案例研究

英特尔® 傲腾™ 持久内存
百度 BigSQL 数据处理平台
Spark* 优化分析包

百度 BigSQL 借助 OAP 和英特尔® 傲腾™ 持久内存为 Spark 交互式查询加速

在近年来全球数据规模指数级增长的背景下，如何满足用户对服务时间的要求成为了摆在众多企业，特别是科技企业面前的严峻挑战。应运而生的 Apache Spark* 是大规模数据处理的快速通用计算引擎，专注于帮助企业应对这一挑战。此外，Apache Spark* 还专为大型数据中心的结构化数据处理开发了 Spark SQL* 模块。百度 BigSQL 数据处理平台正是以 Spark SQL* 为基础，并引入了众多新功能和性能拓展。

其中一个重要的性能拓展针对交互式查询需求。为实现次秒级的交互式查询响应，百度联合英特尔展开了 Spark* 平台优化分析包（OAP）项目合作。OAP 能很好地利用列式数据以及选定列上的用户定义索引，提高数据检索效率。它还采用了细粒度的内存数据缓存策略，以消除磁盘和网络中的 I/O 瓶颈，从而将性能最大化至次秒级。

与此同时，随着百度业务不断发展，热点数据的规模也迅速膨胀。为保障持续高品质的访问性能，满足用户需求，内存扩展势在必行。然而，动态随机存取存储器（Dynamic Random-Access Memory，DRAM）的扩展成本高昂，给企业总拥有成本（Total Cost of Ownership，TCO）带来了巨大压力。为了兼顾良好的性能与合理的 TCO，百度与英特尔合作，利用英特尔® 傲腾™ 持久内存（英特尔® 傲腾™ PMem）替代 DRAM 部署更具成本效益的解决方案。

百度内部测试表明，与未使用 PMem 的解决方案相比，英特尔® 傲腾™ PMem 可有效提高 OAP 缓存性能及成本效率，从而大幅提升业务成效，例如帮助百度即席查询服务图灵减少工作负载、降

“百度 Big SQL* 可以为用户提供高性能的即席查询服务，它需要大内存在计算节点本地缓存热数据，以减少 DFS I/O 对查询性能的影响。我们使用来自英特尔的傲腾持久内存，在缓存质量得到保证的同时，极大地提升了集群的处理能力，获得明显的 TCO 收益。”

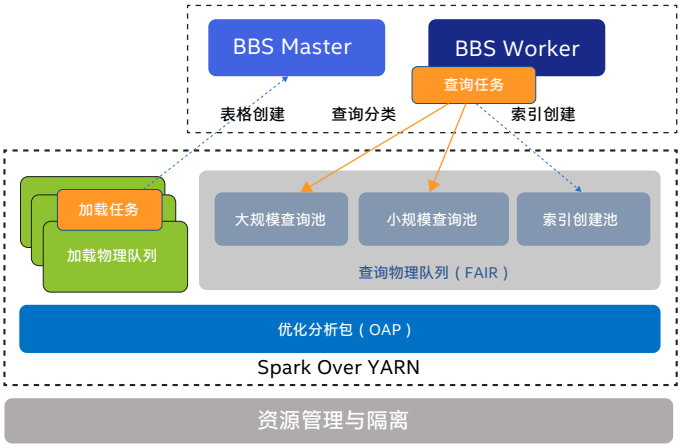
黎世勇
资深系统工程师
百度

低平均查询延时等。

百度 BigSQL 与 OAP

Spark SQL* 的一个核心特征是为批处理提供优化性能。然而，百度面对的一些查询需求具有与批处理完全不同的特征，它们被称作交互式查询。交互式查询虽然同样要访问大型数据集，但却具有非常特定且具体的筛选条件，仅以查询相对少量的数据为目的。因此，与批处理所需的几分钟或几小时不同，用户期望这些少量查询数据能在几秒内甚至几毫秒内返回。显然，已有的百度 Spark SQL* 无法实现交互式查询所要求的性能。

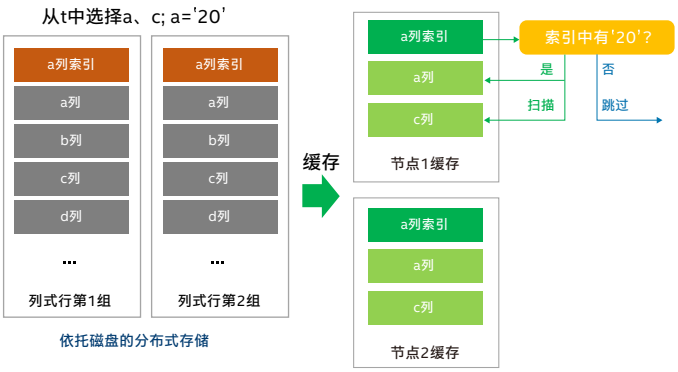
为解决这一难题，百度与英特尔合作实施了 OAP 项目，使用索引和缓存技术来加速交互式查询响应，推动百度 BigSQL 实现令人满意的交互式查询性能。



