Life is short, you need Spark!

从零开始

不需要任何基础，带领您无痛入门 Spark

云计算分布式大数据 Spark 实战高手之路

王家林著

Spark 亚太研究院系列丛书

伴随着大数据相关技术和产业的逐步成熟，继 Hadoop 之后，Spark 技术以其无可比拟的优势，发展迅速，将成为替代 Hadoop 的下一代云计算、大数据核心技术。

本书特点

- 云计算分布式大数据 Spark 实战高手之路三部曲之第一部
- 网络发布版为图文并茂方式，边学习，边演练
- 不需要任何前置知识，从零开始，循序渐进

本书作者

Spark 亚太研究院院长和首席专家，中国目前唯一的移动互联网和云计算大数据集大成者。在 Spark、Hadoop、Android 等方面有丰富的源码、实务和性能优化经验。彻底研究了 Spark 从 0.5.0 到 0.9.1 共 13 个版本的 Spark 源码，并已完成 2014 年 5 月 31 日发布的 Spark1.0 源码研究。

Hadoop 源码级专家，曾负责某知名公司的类 Hadoop 框架开发工作，专注于 Hadoop 一站式解决方案的提供，同时也是云计算分布式大数据处理的最早实践者之一。

Android 架构师、高级工程师、咨询顾问、培训专家。

通晓 Spark、Hadoop、Android、HTML5，迷恋英语播音和健美。

“真相会使你获得自由。”

—— 耶稣《圣经》约翰 8:32KJV

“所有人类的不幸都来源于不能直面事实。”

—— 释迦摩尼

“道法自然”

—— 老子《道德经》第 25 章

《云计算分布式大数据 Spark 实战高手之路》

系列丛书三部曲

《云计算分布式大数据 **Spark** 实战高手之路---从零开始》：

不需要任何基础，带领您无痛入门 Spark 并能够轻松处理 Spark 工程师的日常编程工作，内容包括 Spark 集群的构建、Spark 架构设计、RDD、Shark/SparkSQL、机器学习、图计算、实时流处理、Spark on Yarn、JobServer、Spark 测试、Spark 优化等。

《云计算分布式大数据 **Spark** 实战高手之路---高手崛起》：

大话 Spark 源码，全世界最有情趣的源码解析，过程中伴随诸多实验，解析 Spark 1.0 的任何一句源码！更重要的是，思考源码背后的问题场景和解决问题的设计哲学和实现招式。

