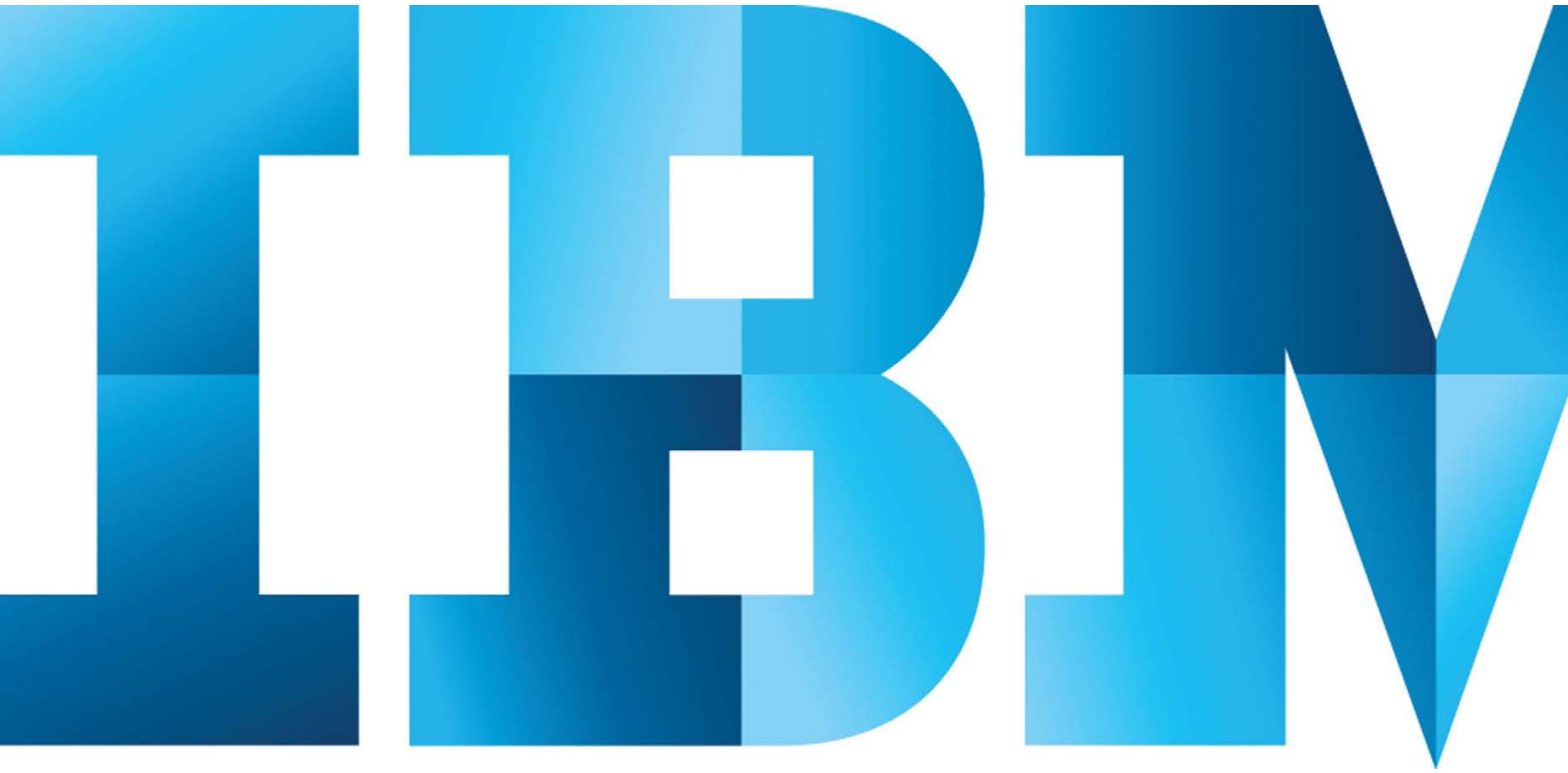
PureSystems

Hadoop 设备：在大数据环境中实现简便性、速度、可扩展性和稳定性的关键

克服 Hadoop 早期采用者所遇到的挑战



IBM

内容

简介

各地企业正在将 Hadoop 作为关键任务的平台。Hadoop 作为一种基于开源的方法和新兴的市场正在迅速日臻成熟。然而，由于下列原因仍处于开发过程中：

- Apache Hadoop 社区仍在为有力的企业部署开发关键规格。
- Hadoop 行业缺乏清晰的参考实施标准、编程接口和跨供应商的互操作性。
- 需要相当大的自定义集成为有利的生产部署准备好 Hadoop 的实施。
- 仅有少数家知名解决方案供应商已经开始将 Hadoop 与其平台、应用程序和工具组合集成起来。
- 获得一致赞同的实现 Hadoop 实施、管理和管控的最佳实践尚未出现。
- 企业数据分析机构缺少深度的 Hadoop 技能集。

为此，许多早期采用 Hadoop 的用户在向 Hadoop 过渡的过程中经历了惨痛的教训。企业必须具有高度专业化的要求和才能出众的技术人才方能取得成功。对于大多数用户而言，在商业硬件上部署 Hadoop 是一种必需，但这种情况在最近有了改观。这是因为市面上提供了少数鲁棒、可重复的 Hadoop 解决方案之故，企业可以将其快速部署到任务关键的环境中。尽管大多数 Hadoop 供应商都采用了核心的开源技术，但在所包含的特定组件、实施的版本、提供的连接器以及支持的编程接口方面有着很大的不同。

过去，回滚自己的 Hadoop 基础架构难度大、成本高、风险大。首先，要求整合不同 Hadoop 技术，包括开源技术和供应商专有技术，来满足特定大数据需求。对于早期用户来说，Hadoop 社区并没有使其变得容易，由于仍然没有统一的参考框架或将不同技术整合到统一平台的互操作性证书。