图一 百度 BigSQL 结合 OAP 优化

当查询具有非常特定的筛选条件时，OAP 可以在符合条件的列上创建索引。通过与列数据文件并排创建与存储完整的 B+Tree 索引，OAP 可快速搜索 B+Tree 索引来识别目标行，同时跳过后端存储（例如 HDFS）上不必要的数据扫描。由于索引文件与原始数据文件保持分离，在创建或删除索引时均无需重写原始数据文件。

在此基础上，为实现查询响应时间从秒级至次秒级的突破，OAP 还借助缓存来优化索引和数据访问进程。通过将索引和数据缓存在内存中，索引加载和数据扫描速度均得到数量级提升，避免了从分布式文件系统读取时磁盘和网络的 I/O 时间。此外，通过将索引和数据单独缓存，其缓存清除和内存空间管理都实现了彼此独立。



图二 OAP 缓存与索引概念

由于采用列式缓存，OAP 只需缓存查询所需的列。基于最近最少使用（LRU）策略，当缓存达到最大容量时，那些最近最少使用的数据项将被淘汰，为缓存最新数据释放空间。依照此策略，百度 BigSQL 启用了一个高级缓存管理器，可以主动填充热点列，并清除缓存中不再需要的列。

英特尔® 傲腾™ 持久内存为百度 BigSQL 提供性能优化

当数据规模较小时，百度 BigSQL 可以通过 DRAM 缓存索引或数据，并实现最佳性能。然而随着百度业务的不断发展，数据集规模日趋庞大，热点数据量超过缓存空间容量，最终导致性能下降。

最简单的解决方案是添加更多的 DRAM，但这一方案有几大缺陷。第一，是每 GB 价格过高，给 TCO 带来了巨大压力。第二，DRAM 内存本身是用于计算的宝贵资源，尤其是在 Spark 环境中，因为每个节点上可配置的总 DRAM 容量有限。第三，DRAM 的一部分优势在于较高的随机访问带宽和较低的延迟，而将其用于大量数据缓存和顺序数据的读取无疑是浪费了 DRAM 的优势。为寻找更具成本效益的替代方案，百度与英特尔携手引入了英特尔® 傲腾™ PMem。

英特尔® 傲腾™ PMem 是一种将高容量、经济实用性和数据持久性相结合的突破性创新技术。作为一种全新类型的内存和存储技术，英特尔® 傲腾™ PMem 专为数据中心设计，其特色优势恰好能够满足百度 BigSQL 的需求，包括：

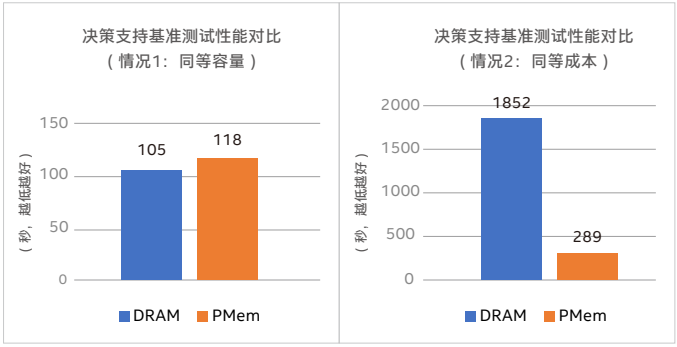
- 适合顺序读取的高带宽
- 大容量和更低的成本

英特尔® 傲腾™ PMem 拥有两种工作模式。在“内存模式”下，应用程序能利用英特尔® 傲腾™ PMem 作为扩展的易失性系统内存，无需重新编写软件，而 DRAM 将起到缓存作用。在“应用直接访问模式”下，经过专门改进的应用程序可从产品固有的持久性中充分获取价值并获得更大的容量。鉴于 OAP 缓存具有索引和输入数据的特定目的，因此 OAP 采用了“应用直接访问模式”，以确保应用程序能完全决策如何使用设备空间。此外，缓存可以从后端存储中重新填充，而无需保持持久性。OAP 使用 memkind 库访问 PMem，无需持久性也不会导致相应的性能损失。

为更好地利用 PMem 代替 DRAM，英特尔还对 OAP 进行了扩展，加入了内存管理器插件，并采用了基于 PMem 的内存管理器以允许在 PMem 中分配缓存空间。用户可以在 DRAM 和 PMem 之间切换，甚至两者兼用，例如使用 DRAM 缓存索引，而使用 PMem 缓存数据。