《云计算分布式大数据 **Spark** 实战高手之路---高手之巅》：

通过当今主流的 Spark 商业使用方法和最成功的 Hadoop 大型案例让您直达高手之巅，从此一览众山小。

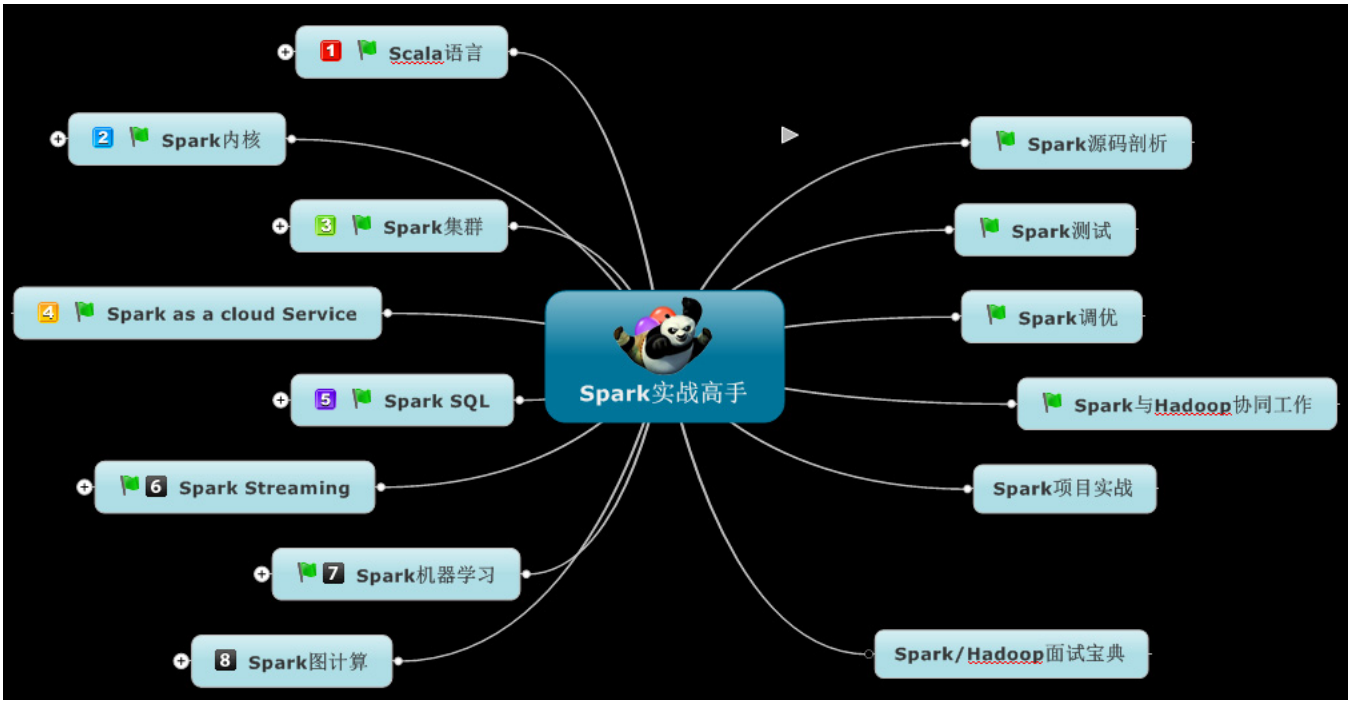
www.sparkinchina.com
TEL: 4006-998-758

《前言》

Spark采用一个统一的技术堆栈解决了云计算大数据的如流处理、图技术、机器学习、NoSQL查询等方面的所有核心问题，具有完善的生态系统，这直接奠定了其一统云计算大数据领域的霸主地位；

要想成为Spark高手，需要经历六大阶段

Spark 实战高手之核心技能点



第一阶段：熟练的掌握Scala语言

1. Spark 框架是采用 Scala 语言编写的，精致而优雅。要想成为 Spark 高手，你就必须阅读 Spark 的源代码，就必须掌握 Scala,；
2. 虽然说现在的 Spark 可以采用多语言 Java、Python 等进行应用程序开发，但是最快速的和支持最好的开发 API 依然并将永远是 Scala 方式的 API，所以你必须掌握 Scala 来编写复杂的和高性能的 Spark 分布式程序；
3. 尤其要熟练掌握 Scala 的 trait、apply、函数式编程、泛型、逆变与协变等；

推荐课程：" 精通Spark的开发语言：Scala最佳实践"

第二阶段：精通Spark平台本身提供给开发者API

1. 掌握Spark中面向RDD的开发模式 掌握各种transformation和action函数的使用；
2. 掌握 Spark 中的宽依赖和窄依赖以及 lineage 机制；
3. 掌握 RDD 的计算流程，例如 Stage 的划分、Spark 应用程序提交给集群的基本过程和 Worker 节点基础的工作原理等

推荐课程： "18 小时内掌握Spark：把云计算大数据速度提高 100 倍以上!"

第三阶段：深入Spark内核

此阶段主要是通过 Spark 框架的源码研读来深入 Spark 内核部分：

1. 通过源码掌握 Spark 的任务提交过程；
2. 通过源码掌握 Spark 集群的任务调度；
3. 尤其要精通 DAGScheduler、TaskScheduler 和 Worker 节点内部的工作的每一步的细节；

推荐课程： “Spark 1.0.0 企业级开发动手：实战世界上第一个Spark 1.0.0 课程，涵盖Spark 1.0.0 所有的企业级开发技术”

第四阶级:掌握基于Spark上的核心框架的使用

Spark 作为云计算大数据时代的集大成者，在实时流处理、图技术、机器学习、NoSQL 查询等方面具有显著的优势，我们使用 Spark 的时候大部分时间都是在使用其上的框架例如 Shark、Spark Streaming 等：

1. Spark Streaming 是非常出色的实时流处理框架，要掌握其 DStream、transformation 和 checkpoint 等；
2. Spark 的离线统计分析功能，Spark 1.0.0 版本在 Shark 的基础上推出了 Spark SQL，离线统计分析的功能的效率有显著的提升，需要重点掌握；
3. 对于 Spark 的机器学习和 GraphX 等要掌握其原理和用法；

推荐课程： “Spark企业级开发最佳实践”

第五阶级:做商业级别的Spark项目

通过一个完整的具有代表性的 Spark 项目来贯穿 Spark 的方方面面，包括项目的架构设计、用到的技术的剖析、开发实现、运维等，完整掌握其中的每一个阶段和细节，这样就可以让您以后可以从容面对绝大多数 Spark 项目。

推荐课程： “Spark架构案例鉴赏：Conviva、Yahoo！、优酷土豆、网易、腾讯、淘宝等公司的实际Spark案例”

第六阶级：提供Spark解决方案

1. 彻底掌握 Spark 框架源码的每一个细节；
2. 根据不同的业务场景的需要提供 Spark 在不同场景的下的解决方案；
3. 根据实际需要，在 Spark 框架基础上进行二次开发，打造自己的 Spark 框架；

推荐课程： “精通Spark：Spark内核剖析、源码解读、性能优化和商业案例实战”

《第三章：Spark 架构设计与编程模型》

Spark 是大数据时代通用而高效的计算平台 基于 RDD 成功实现了“One stack to rule them all” 理念。

目前 SPARK 已经构建了自己的整个大数据处理生态系统，如流处理、图技术、机器学习、NoSQL 查询等方面都有自己的技术 ,并且是 Apache 顶级 Project ,可以预计的是 2014 年下半年到 2015 年在社区和商业应用上会有爆发式的增长。

