Spark和Hadoop优劣

Spark已经取代Hadoop成为最活跃的开源大数据项目。但是，在选择大数据框架时，企业不能因此就厚此薄彼。近日，著名大数据专家Bernard Marr在一篇文章(http://www.forbes.com/sites/bernardmarr/2015/06/22/spark-or-hadoop-which-is-the-best-big-data-framework/)中分析了Spark和Hadoop的异同。

Hadoop和Spark均是大数据框架，都提供了一些执行常见大数据任务的工具。但确切地说，它们所执行的任务并不相同，彼此也并不排斥。虽然在特定的情况下，Spark据称要比Hadoop快100倍，但它本身没有一个分布式存储系统。而分布式存储是如今许多大数据项目的基础。它可以将PB级的数据集存储在几乎无限数量的普通计算机的硬盘上，并提供了良好的可扩展性，只需要随着数据集的增大增加硬盘。因此，Spark需要一个第三方的分布式存储。也正是因为这个原因，许多大数据项目都将Spark安装在Hadoop之上。这样，Spark的高级分析应用程序就可以使用存储在HDFS中的数据了。

与Hadoop相比，Spark真正的优势在于速度。Spark的大部分操作都是在内存中，而Hadoop的MapReduce系统会在每次操作之后将所有数据写回到物理存储介质上。这是为了确保在出现问题时能够完全恢复，但Spark的弹性分布式数据存储也能实现这一点。

另外，在高级数据处理（如实时流处理和机器学习）方面，Spark的功能要胜过Hadoop。在Bernard看来，这一点连同其速度优势是Spark越来越受欢迎的真正原因。实时处理意味着可以在数据捕获的瞬间将其提交给分析型应用程序，并立即获得反馈。在各种各样的大数据应用程序中，这种处理的用途越来越多，比如，零售商使用的推荐引擎、制造业中的工业机械性能监控。Spark平台的速度和流数据处理能力也非常适合机器学习算法。这类算法可以自我学习和改进，直到找到问题的理想解决方案。这种技术是最先进制造系统（如预测零件何时损坏）和无人驾驶汽车的核心。Spark有自己的机器学习库MLib，而Hadoop系统则需要借助第三方机器学习库，如Apache Mahout。

实际上，虽然Spark和Hadoop存在一些功能上的重叠，但它们都不是商业产品，并不存在真正的竞争关系，而通过为这类免费系统提供技术支持赢利的公司往往同时提供两种服务。例如，Cloudera就既提供Spark服务也提供Hadoop服务，并会根据客户的需要提供最合适的建议。

Bernard认为，虽然Spark发展迅速，但它尚处于起步阶段，安全和技术支持基础设施方还不发达。在他看来，Spark在开源社区活跃度的上升，表明企业用户正在寻找已存储数据的创新用法。

本文转载自：http://www.infoq.com/cn/news/2015/12/Spark-Hadoop-HDFS