最后，为了确保 PMem 与百度独特的 OS 环境无缝集成，百度还与英特尔在硬件、操作系统和库等领域进行了一系列广泛的合作优化。

为了验证 OAP 项目及英特尔® 傲腾™ PMem 的性能表现和优势，百度分两步展开了多次评估和内部测试：第一步是决策支持基准测试，第二步使用百度线上业务的真实查询数据。测试的目的是了解 PMem 的表现及成本效率。



图三 DRAM 与英特尔® 傲腾™ PMem 对比测试[1]

在决策支持基准测试中，首先将数据集大小控制为 1TB，并使用相同容量的 DRAM 和 PMem。测试结果表明，二者都能容纳所有的缓存数据，PMem 的性能仅略微低于 DRAM（11.7%），但成本却明显更低[2]。当数据集达到 3TB，且使用相同成本的 DRAM 和 PMem 时，前者的容量已不足以缓存所有数据。相比之下，PMem 不仅具有缓存所有数据的容量，其性能也超出 DRAM 高达 6 倍[3]。此时的 DRAM 性能较差是因为：当数据规模大大超过缓存容量时，DRAM 需要频繁地从后端存储读取数据，从而延长了响应时间。决策支持基准测试清楚表明，在成本相同的情况下，英特尔® 傲腾™ PMem 可提供比 DRAM 更高的容量和更出色的性能。

测试的第二阶段使用了百度线上业务的真实查询数据，依然基于以上两种情况，但方法略有不同。在第一种情况下，DRAM 和 PMem 被设置为缓存 50% 的常用数据列，此时 PMem 的缓存速度仅比 DRAM 低约 12%[4]，但由于成本优势更为显著，因此整体性价比更高。在第二种情况下（即 DRAM 和 PMem 成本相同），只有 PMem 拥有足够容量来缓存所有热点数据，且性能较 DRAM 高出 22%，同时避免了 30% 的底层系统 I/O 请求[5]。

测试结果表明，在百度 BigSQL 中用英特尔® 傲腾™ PMem 取代 DRAM 是更具成本效益的缓存解决方案。此后，百度

[1,2,3,4,5] 测试性能数据于 2019 年 1 月 31 日获得。如欲了解这些测试的更多细节，请与百度联络。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

在BigSQL中成功部署了PMem，并以此为基础优化了百度即席查询服务图灵。在英特尔®傲腾™PMem的加持下，图灵集群的工作负载降低了30%[6]，平均查询延时降低了 20%[7]，每个PMem 服务器实例Spark/OAP 性能提高了50%，而成本仅增加了20%[8]。

展望

目前，众多新兴趋势正推动大数据技术不断进化和演变，其核心也从提供关键功能转变为提供云端解决方案，并通过深度优化来满足性能需求并降低成本。未来，随着百度 BigSQL成为云端解决方案，英特尔®傲腾™ PMem将在性能和TCO方面为其带来更显著的优势。

此外，除了为 Spark SQL* 输入数据提供缓存加速之外，PMem 高容量和高带宽的优势还能在基于 Spark 的机器学习和深度学习场景中发挥更大作用，因为这些场景要求在规模庞大的数据集上反复进行多次计算。另外，Spark shuffle 可以通过优化，使用 RDMA 技术访问远程节点上的 PMem 并将其用作随机存储，从而进一步减少随机延迟并提高性能。

展望未来，百度将与英特尔携手为Spark*实施一系列更深度的优化。其中，英特尔®傲腾™ PMem和第二代英特尔® 至强® 可扩展处理器将发挥核心作用。随着这些技术和产品变得更为先进，它们将为 Spark* 引入更多更强大的加速功能，推动性能和成本效率迈上全新台阶。

[6,7,8]性能表现数据于2019年8月16日获得。如欲了解这些测试的更多细节，请与百度联络。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。
性能测试中使用的软件和工作负荷可能仅在英特尔微处理器上进行了性能优化。
诸如 SYSmark 和 MobileMark 等测试均系基于特定计算机系统、硬件、软件、操作系统及功能。上述任何要素的变动都有可能导致测试结果的变化。请参考其他信息及性能测试（包括结合其他产品使用时的运行性能）以对目标产品进行全面评估。更多信息，详见www.intel.com/benchmarks。
性能测试结果基于截至配置中显示日期进行的测试，且可能并未反映所有公开可用的安全更新。没有任何产品或组件是绝对安全的。详情请参阅配置讯息披露。
您的费用和结果可能会有所不同。
英特尔技术可能需要支持的硬件、软件或服务得以激活。
英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。