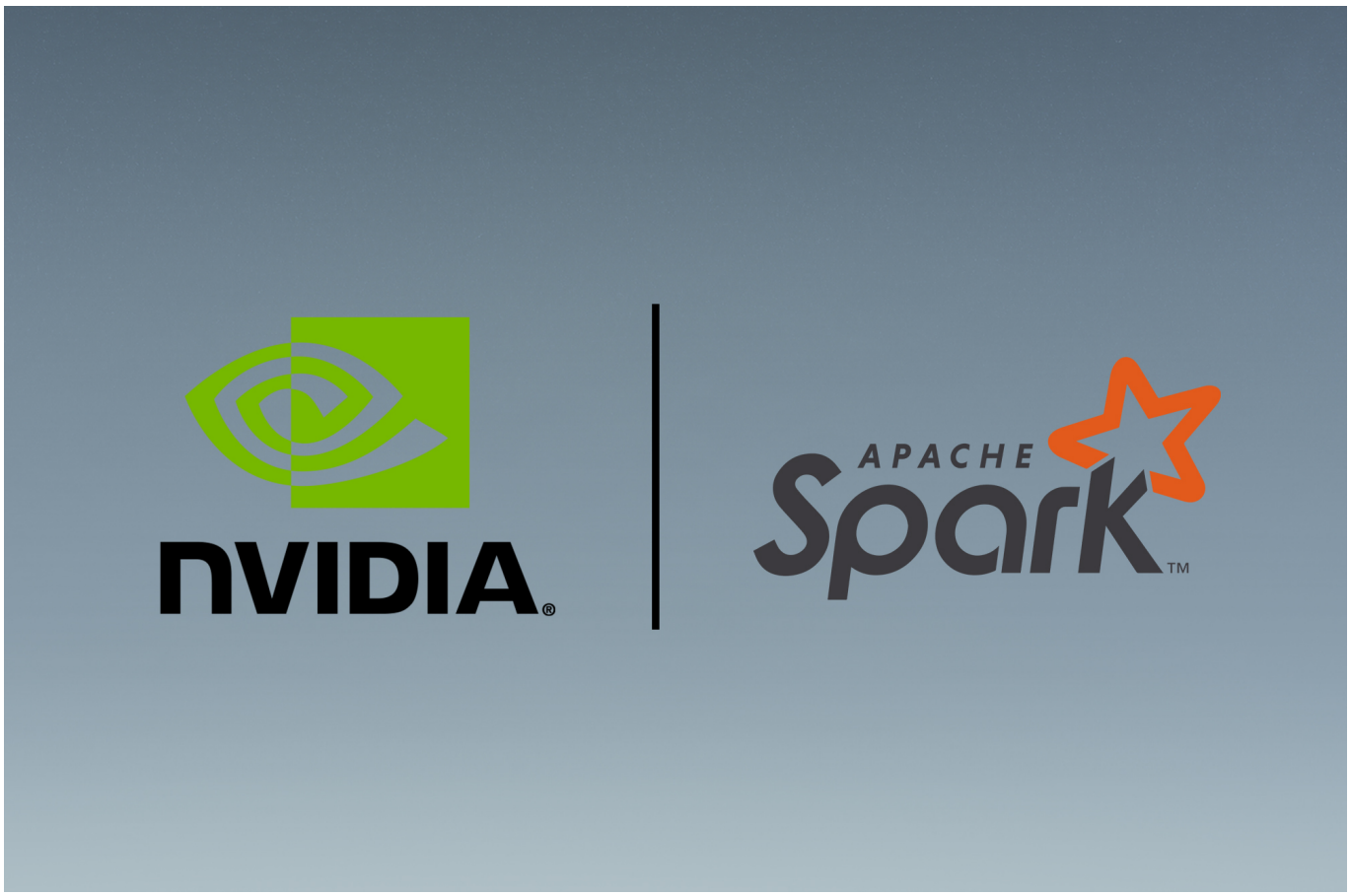
NVIDIA 与数砖合作，将 GPU 加速带入 Apache Spark 3.0

NVIDIA (辉达) 于2020年5月15日宣布将与开源社群携手合作，将端到端的 GPU 加速技术导入 Apache Spark 3.0。全球超过五十万名资料科学家使用 Apache Spark 3.0 分析引擎处理大数据资料。