国外一些大型互联网公司已经部署了 Spark。甚至连 Hadoop 的早期主要贡献者 Yahoo 现在也在多个项目中部署使用 Spark ;国内的淘宝、优酷土豆、网易、搜狐、Baidu、腾讯等已经使用 Spark 技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark 正在逐渐走向成熟，并在这个领域扮演更加重要的角色。

刚刚结束的 2014 Spark Summit 上的信息，Spark 已经获得世界 20 家顶级公司的支持，这些公司中包括 Intel、IBM 等，同时更重要的是包括了最大的四个 Hadoop 发行商都提供了对非常强有力的支持 Spark 的支持。

本章首先介绍为什么 Spark 是大数据必然的现在和未来，接着讲解 Spark 架构和生态系统，然后细致解析 Spark 的编程模型，最后通过众多的案例动手实践 Spark 编程，从零开始，循序渐进，希望助力诸位 Spark 爱好者能够顺利入门 Spark。

Spark 架构设计与编程模型实战共分四个部分：

- 第一部分：为什么 Spark 是大数据必然的现在和未来？！
- 第二部分：Spark 架构设计
- 第三部分：Spark 编程模型
- 第四部分：动手实战 Spark 编程

本章将是 **Spark** 架构设计与编程模型的第一部分：为什么 **Spark** 是大数据必然的现在和未来，具体内容如下所示：

- 1，MapReduce 已死，Spark 称霸 ；
- 2，企业为什么需要 Spark；
- 3，你为什么需要 Spark；
- 4，如何成为云计算大数据 Spark 高手(含思维导图和每个阶段的课程推荐)；

不需任何前置知识，从零开始，循序渐进，成为 Spark 高手！

Life is short, you need Spark!

----Spark 亚太研究院

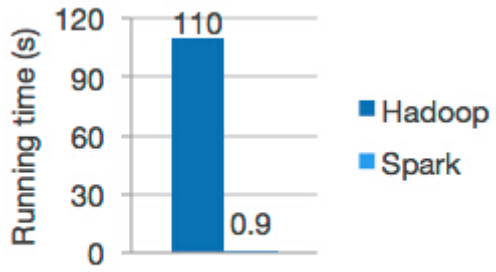
一个工程师最大的幸运是在年轻的时候，遇到一些有挑战的技术和实际有影响力的问题，对解决这些问题的渴望会给人带来持续的动力和成就感。

Spark 给当今的 IT 工程师尤其是云计算大数据工程师提供了这样的一种最大的幸运！

工业和信息化部电信研究院于 2014 年 5 月发布的“大数据白皮书”中指出：

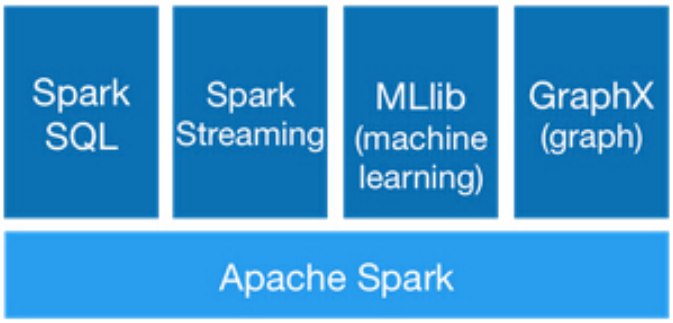
“2012 年美国联邦政府就在全球率先推出“大数据行动计划(Big data initiative)”，重点在基础技术研究和公共部门应用上加大投入。在该计划支持下，加州大学伯克利分校开发了完整的大数据开源软件平台“伯克利数据分析软件栈(Berkeley Data Analytics Stack)，其中的内存计算软件 Spark 的性能比 Hadoop 提高近百倍，对产业界大数据技术走向产生巨大影响”

----来源：工业和信息化部电信研究院



Logistic regression in Hadoop and Spark

Spark 是继 Hadoop 之后，成为替代 Hadoop 的下一代云计算大数据核心技术。目前 SPARK 已经构建了自己的整个大数据处理生态系统，如流处理、图技术、机器学习、Interactive Ad-Hoc Query 等方面都有自己的技术，并且是 Apache 顶级 Project，可以预计的是 2014 年下半年到 2015 年在社区和商业应用上会有爆发式的增长。



国外一些大型互联网公司已经部署了 Spark。甚至连 Hadoop 的早期主要贡献者 Yahoo 现在也在多个项目中部署使用 Spark；国内的淘宝、优酷土豆、网易、Baidu、腾讯、皮皮网等已经使用 Spark 技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark 正在逐渐走向成熟，并在这个领域扮演更加重要的角色。

刚刚结束的 2014 Spark Summit 上的信息，Spark 已经获得世界 20 家顶级公司的支持，这些公司中包括 Intel、IBM 等，同时更重要的是包括了最大的四个 Hadoop 发行商（Cloudera, Pivotal, MapR, Hortonworks 都提供了对非常强有力的支持 Spark 的支持，尤其是是 Hadoop 的头号发行商 Cloudera 在 2014 年 7 月份宣布“Impala’ s it for interactive SQL on Hadoop; everything else will move to Spark”，具体链接信息 http://t.cn/Rvdsukb ，而其实在这次 Spark Summit 之前，整个云计算大数据就已经发声巨变：

