窘境：将数据整理成型

如今，有太多数据分析项目停滞不前、延期压后或半途而废，究其原因就是数据问题，比如数据不完整、不准确或不相关等。事实上，根据 Forrester 调研结果，竟有 42% 的业务专家在分析数据之前，要耗费近一半的工作时间（40% 以上）对手头的数据进行修复和验证。[2] 其本质问题就是数据质量，要应对这一挑战，我们可以采用一些新方法来完成数据准备。

数据准备指：通过实现广泛互联，以自助服务的方式安全地访问任何位置的数据。数据准备基于传统的 ETL（提取、转换、加载）概念，先优化数据质量和完整性，再通过分析数据来获得业务洞察力。字符串和整数属于技术用户掌控的领域；业务用户希望以尽可能简单的方式了解数据背后的含义。即便传统流程让业务用户长期依赖 IT 部门来为他们筛选数据集，但现如今的新数据需求却要求企业降低数据准备的门槛。业务用户必须能够自己完成数据准备工作。

为应对这一挑战，DataWorks 提供了易于使用的云端数据准备及数据迁移服务，不论是技术用户还是非技术用户，均可访问这些服务。数据准备技术异常复杂，对此我们仅有一个难以实现的终极承诺：让任何业务线用户无需掌握深厚的专业知识，就能扮演数据科学家的角色。促使我们开发 DataWorks 的强大动力是：让用户无需成为已获认证的数据科学家，就能运用 DataWorks 获取高级数据洞察力；Excel 高级用户能够掌控数据集，在更短的时间内制定出更完善的报告，而无需具备深厚的数据库及数据结构知识。

对于业务分析师和 Excel 高级用户，DataWorks 十分便捷的电子表格式界面，让用户能够轻松地将数据整理成型，并巧妙地完成可视化操作。通过遵循交互式指南，用户可以快速地构建活动，并将其运行于任何规模的数据集之上，不论是少数小型的电子表格，还是 TB 级数据库，均可适用。这种在批处理方式下按需创建、迁移和处理数据的能力意味着：不了解技术的业务用户不需要等待 IT 人员或数据库管理库人员提供协助，即可继续开展高级分析项目。一旦数据管理人员设定治理策略并建立连接，任何业务用户都可以利用自助式数据准备及整理工具，让数据摆脱未经处理的停滞状态。

Apache Spark：实力与性能的神秘催化剂

DataWorks 是一款功能强大的数据迁移解决方案，因为它拥有多种连接器，适于不同的数据源，包括 dashDB、Salesforce.com、IBM DB2®、Cloudera Impala、Apache Hive 和 Sybase。不过，要保持性能及可扩展性，同时在多个数据源之间稳定运行，DataWorks 需要一种强劲的助力，所以它要借助于领先的开源大数据处理引擎 - Apache Spark。

Spark 是一款免费的易用型工具，随着其大型开源社区的不断发展，Spark 在功能方面与日俱进，能为数据处理和机器学习提供更广泛的支持。Spark 利用集群计算模型实现了 Apache Hadoop 数据处理模型的扩展和改进，其编程界面使用简便，非常适于处理当今 Web 应用和移动应用中常见的流动数据，以及持续的查询工作负载。Spark 所具备的性能、灵活性及易用型，使其成为了快速查询大型数据集的理想之选。

在 DataWorks 中，Spark 引擎采用"幕后"工作方式，实时支持快速的大规模数据操控。用户只需在登录后创建连接，指定安全网关，便可连接本地或云端的数据。整个流程完全对用户可见，DataWorks 先连接 Spark 集群，以快速加载源数据，然后完成一系列的前期准备工作，再执行数据排序、重组、列式操控等其他操作。之后，DataWorks 会将 Spark 驱动的流程存储为活动，以供任何调度计划重复使用。由此，用户可以专注于利用新获得的数据洞察力快速交付业务成果，而非长时间进行手动数据验证。借助 DataWorks 和 Spark，即便是初级用户也能够简便、轻松、安全地管理大量的本地数据或云数据。

IBM Watson Analytics：DataWorks 发挥成效

谈及早期成功案例，DataWorks 集成数据准备功能和云服务集成曾为业内领先的数据分析及可视化工具"IBM Watson Analytics"提供了强劲支持。对于需要在分析和报告之前提升数据质量的业务分析师，DataWorks 通过嵌入 Watson Analytics，为他们提供了单一、集成式的情境体验。通过与 DataWorks 集成，Watson Analytics 获得了多种新功能，其中包括：

- 访问多个企业数据源：目前，用户可以访问多个数据源（不论其位于本地还是云端），以在 Watson Analytics 中进行更深入的分析并制定 BI 报告，这些数据源包括 Amazon Redshift、Apache Hive、Cloudera Impala、IBM DB2、IBM Informix®、IBM Netezza®、IBM SQL Database、IBM dashDB、Microsoft Azure、Microsoft SQL Server、MySQL、Oracle、Pivotal Greenplum、PostgresSQL、Salesforce.com、Sybase 和 Sybase IQ。
- 预载成型：现在，用户无需在加载数据之前对其进行任何的修改或成型操作，便可决定是否将其数据源中的数据加载至 Watson Analytics 中。借助成型技术，用户可以评估数据质量、预览数据、根据列值筛选数据、删除不合要求的列，还可以融合多个来源的数据。
- 安全访问防火墙下游的数据：利用 DataWorks 安全网关，用户可以访问仅在防火墙下游可用的数据。这让管理人员能够在可控访问环境中建立通向服务器的 SSH 隧道，同时建立与本地数据源及其他安全数据源的连接。

