Apache Spark 不过时的六大理由

在极短的时间内，Apache Spark 迅速成长为大数据分析的技术核心。这就使得保守派担心在这个技术更新如此之快的年代它是否会同样快的被淘汰呢。我反而却坚信，spark仅仅是崭露头角。

在过去的几年时间，随着Hadoop技术爆炸和大数据逐渐占据主流地位，几件事情逐渐明晰：

1、对所有数据而言，Hadoop分布式文件系统（HDFS）是一个直接存储平台。
2、YARN(负责资源分配和管理)是大数据环境下一个适用的架构。
3、或许是最为重要的一点，目前并不存在一个能解决所有问题的框架结构。尽管MapReduce是一项非常了不起的技术，但是它仍不能解决所有问题。

然而，Spark却可以解决大数据时代中很多关键问题，推动大数据以惊人的速度发展。这就是尽管其还很年轻，我们的“Big Data Discovery”平台依旧使用Apache spark
作为底层技术来处理和分析大数据的原因。

Spark时代即将到来

在寻找关键问题的答案时，基于Hadoop的架构需要调用的多种基础设施和进程来进行分析。他们需要已有的数据，描述性的分析，搜索和更先进的技术，如机器学习，甚至是图形处理。

公司需要这样一个工具，该工具可以让他们充分利用现有技术和资源。至今，尚未存在可以满足上述所有标准的单一处理框架结构。然而，这却是
Spark的最为基本优势，为处理大数据业务的公司提供跨越六个关键领域的技术支持。

1、高级分析

许多大型的创新性公司正在寻求增强他们的高级分析能力。然而,在最近纽约的一次大数据分析会议中，只有20%的参与者表示目前正在公司里部署高级分析。

剩下的80%表示他们正忙于准备数据和提供基本分析。少数科学家花费了大量时间来实施和管理描述分析。Spark为高级分析提供了一个开箱即用的框架，包括加速查询工具，机器学习库，图形处理引擎和流分析引擎。

与MapReduce试图实现这些分析相比——MapReduce几乎不可能实现，甚至说很难找到此类数据科学家——Spark提供了更容易且更快上手的预编译库。这就使得数据科学家可以把任务放在准备数据和保障数据质量之外了。通过Spark他们甚至可以确保分析结果的正确解释。

2、简化

最早对Hadoop的批评不仅仅是它很难使用，而是更难找到会使用它的人。尽管进过后续的迭代后，它变的更加简化和强大，但抱怨声至今未息。

相对于要求用户理解各类复杂的情况，例如Java和MapReduce编程模式，凡具有一些数据库基本知识和一些脚本技能（在Python或者Scala）均可以使用Spark。对于企业而言，能够更容易的找到理解数据并使用工具处理数据的工程师。对供应商而言，我们可以在Spark的上层有所发展并给企业带来更快的创新。

3、多种语言

SQL 语言无法应对大数据分析的面临的所有挑战，至少但依靠它是无法应对的。因此我们需要在解决这个问题上保持更多的灵活性，在组织和检索数据中应有更多的选项，并能快速的将其移动到另一个分析框架中。

Spark保留了SQL语言的模式，采用最快最简洁的方式进行数据分析，不管是什么类型的数据。

4、更快的结果

随着商业业务的不断加快，所以对实时结果的要求是十分必要的。在内存处理上，Spark提供了并行处理的方式使得返回的结果比其他任何其他访问磁盘的方法快了几倍。实时结果去掉延迟后可以显著的减缓商业进程和增量分析。

供应商开始在sparkj上开发应用程序，在工作流程分析上将会出现巨大的进步。加速周转时间意味着分析师可以迭代工作，使得答案更加完整精确。Spark让分析师去做他们的本职工作----——更快且好的寻求答案。

5、不歧视或偏爱的Hadoop供应商

Spark兼容现行所有的Hadoop版本，并有很好的缘由：它是中立的供应商，这意味着它不需要用户去绑定任何特定的供应商。

由于Spark的开源特性，企业可以自由创建基于Spark析基础设施而不用担心会其他事情发生什,即便他们改变Hadoop供应商。如果他们做了什么改变，分析架构也会随之变化。

6、高增性

Apache Spark在极短的时间内取得极大的增长。到2014年为止，Spark在 Daytona Gray Sort 100TB Benchmark.中世界第一。不管是服务、产品抑或技术一旦被迅速关注后，人们通常急于将其搞清楚——如何抑制其炒作，揭示其缺陷或揭穿其的承诺。但根据最近的一项调查显示，人们对Spark的关注仍在增长。

覆盖超过2100产品开发人员的报告显示，71%的受访者有过Spark框架开发经验。如今，它已经拥有多达500多个不同规模的组织，成千上万的开发者和广泛的资源项目参与其中。Spark作

为大数据分析的基本技术之一尚未确定自身的地位，但它已着手去做。换句话说，这仅仅只是开始。

本文转载自：http://www.csdn.net/article/2015-08-26/2825542
英文原文：《6 Reasons That Apache Spark Isn't Flickering Out》http://readwrite.com/2015/08/24/big-data-apache-spark