1，2014 年 5 月 24 日 Pivotal 宣布了会把整个 Spark stack 包装在 Pivotal HD Hadoop

发行版里面。这意味这最大的四个 Hadoop 发行商（Cloudera，Pivotal，MapR，Hortonworks）都提供了对 Spark 的支持。http://t.cn/RvLF7aM 星火燎原的开始；

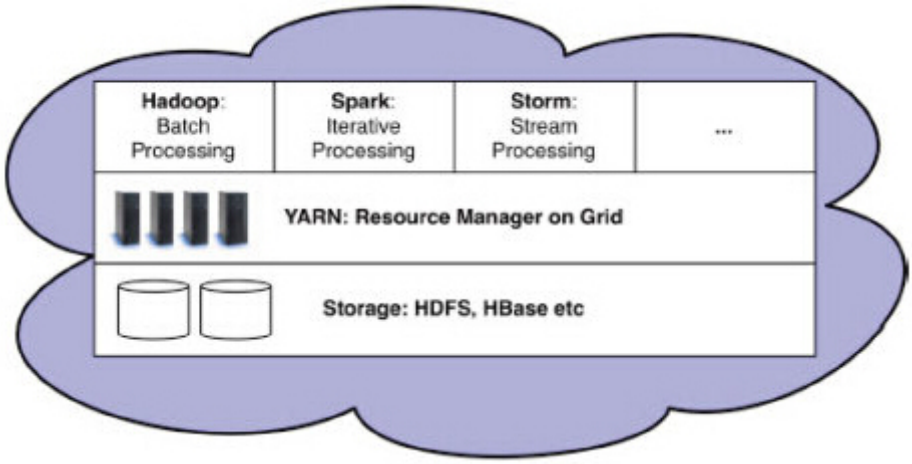
2，Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark；

3，Cloudera 的机器学习框架 Oryx 的执行引擎也将由 Hadoop 的 MapReduce 替换成 Spark；

4，Google 已经开始将负载从 MapReduce 转移到 Pregel 和 Dremel 上；

5，FaceBook 则将原来使用 Hadoop 的负载转移到 Presto 上；

现在很多原来使用深度使用 Hadoop 的公司都在纷纷转向 Spark，国内的淘宝是典型的案例，国外的典型是 Yahoo！，我们以使用世界上使用 Hadoop 最典型的公司 Yahoo！为例，大家可以从 Yahoo！的数据处理的架构图看出 Yahoo！内部正在使用 Spark：



不得不提的是 Spark 的“One stack to rule them all”的特性，Spark 的特点之一就是用一个技术堆栈解决云计算大数据中流处理、图技术、机器学习、交互式查询、误差查询等所有的问题，此时我们只需要一个技术团队通过 Spark 就可以搞定一切问题，而如果基于 Hadoop 就需要分别构建实时流处理团队、数据统计分析团队、数据挖掘团队等，而且这些团队之间无论是代码还是经验都不可相互借鉴，会形成巨大的成本，而使用 Spark 就不存在这个问题；

目录

一、MapReduce 已死，Spark 称霸

由于 Hadoop 的 MapReduce 高延迟的死穴 ,导致 Hadoop 无力处理很多对时间有要求的场景，人们对其批评越来越多，Hadoop 无力改变现在而导致正在死亡。正如任何领域一样 ,死亡是一个过程 ,Hadoop 正在示例这样的一个过程 ,Hadoop 的死亡过程在 2012 年已经开始

1，原先支持 Hadoop 的四大商业机构纷纷宣布支持 Spark；

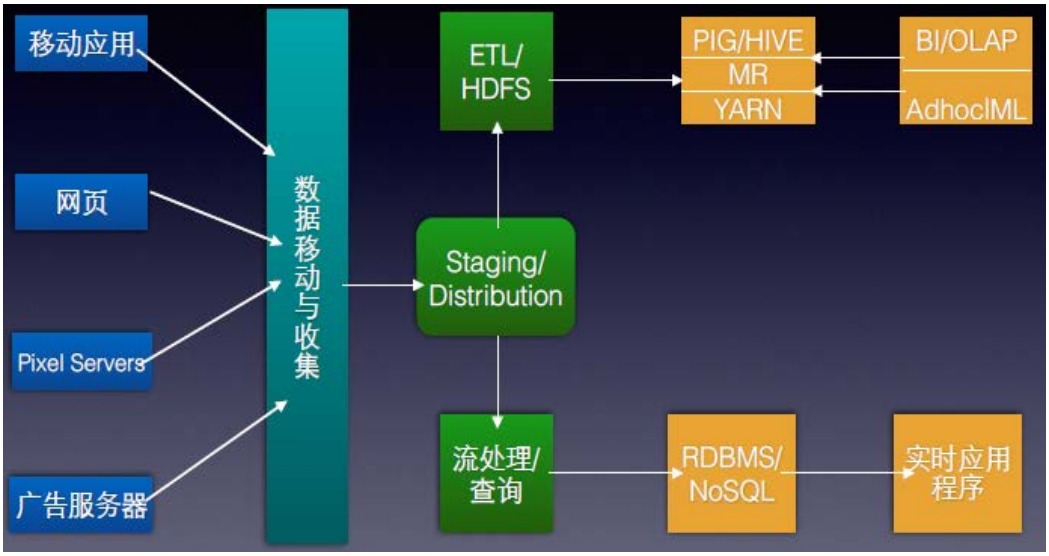
2 ,Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark；

