Spark简介

大家好，在这一章中，我们将向同学们介绍Spark技术，学习Spark ，首先应该了解Spark的发展及特点，本节我们将带领大家走进Spark，了解Spark的发展历史，理解Spark的特点，同时认识Spark的运行架构基本概念并了解Spark的应用场景。首先，我们带大家认识一下Spark，Spark英文的意思是火花、火星，正如它的Logo显示的那样，Spark是处理大数据的电花火石，2009年，Spark诞生于伯克利大学AMPLab，最初属于伯克利大学的研究性项目。实验室的研究人员之前基于Hadoop MapReduce工作，但他们发现MapReduce对于迭代和交互式计算任务效率不高，因此他们研究的Spark主要为交互式查询和迭代算法设计，支持内存存储和高效的容错恢复。我们一起来看看Spark的发展过程。2009年Spark诞生，2010年成为Apache开源的一部分，从2013年6月到2014年2月，仅用了不到一年时间就成为了Apache顶级项目，到目前为止，发布的最新版本为Spark1.4.1

Spark在最近6年内发展迅速，相较于其他大数据平台或框架而言，Spark的代码库最为活跃

参与贡献的开发人员从原来的68位增长到255位

参与贡献的公司也从17家上升到50家。

在这50家公司中，有来自中国的阿里、百度、网易、腾讯、搜狐等公司。最大的集群来自腾讯——8000个节点，广点通是最早使用Spark的应用之一。腾讯大数据精准推荐借助Spark快速迭代的优势，围绕“数据+算法+系统”这套技术方案，实现了“数据实时采集、算法实时训练、系统实时预测”的全流程实时并行高维算法，最终成功应用于广点通PCTR投放系统上，支持每天上百亿的请求量。从图中可以看出，已经有超过1000家企业开始使用该平台，其中包括传统工业厂商 TOYOTA 和著名 O2O 公司 Uber 与 airbnb，说明 Spark 用户领域延伸到传统工业界和互联网与传统行业交叉领域。

此外，越来越多的大数据商业版发行商（如曾今的Hadoop发行商hortonworks，cloudera）也开始将 Spark 纳入其部署范围，这无疑对 Spark 的商业应用和推广起到巨大作用。

Spark具有以下几个特点：首先是具有先进架构。spark使用scala编程语言，scala是一种面向对象的编程语言，无缝地结合了命令式和函数式的编程风格。 scala运行于JVM之上，可以与Java互操作，语言简单，效率高。基于DAG图的执行引擎，减少了多次计算之间中间结果写到Hdfs的开销。

建立在统一抽象的RDD（分布式内存抽象）之上,使得它可以以基本一致的方式应对不同的大数据处理场景。那么什么是DAG呢？

DAG全称 DirectedAcyclicGraph，有向无环图。简单的来说，就是一个由顶点和有方向性的边构成的图中，从任意一个顶点出发，没有任何一条路径会将其带回到出发的顶点。

第二是高效，spark提供Cache机制来支持需要反复迭代的计算或者多次数据共享，减少数据读取的IO开销。

与Hadoop的MapReduce相比，Spark基于内存的运算比MR要快100倍；而比较基于硬盘的运算也要快10倍！

三是易用，Spark提供广泛的数据集操作类型（20+种），不像Hadoop只提供了Map和Reduce两种操作。

Spark有多种交互方式，它支持Java，Python和Scala API，支持交互式的Python和Scala的shell。

四是提供了整体解决方案：主要包括Spark内存中批处理，Spark SQL交互式查询，Spark Streaming流式计算， GraphX和MLlib提供的常用图计算和机器学习算法。最后Spark与Hadoop可以实现无缝衔接，Spark可以使用YARN作为它的集群管理器

读取HDFS,HBase等一切Hadoop的数据

现在Apache Spark已经形成了一个丰富的生态系统，包括官方和第三方开发的组件或工具。Spark生态圈也称为BDAS（伯克利数据分析栈），是伯克利AMPLab实验室打造的，力图在算法、机器、人之间通过大规模集成来展现大数据应用的一个 平台。Spark生态圈以Spark Core为核心，从HDFS、AmazonS3和Hbase等持久层读取数据，以Mesos、Yarn或自身携带的Standalone为资源管理器来调度Job完成Spark应用程序的计算。这些应用程序可以来自于不同的组件，如Spark Shell/Spark Submit的批处理，Spark Streaming的实时处理应用、SparkSQL的即席查询、BlinkDB的权衡查询、Mlib、MLBase的机器学习、GraphX的图处理和SparkR 的数学计算等。

下面向大家详细介绍一下Spark的重要组件。

Spark Core：Spark核心，提供底层框架和核心支持。

BlinkDB：一个用于在海量数据上运行交互式SQL查询的大规模并行查询引擎，它允许用户通过权衡数据精度来提升查询响应时间，其数据的精度被控制在允许的误差范围内。

SparkSQL： 可以执行SQL查询，包括基本的SQL语法和HiveQL语法。读取的数据源包括Hive表、Parquent文件、JSON数据、关系数据库（如MYSQL等）。

Spark Streaming：流式计算。比如，一个网站的流量是每时每刻都在发生的，如果需要知道过去15分钟或一个小时的流量，则可以使用Spark Streaming来解决这个问题。

MLBase：MLbase是Spark生态圈的一部分，专注于机器学习，让机器学习的门槛更低，让一些可能并不了解机器学习的用户也能方便地使用MLbase。

Mlib：MLbase的一部分，MLlib是Spark的数据挖掘算法库，实现了一些常见的机器学习算法和实用程序。

GraphX：图计算的应用在很多情况下处理的数据都是很庞大的，比如在移动社交上面的关系等都可以用图相关算法来进行处理和挖掘，但是如果用户要自行编写相关的图计算算法，并且要在集群中应用，那么难度是非常大的。而使用Spark GraphX就可以解决这个问题，它里面内置了很多的图相关算法。