透过预计于今年春末正式发表的 Spark 3.0，资料科学家与机器学习工程师将能首次把革命性的 GPU 加速技术应用于 ETL (撷取、转换、载入) 资料处理作业负载，这些作业普遍都是透过操作 SQL 资料库来进行。



如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop

另一项创举便是人工智慧 (AI) 模型可以在同一个 Spark 丛集上进行训练，而非将作业负载视为单独的流程，在单独的基础架构上进行训练。这麼一来便能在整个资料科学作业管道上以高效能的方式分析庞大资料，加快处理从资料湖泊 (Data Lake) 到模型训练的数万 TB 资料量，又无需修改用于本地及云端运行之 Spark 应用程式的程式码。

NVIDIA 企业运算部门主管 Manuvir Das 表示：「资料分析正是当前企业与研究人员在高效能运算领域所面临到的最大挑战。从 ETL、训练再到推论，用于整个 Spark 3.0 资料处理管道的本地 GPU

加速技术，提供最终将大数据的潜力与人工智慧的力量串连起来所需的效能和规模。」

基于与 NVIDIA 的策略性 AI 合作伙伴关係，Adobe 是首批推出在 Databricks 上运行 Spark 3.0 预览版本的公司之一。经初步测试，于 Adobe Experience Cloud 中使用 GPU 加速资料分析技术来开发产品，并支援数位业务相关功能，成果显示 Spark 3.0 的运算效能提升了七倍，并省下了 90% 的成本。

Spark 3.0 的效能提升让科学家们可以使用更庞大的资料集来训练模型，并且更频繁地重新训练模型，进而提高模型的准确性。如此一来，科学家们每天都能处理多达数 TB 的新资料，这对于支援线上推荐系统或分析新研究资料的资料科学家们来说十分重要。此外，更快的处理速度也代表著能够减少取得结果所需的硬体资源，大幅节省了成本。

Adobe 机器学习部门资深总监 William Yan 表示：「与 CPU 相比，我们发现 Spark 3.0 使用 NVIDIA 加速技术的运算表现有显著提升。透过改写游戏规则的 GPU 效能表现，为我们整套 Adobe Experience Cloud 应用程式中增强的 AI 驱动功能，开启全新的可能。」

Databricks 与 NVIDIA 加快 Spark 的运算速度

Apache Spark 是由 Databrick 的创办人所打造，该公司基于云端的整合资料分析平台 (Unified Data Analytics Platform) 每天在超过一百万台虚拟机器上运行。NVIDIA 与 Databricks 合作使用 RAPIDSTM 套装软体为 Databricks 优化 Spark，将 GPU 加速技术应用在于 Databricks 上运行的资料科学和机器学习作业附载，其横跨医疗、金融、零售等各种产业。

Apache Spark 的原始建立者，也是 Databricks 的首席技术专家 Matei Zaharia 表示：「我们持续与 NVIDIA 合作，使用针对 Apache Spark 3.0 和 Databricks 的 RAPIDS 优化内容来提升运算效能，我们双方共有的客户如 Adobe 便因此而受惠。这些贡献加快了资料管道处理、模型训练和评分的速度，直接为资料工程师及资料科学家带来更多的突破和崭新的见解。」

借助 NVIDIA GPU 加快 Spark 中的 ETL 及资料传输速度

NVIDIA 正在为 Apache Spark 提供全新的开源 RAPIDS 加速器，以协助资料科学家提高从端到端的资料管道效能表现。此加速器拦截了过去由 CPU 运行的功能，改由 GPU 来执行：

在无需修改任何程式码的前提之下，大幅提升 Spark SQL 和 DataFrame 的运算表现，以加快 Spark 中处理 ETL 资料的速度 在同一套基础架构上加快资料准备及模型训练的速度，机器学习与深度学习则无需使用另外的丛集 加快 Spark 分散式丛集中跨节点的资料传输效能。这些函式库利用 UCF Consortium 的开源 Unified Communication X (UCX) 框架，让资料直接在 GPU 的记忆体之间移动，将延迟情况降到最低。现在可以透过 Apache 软体基金会取得 Spark 3.0 的预览版本，预计将在未来几个月内全面推出。欲了解更多资讯，请参考 www.nvidia.com/spark。

本博客文章除特别声明，全部都是原创！

原创文章版权归过往记忆大数据（过往记忆）所有，未经许可不得转载。
本文链接: 【】（）