IBM 分析

白皮书

基于 Apache Spark 的下一波智能应用

Brandon MacKenzie 和 Joel Horwitz

IBM Corporation

生活处处都会涉及到预测。举例来说，您的上班路线、第二天要去往的地方、是否要继续往下阅读本句子都属于预测。有关未来的预测与进展密切相关。我们可利用预测来帮助我们规划生活，进而提高获得成功的可能性。

不过，人类的判断从本质上来说是不可靠的。由于当今的数据量如此之大，因此没有人能够处理所有的问题。举例来说，很多公司都拥有大量的数据，从中可发现其客户的实际感受以及客户何时、为何会转向竞争对手。问题在于大多数公司并不知道它们所不知道的事情。

在这一点上，机器学习这一领域可提供相关帮助。机器学习不仅正在改变我们与机器互动的方式，而且也在改变着我们与周围世界的关系。在过去的十年中，机器学习已为我们提供了自动驾驶汽车、语音识别和高效 Web 搜索等等，而且大大提高了我们对于人类基因的了解。

机器学习是指*可从数据中学习的系统*。数据即老师。您可以为机器提供学习算法以及可用于其学习特定模式的示例，而无需通过计算机编程下达明确的指令。这些示例通常被称为"*训练示例*"。学习算法通过这些训练示例运行，从而建立模型系数，类似于人类在特定情境下建立行动的肌肉记忆。

不同的机器学习算法可按不同的速度进行学习，而且如同人类一样，有时可从额外投入中获益。不过，很多机器学习算法会消耗大量的计算资源，而且需要较长的学习时间，因而被称为计算密集型机械学习。当向机器学习算法提供更多资源和机会来学习更大、更全面的数据集时，便可作出更佳决策。

截至目前，仅有少数用户拥有机器学习方面的最具创新性应用。对于大多数企业来说，机器学习的开发和产品化存在太大的障碍。大多数公司仅仅是因为不具备正确的技能或必要的技术。不过，该领域的下一波较大的浪潮均与大众化机器学习有关。它可帮助每个人构建可服务于现实世界，并与现实世界进行互动的智能应用。

IBM 致力于使 Apache Spark™ 成为助力下一波机器学习的引擎。Apache Spark 是一个开放源码项目，而非一种产品。简单来说，它是一种进行高度迭代分析的应用框架，该迭代分析可扩展至大量数据。Apache Spark 提供了一种在易于使用的统一环境中聚集应用开发者、数据科学家和数据工程师的平台。它是一种开放源码内存计算引擎，可助力各种高级别工具，如 Spark SQL、用于机器学习的 MLlib、GraphX 和 Spark Streaming。您可在同一应用中将这些库进行无缝结合（见图 1）。

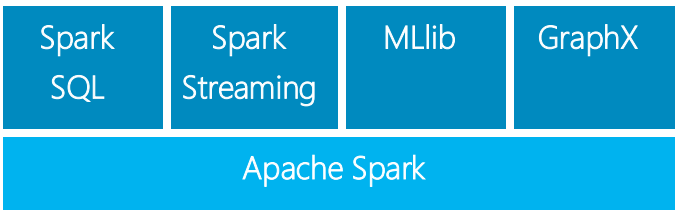

图 1. Apache Spark 堆栈

Apache Spark 的核心引擎和应用编程界面 (API) 是早期分布式计算处理框架的重大改进，如 MPI 和 MapReduce。与之前的低级别选项相比，Spark 高级别 API 更易使用，而且其内存计算引擎经过重新设计，可用于快速的分布式计算。该引擎非常适合于机器学习等迭代算法。借助 Spark，这些算法的执行速度可比 MapReduce 快 100 倍。[1]

借助其强大的引擎和工具，Apache Spark 可大大减少构建分析应用的障碍。它可减少开发分析工作负载的时间和复杂性。通过与数据、设备和人类进行互动，应用的智能化、个性化水平越来越高，而之前未利用的机会也可得到充分利用。通过使用我们周围的所有信息并在需要时获取洞察力，我们可解决原本会被视为无法解决的问题。

在未来五年里，机器学习应用将推动实现有助于增强人类能力的新突破，帮助我们做出更佳选择，并帮助我们以有趣的新方法对我们的世界进行导航。以下给出了一些示例，指导您如何即刻开始使用 Apache Spark 来构建智能分析应用。

自然语言处理：以非结构化的形式捕捉您与客户之间进行的印象最深、最具洞察力的互动。对话时，客户通常会透露您需要向其提供的个性化互动体验相关信息。公司通常可捕捉其客户所思所想的重要信息，但并未利用这些信息。

词频 - 逆文档频率 (TF-IDF) 等自然语言处理技术可将文本转化为可用来训练机器学习算法的信息。TF-IDF 是您可经常在搜索引擎中看到的一种技术。借助 Spark MLlib，您可将自然语言处理直接拷贝至您的应用中，以便您主动管理客户互动。

指导性分析：借助指导性分析，您还可实现更多功能。指导性分析不仅可预测未来会发生什么，还可预测发生原因以及您应该采取的行动。举例来说，您可利用机器学习来决定哪些属性在预测客户行动方面最具预测性意义（即属性重要性）。当您了解了客户行事方式背后的原因，您便可通过互动系统以个性化的方式介入。简言之，机器学习可在需要时帮助您提供定制化的下一步最佳行动。

军队智能：一群专家如同个人一样智能和强大，可更为高效地通力合作，在战争中取胜。机器学习也不例外。Spark MLlib 支持我们称之为"集成 (ensemble)"的机器学习技术。借助集成，很多不同的模型可彼此协作，进而作出更佳预测。该技术非常适合于 Apache Spark 的大规模并行处理。

实时机器学习：借助 Apache Spark，您可以开发和部署可实际上进行实时学习的应用。Spark Streaming 和 MLlib 可使您的应用更具动态适应性。举例来说，"MLlib 流 K 方法实施"是一种可进行动态学习的技术，该技术在数据随时间推移而发生变化时非常有用。这一方法有助于您的应用专注于当前重要的事项上。

自动化：机器学习应用需要自动化和优化。自动化机器学习是 Apache Spark 的实际作用领域之一。举例来说，借助 Spark，您可自动决定训练学习算法的最佳方法，即通常被称为"*超参数调优*"的一种技术。Spark 社区在该领域一直处于领先地位，而 IBM 也很乐意贡献其在自动化和优化领域的专业知识，以推动 Apache Spark 的发展。

问题是人类大脑无法获取大数据方面的所有洞察力。这一点的解决方法便是机器学习。机器学习可改善我们的决策能力，进而交付颠覆性的业务成效。我们已经看到基于 Apache Spark 的机器学习正在改变 IBM 创新的面貌。我们希望其他企业也能与我们保持同步。

有关更多信息

了解如何利用我们的免费训练材料自行创建机器学习模型，敬请访问：BigDataUniversity.com，或前往 Spark 技术中心：http://www.spark.tc ，获取更多相关技术信息。

作者

Brandon MacKenzie 是 IBM Analytics Platform 全球技术销售团队的 Hadoop 数据科学领导者。Brandon 是 Hadoop 和 HPC 环境统计流程领域的专家。他拥有爱丁堡大学的硕士学位。

Joel Horwitz 目前担任 IBM Analytics Platform 的组合营销全球总监。他毕业于西雅图华盛顿大学的纳米技术专业，获得了分子电子学硕士学位。他还拥有匹兹堡大学的产品营销和金融管理国际工商管理硕士学位。

IBM Analytics
Route 100
Somers, NY 10589

美国印刷
2015 年 6 月

1 https://spark.apache.org/