此外，完整的 Hadoop 部署通常包括编写自定义代码来将集群与现有的数据库、中间件、分析应用程序和其他技术的投资集成。相当数量的 Hadoop 开发集中在 Java 和 Pig 上以在此平台范型中心构建 MapReduce 模型。然而，通常十分有必要使用其他语言以及许多一次性编码来实现所有必要的数据发现、准备、集成和分析功能。自定义 Hadoop 开发的复杂性令人生畏，如果没有强有力的可视化工具和自动化，可能会使全面部署延迟数周或数月。对于早期用户来说，经常就像没有明确的路径图或里程碑来艰难穿越荒原一样。

图 1 定制 Hadoop 开发需求

此外，作为企业大数据范式，Hadoop 尚未达到我们在企业数据仓储 （EDW） 中习以为常的成熟水平。更关键的是，Hadoop 市场尚未集中与平台无关的高可用性、安全性、管控性、联合性和其他强劲的能力的规格。尽管 Hadoop 市场正在迅速所有这些问题，但是如今需要这些功能的用户几乎别无选择，只能使用专有的、即使复杂的功能，IBM 和其他供应商已将这些功能整合到其各自的 Hadoop 产品中。

使用 Hadoop 的十大问题

对于用户来说，使用 Hadoop 的十大最具挑战性问题如下：

1. Hadoop 分布式文件系统 （HDFS） 具有单一故障点。
2. Hadoop 集成、部署、配置管理和性能优化既耗费时间又十分费力。
3. Hadoop 占用大量内存、存储器、输入/输出和网络带宽。
4. Hadoop 集群管理工具在功能方面存在很大差异。
5. Hadoop 混合工作负载管理十分棘手。
6. HDFS 和 Hive 不是实时应用程序的最佳选择。
7. 缺少用于管理 Hadoop、NoSQL、RDBMS 和 EDW 平台的集成工具。
8. Hadoop 技能十分短缺，并且 Hadoop 社区（不是供应商）通常是解决问题和优化的最佳支持来源。
9. 针对 Hadoop 用来检测、诊断和解决问题的工具经常供不应求。
10. 大多数针对 Hadoop 的安全、管控、利用率统计和信息生命周期工具都不存在。

早期采用者也不得不克服技术的陌生感与复杂性。大多数企业的数据分析组织仍缺乏专业的 Hadoop 技能，尽管他们中许多已以尽可能快的速度进行了自学。同时，他们已很难以合理的价格找到完成所有集成并执行管理实施中的 Hadoop 部署所需要的人才。一份最新行业调查表明，约 60% 的公司的核心数

