一体化的威力：IBM + Hortonworks

通过一体化解决方案克服企业数据挑战

HORTONWORKS 白皮书
2018 年 5 月

IBM® 与 Hortonworks® 强强联手，不仅让企业可以轻松享受 Apache™ Hadoop® 的强大能力、可扩展性和经济优势，还能提供额外的监管与安全功能，以及数据联合、高级查询和数据管理工具。他们联合打造出面向未来、面向企业的开源式分析解决方案。

挑战：运用现有数据推动实现分析优势

访问和分析整个企业范围的数据，是一项至关重要的能力。我们目前面临的挑战在于，必须适应云计算、人工智能 (AI) 和物联网 (IoT) 等新技术不断出现以及由此推动的数据数量、数据产生速度以及数据种类爆炸式增长的局面。最大的挑战在于，社交媒体、流式音频 / 视频、日志数据、图像、点击流以及其他渠道所产生的半结构化和非机构化数据越来越多。这种以前所未有的速度呈指数级增长的数据量产生了级联效应。显然，正如研究表明，目前 85% 的数据未被利用，挑战所涉及的范围变得越来越清晰。

为使企业始终保持竞争优势，数据科学家、业务分析师和开发人员必须广泛借鉴各种相关数据源，以便全面了解客户行为，推动内部运营和流程创新。新数据日益多样化，用户对于洞察的需求持续攀升，在这些挑战的综合作用下，刺激了企业采用数据湖的热情，因为这是一种强大灵活的数据管理和分析洞察平台。

Apache Hadoop 是最受欢迎的数据湖技术之一，这种高度可扩展的开源平台旨在处理覆盖数百乃至数千并行计算节点的超大型数据集。它为数据消化提供了一种经济有效的存储解决方案，没有初始格式要求。它作为开源平台，得益于全球最出色的开发人员社区贡献的代码。 大批协作者的聪明才智持续推动创新，以社区为后盾，提供强大支持。

然而，随着新业务模式及创新成果的不断涌现，Hadoop 必须重塑自身，致力打造新一代数据平台，积极推动技术进步。2018 年，更新更强大的 Hadoop 版本问世，IBM 与 Hortonworks 合作，进一步扩展了 Hadoop 的功能，推动数据科学和机器学习提升到全新高度，同时将 Hadoop 深度整合到具有高级分析功能的企业级数据平台。结果就是形成了集 Hortonworks Data Platform (HDP®)、Hortonworks DataFlow (HDF) 和 IBM 技术于一身的解决方案。

IBM 与 Hortonworks 携手推出组合解决方案

IBM 与 Hortonworks 的解决方案建立了基于数据湖的 Hadoop 基础架构，改进了数据探索、数据发现、数据测试和高级数据查询功能。此外，该解决方案不仅可以实现大规模的可扩展性、安全性和监管，还能够联合整个企业范围内静态存储的数据和动态传输的数据。用户可以轻松查询本地或云端的关系数据库和 Hadoop。用户不但可以从自助式数据访问受益，而且还能执行专门查询和实时查询。归根结底，IBM 与 Hortonworks 的解决方案旨在更有效地支持企业规模的机器学习和数据科学发展。

- 排名第一的纯开源 Hadoop 分发版
- 1300 多家客户，2100 多个生态系统合作伙伴
- 广泛汇集 Hadoop（从前隶属于 Yahoo 旗下）的原始架构师、开发人员和运营商的专业知识

- 处理复杂分析工作负载方面最出色的 SQL 引擎
- 排名第一的数据科学平台（根据 Gartner 的评估报告）
- 在本地和混合云解决方案领域处于领先地位
- OpenPOWER 性能名列前茅
- 灵活的软件定义存储

Hortonworks Data Platform 和 Hortonworks DataFlow

Hortonworks Data Platform (HDP) 是业界唯一真正安全的面向企业的开源 Apache Hadoop 分发版，基于集中式资源分配架构 (YARN)。YARN 旨在最大程度消化数据，帮助企业分析数据，以支持高级用例；同时协调集群范围的运营、数据监管和安全服务。HDP 可全方位满足对于静态存储数据和动态传输数据的分析需求，为客户应用提供实时技术支持，并在本地或云端为数据科学家、分析师和开发人员提供实时分析能力。