3，Cloudera 的机器学习框架 Oryx 的执行引擎也将由 Hadoop 的 MapReduce 替换成 Spark；

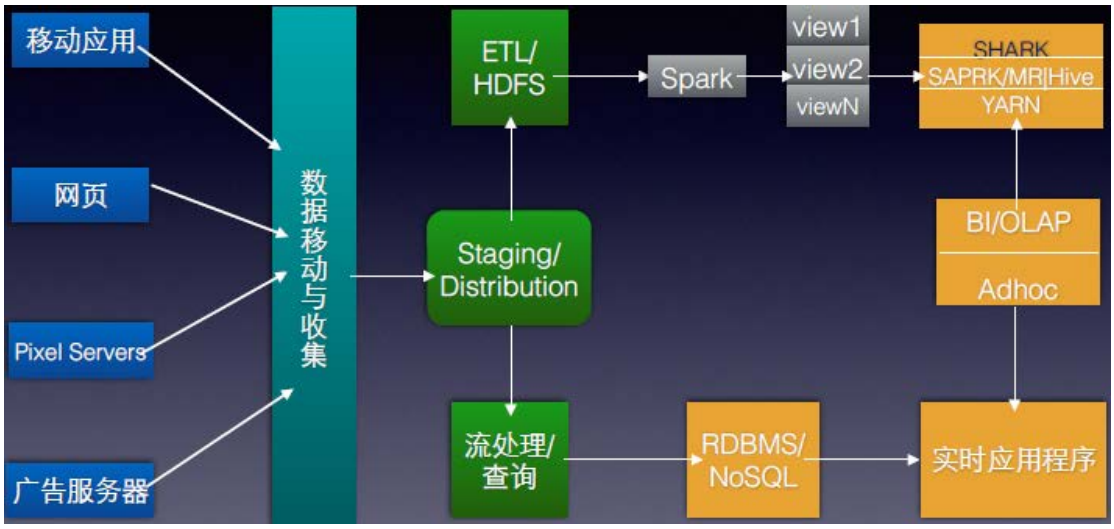
4，Google 已经开始将负载从 MapReduce 转移到 Pregel 和 Dremel 上；

5，FaceBook 则将负载转移到 Presto 上；

现在很多原来使用深度使用 Hadoop 的公司都在纷纷转向 Spark，国内的淘宝是典型的案例。在此，我们以使用世界上使用 Hadoop 最典型的公司 Yahoo！为例，大家可以看一下其数据处理的架构图：



而使用 Spark 后的架构如下：



大家可以看出，现阶段的 Yahoo！是使用 Hadoop 和 Spark 并存的架构，而随着时间的推进和 Spark 本身流处理、图技术、机器学习、NoSQL 查询的出色特性，最终 Yahoo！可能会完成 Spark 全面取代 Hadoop，而这也代表了所有做云计算大数据公司的趋势。

或许有朋友会问，Hadoop 为何不改进自己？

其实，Hadoop 社区一直在改进 Hadoop 本身，但事实是无力回天：

1 ,Hadoop 的改进基本停留在代码层次 ,也就是修修补补的事情 ,这就导致了 Hadoop 现在具有深度的“技术债务”，负载累累；

2，Hadoop 本身的计算模型决定了 Hadoop 上的所有工作都要转化成 Map、Shuffle 和 Reduce 等核心阶段，由于每次计算都要从磁盘读或者写数据，同时真个计算模型需要网络传输，这就导致了越来越不能忍受的延迟性，同时在前一个任务运行完之前，任何一个任务都不可以运行，这直接导致了其无力支持交互式应用；