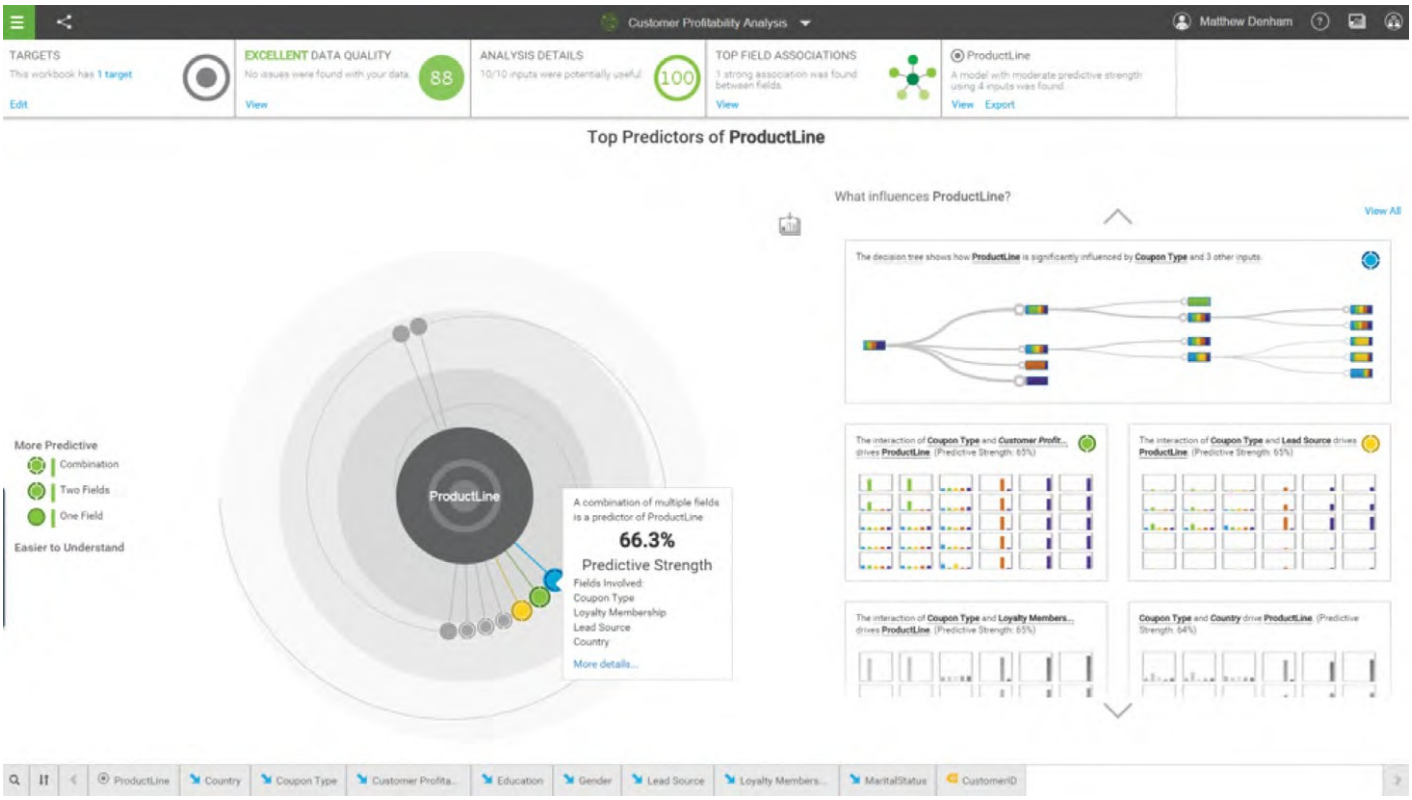

图 4：利用 Watson Analytics 明确客户行为的预测指标

开始使用：获取数据价值

现在，您就可以马上免费使用 DataWorks。您只需访问 Bluemix.net，然后创建一个账户。Bluemix 是 IBM 推出的平台即服务 (PaaS) 产品；通过 Bluemix，您可以访问多种云数据服务，其中就包括与 DataWorks 紧密集成的云数据服务，比如 dashDB 云数据仓库、Cloudant NoSQL 数据库服务等等。对于 1,000 行以内的数据，使用 DataWorks 将不收取任何费用。为此，您可以马上开始加载数据，并将其整理成型，而无需担心任何财务风险，然后以此为起点，再作下一步计划。对于更大的数据集，DataWorks 遵循“即用即付”策略，您无需担心将资金浪费在日后不会投入使用的基础架构上。

有关更多信息及产品体验，敬请访问 ibm.biz/IBMDataWorks。

关于 IBM Cloud Data Services

IBM Cloud Data Services 可向开发者和数据专业人员提供各种丰富的集成式数据服务，服务内容覆盖内容、数据和分析等等。Cloud Data Services 可加快上市速度、延长正常运行时间并帮助 Web 和移动应用开发者实现更多价值。如欲了解 IBM Cloud Data Services 如何改变面向开发人员构建和交付服务的方式，您可以在 Twitter 上通过帐号 @IBMdashDB 和 @IBMcloudant 关注我们，或访问以下网站：ibm.com/analytics/us/en/technology/cloud-data-services。

IBM Corporation
IBM Cloud
Route 100
Somers, NY 10589

美国印刷
2015 年 12 月

1 *Accelerate BI Initiatives With Self-Service Data Discovery And Integration* – Forrester. 2015 年 6 月.

2 *Data Preparation Tools Accelerate Analytics* – Forrester. 2015 年 2 月.(https://www.forrester.com/Brief+Data+Preparation+Tools+Accelerate+Analytics/fulltext/-/E-res119975)

请回收利用

CDW12352-CNZH-00