据分析师缺乏 Hadoop 技能。另一项最新公布的报告则探讨了业务领域正面临日趋加剧的 Hadoop 人才危机。如 2012 年 7 月在线求职网站所示，处于开放市场的 Hadoop 专业人员不仅供不应求，而且收费用也相当高。

最终许多供应商，包括 IBM 在内，为帮助客户将 Hadoop 与他们庞大的数据环境相集成开始提供专业服务、咨询及外包产品。

现存的标准 Hadoop 最佳实践及集成模式仍然处于萌芽阶段。开放市场仅有少数“专家”拥有数年的经验。许多从未将此技术部署至生产环境中。而 Hadoop 培训与专业认证服务仍供不应求。因此，许多企业正在使用新技术与未经认证的团队（拥有梦想与希望）建立其自身的 Hadoop 基础设施。

通过最佳实践降低风险

为在此新型领域中降低风险，Hadoop 早期采用者获悉了下列关键最佳实践：

- 避免从 EDW 将大数据分析工作移至 Hadoop 集群，直至对集群进行强化后，其可提供 24x7 全天候可用性、配置及管理。
- 当商业软件与设备选项可用于内部部署时，不要尝试自行创建集成的 Hadoop 栈。通过底层服务器硬件（设备或商品）考察 Hadoop 软件平台的集成程度。
- 使用可用的最佳设备保持 Hadoop 设备的简便性。
- 优选 Hadoop 设备，让您能从服务器“裸机”存储、CPU 等管理一个完整的部署，甚至还能通过使用统一的管理工具管理数据库和其他软件。
- 设计您的 Hadoop 环境以支持硬件与软件资源快速、弹性的供应，从而支持负载拓展。
- 在上线之前对您目标 Hadoop 部署中的每一层实施回归性测试，以确保您的数据、工作及设备在日常运营中不会遭受破坏或遇到瓶颈。
- 通过负载平衡、故障转移、再同步及热备用执行冗余 HDFS NameNode，确保您 Hadoop 集群的高可用性。
- 将您 Hadoop 解决方案供应商是否能够提供高可用性、安全、资源配置、混合工作负载管理、性能优化、健康监控、政策管理、工作安排及其他集群管理功能作为重中之重。
- 依赖您 Hadoop 解决方案供应商对您 Hadoop 集群的设置与配置。
- 若较小部署已能充分满足短期需求，请不要创建过多的 Hadoop 集群容量。

有希望使 Hadoop 部署快速实现价值的设备

幸运的是，Hadoop 部署的苦差事已经开始逐渐减少。开源式核心 Apache Hadoop 技术日趋成熟。供应商以这一正调整围绕这一处于不断变化的核心的不同产品。许多供应商，如 IBM，正在为提高、加强、简化、及使我们的 Hadoop 平台、工具及设备实现自动化不断努力。而大多数领先的数据仓储平台现在为快速数据集成包含标准 Hadoop 连接器。

成熟企业 Hadoop 解决方案的标志之一是其供应商的集成程度，至少应将下列所有功能组件集成至其商业产品中：

- **Hadoop 子项目**：MapReduce、HDFS、HBase、Pig、及 Hive。
- **Hadoop 建模**：MapReduce、Pig 和 Hive 的建模和开发工具。
- **Hadoop 存储**：支持 HDFS 和 HBase、以及 Cassandra 和至少两个 RDBMS。
- **Hadoop 加速和优化**：针对多个分销商、存储层、和/或硬件平台的 Hadoop 性能加速和优化工具。
- **Hadoop 实时与低延迟**：Hadoop 实时与低延迟功能通过支持 HBase 及其他广泛采用的开源式技术或已经提交至 Apache 的专用工具和中间件实现。
- **Hadoop 集群管理**：与 Apache Hadoop 和众多供应商的专用 Hadoop 集群一起使用的集群管理工具，且包括高可用性、工作负载管理和安全性。
- **Hadoop 包装**：许可软件、硬件设备和云/SaaS 续约产品，但具备适用于所有软件功能的开源和商业许可方案。
- **与 EDW 对接的 Hadoop 分布式文件存储 （HDFS）接口**：企业 Hadoop 解决方案应涵盖在来自相同供应商的至少包括一个具有双向 HDFS 接口的 EDW 解决方案的产品系列中。
- **Hadoop 业务应用程序**：企业 Hadoop 解决方案应涵盖在来自提供其自有业务应用程序、具有 Hadoop 应用程序合作伙伴，及包括在他处开发的 Hadoop 开源应用程序的供应商的产品系列中。