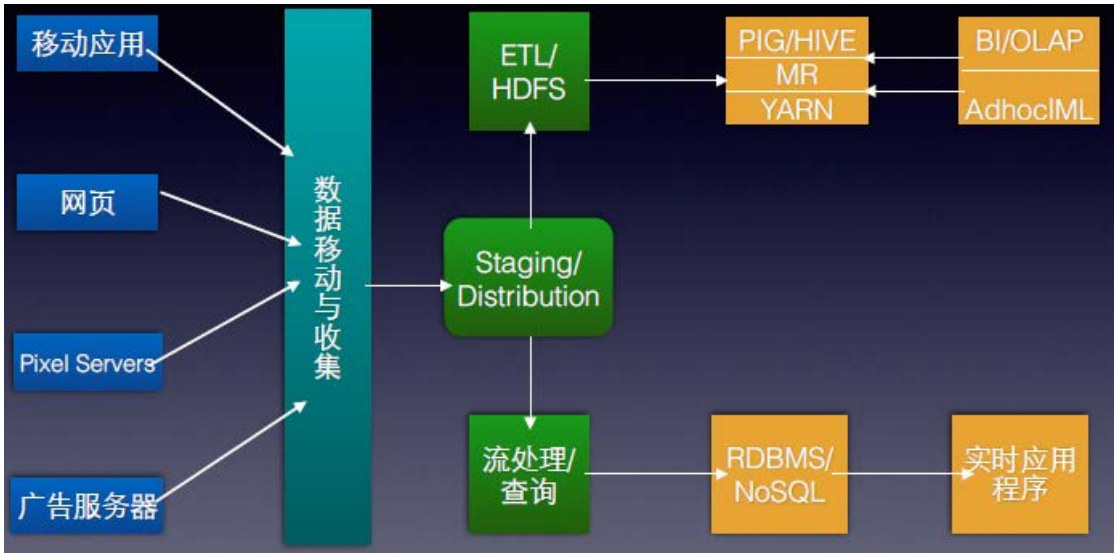
那么，为什么不全部重新写一个更好的 Hadoop 呢?答案是 Spark 的出现使得没有必要这样做了。

Spark 是继 Hadoop 之后，成为替代 Hadoop 的下一代云计算大数据核心技术，目前 SPARK 已经构建了自己的整个大数据处理生态系统，如流处理、图技术、机器学习、NoSQL 查询等方面都有自己的技术，并且是 Apache 顶级 Project，可以预计的是 2014 年下半年到 2015 年在社区和商业应用上会有爆发式的增长。

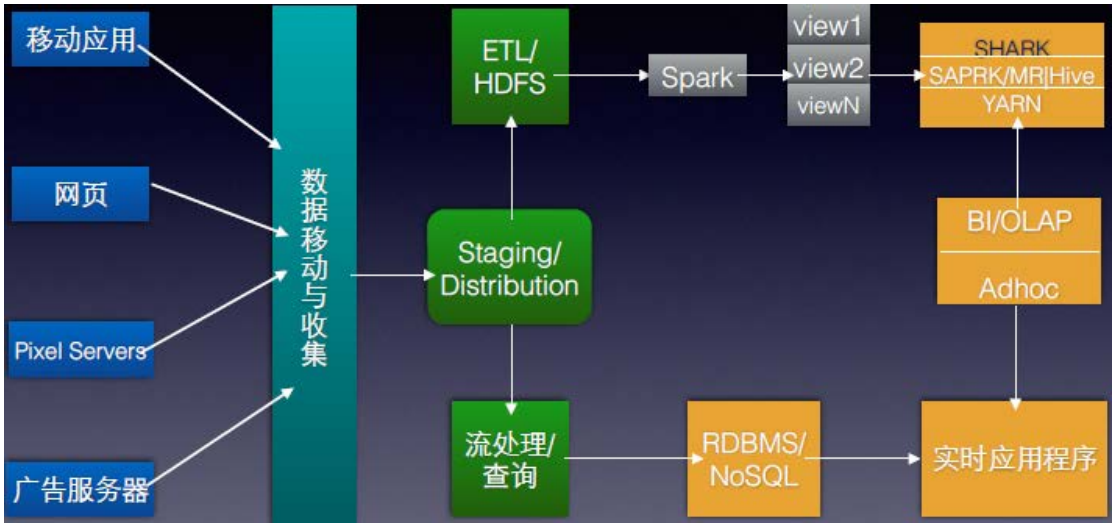
国外一些大型互联网公司已经部署了 Spark。甚至连 Hadoop 的早期主要贡献者 Yahoo 现在也在多个项目中部署使用 Spark；国内的淘宝、优酷土豆、网易、Baidu、腾讯等已经使用 Spark 技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark 正在逐渐走向成熟，并在这个领域扮演更加重要的角色。

二、企业为什么需要 Spark；

1，现在很多原来使用深度使用 Hadoop 的公司都在纷纷转向 Spark，国内的淘宝是典型的案例。在此，我们以使用世界上使用 Hadoop 最典型的公司 Yahoo！为例，大家可以看一下其数据处理的架构图：

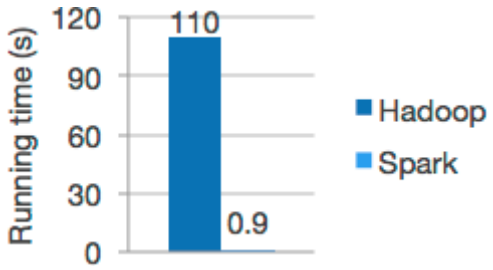


而使用 Spark 后的架构如下：



大家可以看出，现阶段的 Yahoo！是使用 Hadoop 和 Spark 并存的架构，而随着时间的推进和 Spark 本身流处理、图技术、机器学习、NoSQL 查询的出色特性，最终 Yahoo！可能会完成 Spark 全面取代 Hadoop，而这也代表了所有做云计算大数据公司的趋势。