借助 Hortonworks Data Platform，用户可以：

- 部署、整合及处理规模空前的半结构化、非结构化和结构化数据。
- 使用开源平台，摆脱供应商束缚。
- 最大程度降低 IT 基础架构与 Hadoop 连接所需的费用和精力。
- 利用当前 IT 基础架构，节省时间和资金。
- 确保始终如一地管理数据湖安全。

Hortonworks DataFlow (HDF) 提供简单快速的数据采集、安全可靠的数据传输、划分优先级的数据流和清晰的数据可追溯性等功能。这种端到端平台可实时收集、整理、分析和处理内部或云端的动态数据。它旨在处理从数据源至复杂处理系统（如 Hadoop）的各类数据传输，并且支持与其他数据技术整合。

HDF 可以帮助用户：

- 最大程度提高 IoT 等各种来源的动态数据的价值。
- 汇总复杂处理系统（如 Spark、Storm、Google Cloud DataFlow）、Hadoop 及其他数据存储系统传入的所有类型的数据。
- 使用拖放式直观界面管理流式实时分析，包括数据流管理系统、流处理和企业服务。
- 整合 Apache NiFi/MiNiFi、Apache Kafka、Apache Storm 和 Druid。

通过结合 HDF 与 HDP，企业用户就可以更轻松地整合不同来源的众多数据类型，将它们存储在同一位置。借助这种水平的整合，企业可以在本地和云端运行同一款业界领先的开源平台，确保以统一方式在不同的数据访问引擎中管理各种安全措施。

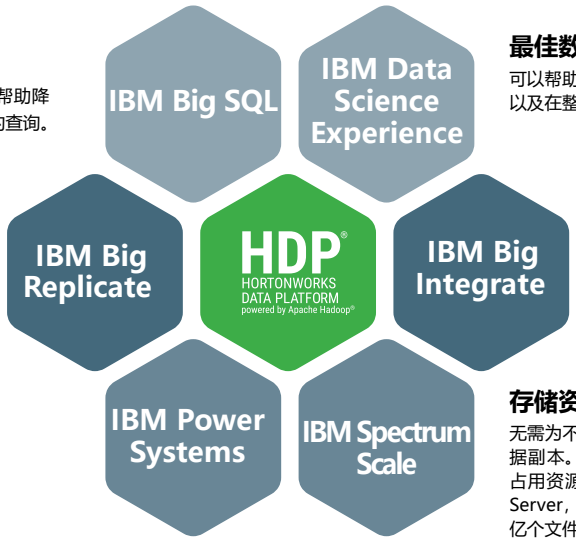
通过结合 HDP 与 HDF，就形成了一种全面的平台，能够轻松应对企业级数据的庞大数量、多样性以及产生速度。辅以 IBM 技术以及专为扩大数据使用范围而创建的 IBM–Hortonworks 生态系统，不仅可以改进查询功能，还能提供高级分析和预测性洞察，提高保持竞争优势所需的业务速度。

IBM 与 Hortonworks 生态系统

面向 Hadoop 的最佳 SQL 引擎
能够在 Hadoop 上处理数据仓库工作负载，帮助降低成本，同时创建性能数据虚拟层，支持复杂的查询。

面向 Hadoop 的最佳灾难恢复解决方案
客户可确保不同 Hadoop 集群的数据一致性和可用性，从而缩短停机时间、缓解风险以及降低成本。

性价比优势提高 3 倍
在 Power Systems 上运行 HDP 的性价比较 x86 高出 3 倍。



最佳数据科学平台
可以帮助数据科学家提高数据生产力、加强协作以及在整个企业范围形成更深入的洞察。

加强 Hadoop 数据监管
准确的数据是唯一可操作的数据。

存储资源最多可减少 60%
无需为不同的应用或数据保护目的而维护多个数据副本。业界最佳的原地分析平台，数据中心占用资源显著减少。使用 IBM Elastic Storage Server，使存储与计算分离，最多可扩展至数十亿个文件。