如需关于进行扩展以实现完整的 Hadoop 解决方案的信息，请参阅 2012 年出版的《Forrester Wave for Enterprise Hadoop Solutions》（Forrester Wave：企业 Hadoop 解决方案），其中介绍了许多实用的信息。

在快速发展的市场中，成就鲁棒的企业部署所需的功能和工具范围也在不断扩大。直到最近，传统 Hadoop 部署一直缺少的内容在当今的数据仓储时代才被认为理所当然，那就是一系列广泛的经现场验证的设备。

如果您是一位经验丰富的 EDW 专业人员，您很可能对这些设备非常熟悉，可能不会对此部署模型的价值所着迷。此类设备是一种统一的平台，将软件与处理器、存储以及其他硬件相集成，可提供功能特定的、性能优化的功能以实现快速部署。该平台包含可重复的模块化基础架构组件，适合特定的部署角色使用，并且支持特定的容量、吞吐量和性能概况。用户可以随着需求的变化添加其他功能组件，以及随着数据量、速率以及种类的增加线性增加容量，从而获得足够的灵活性。

无论您是针对 Hadoop、EDW 还是某些其他目的部署设备，这些集成的节点都可以提供快速的价值、高效的部署以及线性的扩展。企业现在拥有大量的商用 Hadoop 设备，例如 IBM® PureData™ System for Hadoop，在这些设备的帮助下，可通过简单、可扩展、快速、灵活且鲁棒的基础架构快速实现价值。Hadoop 设备可简化围绕此技术开展的各种部署、配置和优化任务。这些设备整合了内置专家集成模式，可帮助您自动利用已确立的 Hadoop 最佳实践。

Hadoop 设备的简便性源自在设备平台上使用统一的工具对应用程序进行部署、管理和优化。这有助于实现弹性的模块化部署，以及在更广范围的 Hadoop 分析应用程序管理中，提高 IT 人员的生产率。此外，还可以极大地简化 Hadoop 与您的 EDW、NoSQL、流计算、在线分析处理、商业情报、预测分析、事务处理以及其他数据、分析和应用程序基础架构的集成。

借助 Hadoop 设备支持大数据功能

Hadoop 设备应该在支持下列核心大数据功能方面具备扩展和加速能力：

功能	描述
大数据存储	Hadoop 设备可用作数据存储架构的核心构建块。 主要用于存档、监管和复制，以及挖掘、获取和管理多结构化内容。 该设备应为这些关键数据整合功能提供高性能应用的模块性、可扩展性和效率。通常而言，它可以通过与由 IBM 提供的高容量存储区域架构等集成以支持上述功能。
大量数据处理	Hadoop 设备应支持高级数据处理、操作、分析和访问功能的海量并行执行。它应当: • 支持全部高级分析，以及一些在传统意义上与 EDW、BI 和 OLAP 相关联的功能。 • 具有处理查询、计算、数据加载和数据集成等此类核心分析功能所需的所有元数据、模型和其他服务。 • 通过与 IBM PureData System for Operational Analytics 或 IBM PureData System for Analytics 等分析平台对接的接口处理上述功能的子集和接口。
大数据开发	Hadoop 设备应支持大数据建模、挖掘、探索和分析。该设备应: • 提供可扩展“沙箱”，其中所包括的工具可帮助数据科学家、预测建模师和业务分析师以互动合作的方式探索丰富的信息资产。 • 包含高性能分析运行时平台，上述团队在寻求深层次统计模式时，可以在该平台上汇总及准备数据集，调整市场细分和决策树，以及通过统计模型进行迭代。 • 为数据科学家提供大型并行 CPU、内存、存储器和 I/O 容量，用于应对复杂性不断增加的分析工作负载。 • 进行沙箱的弹性扩展，将传统统计分析、数据挖掘和预测建模扩展至 Hadoop/MapReduce、R、地理空间、矩阵操作、自然语言处理、观点分析和其他,资源密集型大数据处理的新领域。