2，Spark 是可以革命 Hadoop 的目前唯一替代者，能够做 Hadoop 做的一切事情，同时速度比 Hadoop 快了 100 倍以上：



Logistic regression in Hadoop and Spark

可以看出在 Spark 特别擅长的领域其速度比 Hadoop 快 120 倍以上！

,3，原先支持 Hadoop 的四大商业机构纷纷宣布支持 Spark，包含知名 Hadoop 解决方案供应商 Cloudera 和知名的 Hadoop 供应商 MapR；

4，Spark 是继 Hadoop 之后，成为替代 Hadoop 的下一代云计算大数据核心技术，目前 SPARK 已经构建了自己的整个大数据处理生态系统，如流处理、图技术、机器学习、NoSQL 查询等方面都有自己的技术，并且是 Apache 顶级 Project，可以预计的是 2014 年下半年到 2015 年在社区和商业应用上会有爆发式的增长。

5，国外一些大型互联网公司已经部署了 Spark。甚至连 Hadoop 的早期主要贡献者 Yahoo 现在也在多个项目中部署使用 Spark；国内的淘宝、优酷土豆、网易、Baidu、腾讯等已经使用 Spark 技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark 正在逐渐走向成熟，并在这个领域扮演更加重要的角色。

6，不得不提的是 Spark 的“One stack to rule them all”的特性，Spark 的特点之一就是用一个技术堆栈解决云计算大数据中流处理、图技术、机器学习、交互式查询、误差查询等所有的问题

7，Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark；

8，如果你已经使用了 Hadoop，就更加需要 Spark。Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark，同时，这几年来，Hadoop 的改进基本停留在代码层次，也就是修修补补的事情，这就导致了 Hadoop 现在具有深度的“技术债务”，负载累累；

8，，此时我们只需要一个技术团队通过 Spark 就可以搞定一切问题，而如果基于 Hadoop 就需要分别构建实时流处理团队、数据统计分析团队、数据挖掘团队等，而且这些团队之间无论是代码还是经验都不可相互借鉴，会形成巨大的成本，而使用 Spark 就不

存在这个问题；

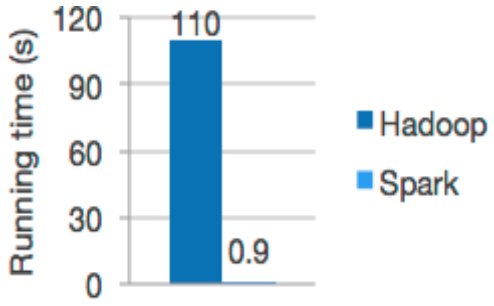
9，百亿美元市场，教授为之辞职，学生为止辍学，大势所趋！

10，Life is short！

三、你为什么需要 Spark；

你需要 Spark 的十大理由：

1，Spark 是可以革命 Hadoop 的目前唯一替代者，能够做 Hadoop 做的一切事情，同时速度比 Hadoop 快了 100 倍以上：



Logistic regression in Hadoop and Spark

可以看出在 Spark 特别擅长的领域其速度比 Hadoop 快 120 倍以上！

2，原先支持 Hadoop 的四大商业机构纷纷宣布支持 Spark，包含知名 Hadoop 解决方案供应商 Cloudera 和知名的 Hadoop 供应商 MapR；

3，Spark 是继 Hadoop 之后，成为替代 Hadoop 的下一代云计算大数据核心技术，目前 SPARK 已经构建了自己的整个大数据处理生态系统，如流处理、图技术、机器学习、NoSQL 查询等方面都有自己的技术，并且是 Apache 顶级 Project，可以预计的是 2014 年下半年到 2015 年在社区和商业应用上会有爆发式的增长。

4，国外一些大型互联网公司已经部署了 Spark。甚至连 Hadoop 的早期主要贡献者 Yahoo 现在也在多个项目中部署使用 Spark；国内的淘宝、优酷土豆、网易、Baidu、腾讯等已经使用 Spark 技术用于自己的商业生产系统中，国内外的应用开始越来越广泛。Spark 正在逐渐走向成熟，并在这个领域扮演更加重要的角色。

5，不得不提的是 Spark 的“One stack to rule them all”的特性，Spark 的特点之一就是用一个技术堆栈解决云计算大数据中流处理、图技术、机器学习、交互式查询、误差查询等所有的问题，此时我们只需要一个技术团队通过 Spark 就可以搞定一切问题，而如果基于 Hadoop 就需要分别构建实时流处理团队、数据统计分析团队、数据挖掘团队等，而且这些团队之间无论是代码还是经验都不可相互借鉴，会形成巨大的成本，而使用 Spark

就不存在这个问题；

6，Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark；

7，如果你已经使用了 Hadoop，就更加需要 Spark。Mahout 前一阶段表示从现在起他们将不再接受任何形式的以 MapReduce 形式实现的算法，另外一方面，Mahout 宣布新的算法基于 Spark，同时，这几年来，Hadoop 的改进基本停留在代码层次，也就是修修补补的事情，这就导致了 Hadoop 现在具有深度的“技术债务”，负载累累；