图 1：IBM 与 Hortonworks 生态系统的价值主张

通过结合 HDP、HDF 和 IBM 的技术，企业可以获得以下优势：

- **企业级 Hadoop 分发版**
  提供大规模的可扩展性、安全性和监管功能。
- **IoT 数据消化**
  支持在网络边缘执行分析，做出明智决策，然后将分析与决策结果发送到数据中心。
- **企业数据移动与混合云**
  支持数据移动：从远程位置移动到数据中心；数据中心间移动；以及在数据中心与云端之间移动。无缝融合数据中心间的数据流。
- **流处理**
  从多个数据流中发掘洞察。
- **通过统一查询，支持及时迅速地获得分析洞察**
  发掘所需的洞察，应对可能影响整个企业的时效性问题。
- **通过工具收集、汇总、联合及查询几乎任何数据**
  从整个企业的本地和云端 Hadoop 及关系数据库中实时挖掘数据洞察。

美国一家主要的休闲游戏企业堪称采用联合解决方案的成功范例。他们结合了不同数据系统的结构化客户数据以及 Hadoop 中存储的半结构化游戏活动日志数据。数据的整合视图产生了深远影响，使得企业可以专注于解答各类关键业务问题，不必花费大量宝贵时间和 IT 资源去从事重复的数据设计任务。此外，由于针对丰富的数据集执行分析，加之向 Hadoop 传输数据的速度较之前的解决方案高出 3 倍，因此数据洞察的获得速度和质量都有显著改善。分析结果不仅帮助提升了企业游戏算法的吸引力，还提供了大量的交叉销售和追加销售商机，直接改善了企业的营收和利润。

一经选择 IBM + Hortonworks 解决方案，企业可以利用以下 IBM 数据管理工具，拓展和丰富自己的数据洞察：

IBM Db2® Big SQL®

面向 Apache Hadoop 的企业级混合 SQL 引擎，具备高度的可扩展性，旨在简化整个企业范围的数据查询。通过单一数据库连接（甚至单次查询），可同时处理 Hive、HBase 和 Spark 数据源。

在 Db2 Big SQL 的帮助下，企业可以通过单一数据库连接甚至单次查询，连接不同的数据源，如 HDFS、RDMS、NoSQL 数据库、对象存储和 WebHDFS。企业可以畅享低延迟、专门查询和复杂查询支持、高性能、强大的安全性、SQL 兼容性以及联合功能带来的价值，充分利用数据仓库和 SQL on Hadoop 的优势。

IBM Data Science Experience

数据科学是跨学科领域，集机器学习、统计、高级分析和编程于一身。IBM Data Science Experience 不仅提供了一系列关键工具，还营造了卓越的协作环境。广大分析人员和开发人员可以在协作环境下运用工具快速轻松地创建新的分析模型。例如，根据 IBM 的测试结果，Data Science Experience 中提供的 IBM 机器学习工具可将构建和部署应用开发分析模型的时间缩短一半。

IBM Big Replicate

在各种受支持的环境、分发版及混合部署中针对 Hadoop 实施主动 / 主动式数据复制。将大数据从实验室复制到生产环境、从生产环境复制到灾难恢复地点，或者从本地复制到受最严苛的业务和监管要求约束的云端对象存储环境。

IBM BigIntegrate

连接性能卓越，数据转换迅速，数据传输功能简便易用，可在 Hadoop 集群的数据节点中执行。这种内存中数据整合解决方案不仅可以提供出色的连接、数据分析功能和元数据管理能力，还与 IBM Streams 整合。

IBM Power® Systems

专为全球要求最严苛的数据密集型计算打造的面向云的服务器。从数据管道中发掘洞察，有效管理任务关键型数据、运营数据存储和数据湖，提供一流的认知计算服务器。

IBM Spectrum Scale ™

为云计算、大数据、分析、对象存储等环境中的非结构化数据提供高级存储管理。凭借安全性、可靠性和高性能优势，全面支持新时代的大数据应用和传统应用。IBM Spectrum Scale 是一项高性能解决方案，运用自身的独特功能大规模管理数据，在数据存储的原地开展存档和分析工作。

更多信息

要了解有关 IBM 与 Hortonworks 解决方案的更多信息，请联系您的 Hortonworks 代表或访问 **Hortonworks** 网站。

关于 Hortonworks

Hortonworks 是领先的企业级全球数据管理平台、服务和解决方案提供商，为超过一半的《财富》100 强企业从任何类型的数据中挖掘切实可行的智能。Hortonworks 致力于推动开源社区创新，为企业客户提供独特价值。Hortonworks 联手合作伙伴提供技术、专业知识与支持，帮助企业客户采用现代化的数据架构。要了解更多信息，请访问 hortonworks.com。

联系方式

如需进一步的信息，请访问
hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

05/2018