您应该向任何销售商用 Hadoop 设备的供应商询问下列关键问题：

- 你们是否针对 Hadoop 设备的支持、维修和维护提供单点联系？
- 你们是否针对 Hadoop 设备的规划、部署、集成、优化、定制和管理提供全球性专业服务？
- 你们是否通过快速现场问题响应为 Hadoop 设备提供全天候的支持？
- 你们是否通过包含直销和合作伙伴销售在内的多种渠道销售 Hadoop 设备并为其提供支持？
- 你们是否曾通过 SaaS/云服务提供商或者以混合配置的方式优化 Hadoop 设备，以实现在企业 IT 组织内的灵活部署？
- 你们能否针对我的特定大数据工作负载优化 Hadoop 设备？
- 你们的设备能否以弹性的方式进行扩展，以解决我们不断增长的大数据工作负载？

IBM PureData System for Hadoop

近几年来，客户已经部署了 IBM 的最佳 Hadoop 软件产品，即 IBM InfoSphere® BigInsights™，用于上述提及的所有应用程序。了解到客户在 Hadoop 部署中对于简洁性、可扩展性、速度、易管理性和使用性都有更高的要求，我们最近研发了 IBM PureData System for Hadoop。它是一种企业级解决方案，能够扩展并加快针对全系列任务关键型 Hadoop 分析应用程序的 BigInsights。

IBM PureData System for Hadoop 的主要功能和优势有哪些？主要分为三大类：

IBM 价值点	IBM PureData System for Hadoop Analytics
Hadoop 设备功能	IBM PureData System for Hadoop 可提供完备的企业级 Hadoop 平台。能够提供您希望从 IBM 的数据库和分析设备中获得的**简便性、可扩展性、速度、可靠性、可用性、安全性、易管理性和实用性**。其关键 Hadoop 设备功能： • 以单一 SKU 提供，是一种预配置的全套优化软件包，可集成所有硬件和软件组件。 • 整合 IBM System x® Series Servers、IBM 存储器和网络交换机。 • 通过强化的操作系统、冗余磁盘、交换机和电源，以及主节点故障转移支持高可用性和故障恢复。 • 借助仅通过边缘节点访问，以及通过轻量目录访问协议进行身份验证和集成启用安全操作。 • 包括图形用户界面、命令行接口和应用程序编程界面，用于跨 BigInsights 集群内的一个或多个节点对硬件、软件、作业、文件和数据库进行集成管理。
Hadoop 分析平台和工具	IBM PureData System for Hadoop 集成**一流 BigInsights Hadoop 分析软件平台和工具**。该设备的关键 Hadoop 分析运行时和开发功能包括： • 整合 IBM InfoSphere BigInsights 2.1 Enterprise Edition 和进行多结构化数据设置、部署、优化、管理和保护所需的所有工具。 • 包括可扩展的内置分析库和工具，可用于建模、开发和部署 MapReduce 和其他模型。 • 支持 IBM 的 JAQL 访问语言。 • 为 BigInsight 的库统计建模、数据挖掘、预测分析、文本分析、机器学习和空间分析功能，以及 IBM Netezza® 分析库中的相关功能提供嵌入式支持。 • 与适用于社交媒体分析、日志分析和其他 Hadoop 应用程序的可选 IBM 解决方案加速器集成。
Hadoop 集成	IBM PureData System for Hadoop 包括用于 **IBM DB2® 和 IBM PureData System for Analytics** 的接口，以启用充分利用 MapReduce 进行并行和扩展的数据传输。该设备的关键 Hadoop 集成功能包括： • 借助通过 JDBC 的内置组件与所有 IBM 和已写入 BigInsights 的合作伙伴应用程序连接。 • 包括与第三方和开源分析工具、应用程序和库对接的接口。 • 与 IBM InfoSphere Information Server 搭配使用，可用于进行快速的高容量数据发现、获取、准备、加载、转型、清理和监管。

最后，IBM PureData System for Hadoop 是 IBM 设备系列中的一员，这些设备能够支持您全面分析和管理事务数据的需求。所有这些设备均在 IBM PureSystems™ 系列平台上构建，提供集成专业模式来加快部署、安装、优化、监控和管理。

所有这些基于 PureSystems 的设备均支持彼此间无缝数据分布，使用全系列 IBM 数据管理和分析产品以及一整系列非 IBM 平台、工具和应用程序。

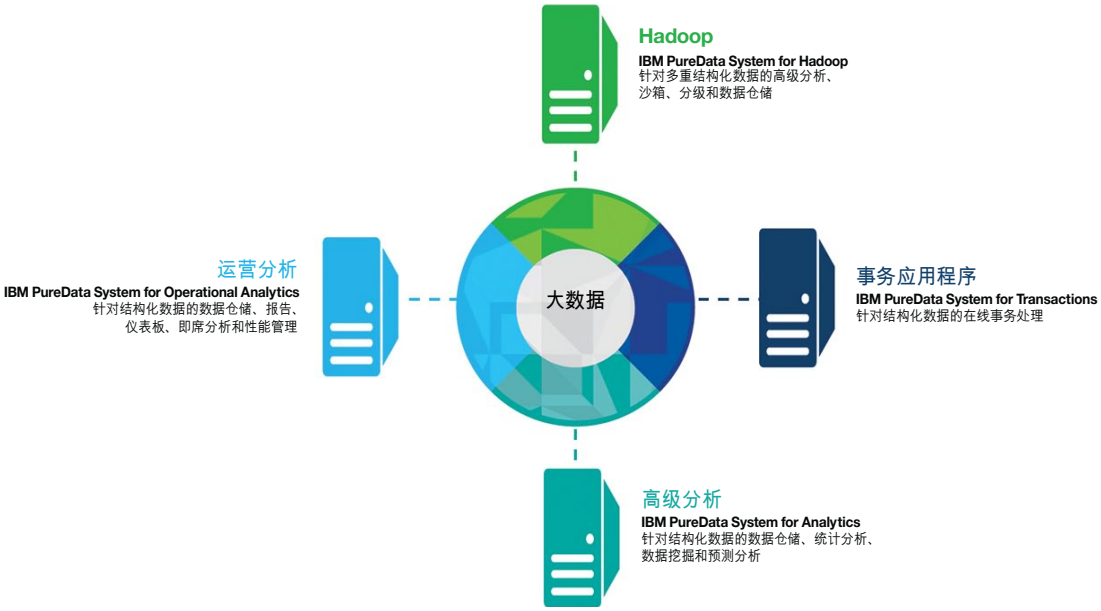


图 2 IBM PureData System 产品

总结

各地企业正在将 Hadoop 作为任务关键型大数据平台。过去，发展自己的 Hadoop 基础架构难度大、成本高、风险大。直到最近，传统 Hadoop 部署一直缺少的内容（即各种设备）在当今的数据仓储时代才被认为理所当然。Hadoop 设备的简洁性来自于为在设备平台上部署、管理和优化设备提供统一的工具。这样就可实现弹性化、模块化部署和更高的 IT 人员生产力。显著简化 Hadoop 与数据仓库、流计算以及其他数据、分析和应用程序基础架构的集成。了解到客户在 Hadoop 部署中对于简洁性、可扩展性、速度、易管理性和使用性都有更高的要求，IBM 最近研发了 IBM PureData System for Hadoop。这是一种企业级别的解决方案，能够扩展并加快针对全系列关键任务大数据应用程序的 IBM Hadoop 平台、InfoSphere BigInsights。

备注

如需更多信息

如需了解 PureData System for Hadoop 的更多详情，请咨询 IBM 代表或 IBM 业务合作伙伴，或访问以下网址：http://www-01.ibm.com/software/data/puredata/hadoop/。

此外，IBM Global Financing 可以帮助您以最经济高效和最具策略性的方式获得您企业所需的软件功能。我们将与符合信用要求的客户合作以定制最适合其业务与发展目标的融资解决方案，实现高效的现金管理，并降低其总拥有成本。IBM Global Financing 可为您的重要 IT 投资筹措资金并推动业务向前迈进。如需更多信息，请访问：**ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

2013 年 9 月

WAW12368-CNZH-00