8，伴随 Spark 技术的普及推广，对专业人才的需求日益增加。Spark 专业人才在未来也是炙手可热，轻而易举可以拿到百万的薪酬；

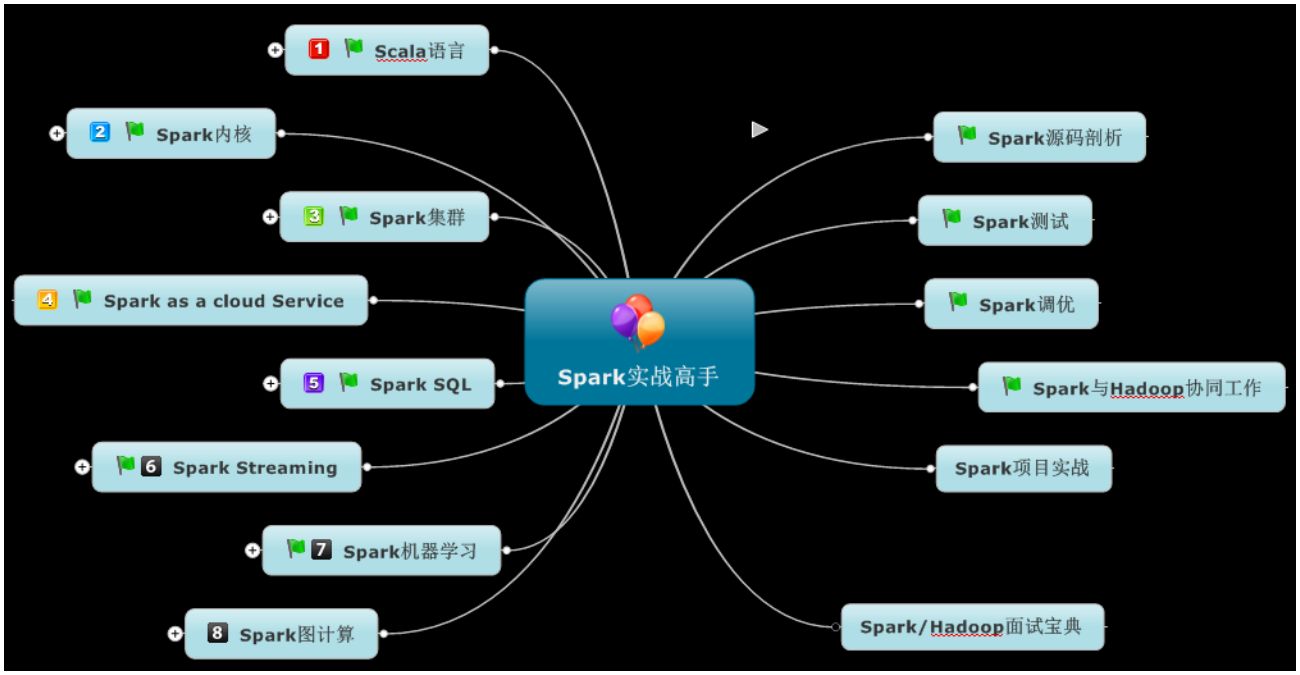
9，百亿美元市场，教授为之辞职，学生为止辍学，大势所趋！

10，Life is short.

四、如何成为云计算大数据 Spark 高手(含思维导图和每个阶段的课程推荐)；

Spark 采用一个统一的技术堆栈解决了云计算大数据的如流处理、图技术、机器学习、NoSQL 查询等方面的所有核心问题，具有完善的生态系统，这直接奠定了其一统云计算大数据领域的霸主地位；

要想成为 Spark 高手，需要经历六大阶段：



第一阶段：熟练的掌握Scala语言

1， Spark 框架是采用 Scala 语言编写的，精致而优雅。要想成为 Spark 高手，你就必须阅读 Spark 的源代码，就必须掌握 Scala,；

2， 虽然说现在的 Spark 可以采用多语言 Java、Python 等进行应用程序开发，但是最快速的和支持最好的开发 API 依然并将永远是 Scala 方式的 API，所以你必须掌握 Scala 来编写复杂的和高性能的 Spark 分布式程序；

3， 尤其要熟练掌握 Scala 的 trait、apply、函数式编程、泛型、逆变与协变等；

推荐课程：”精通Spark的开发语言：Scala最佳实践”

第二阶段：精通Spark平台本身提供给开发者API

1， 掌握 Spark 中面向 RDD 的开发模式，掌握各种 transformation 和 action 函数的使用；

2， 掌握 Spark 中的宽依赖和窄依赖以及 lineage 机制；

3， 掌握 RDD 的计算流程，例如 Stage 的划分、Spark 应用程序提交给集群的基本过程和 Worker 节点基础的工作原理等

推荐课程：“18 小时内掌握Spark：把云计算大数据速度提高 100 倍以上!”

第三阶段：深入Spark内核

此阶段主要是通过 Spark 框架的源码研读来深入 Spark 内核部分：

1， 通过源码掌握 Spark 的任务提交过程；

2， 通过源码掌握 Spark 集群的任务调度；

3， 尤其要精通 DAGScheduler、TaskScheduler 和 Worker 节点内部的工作的每一步的细节；

推荐课程：“Spark 1.0.0 企业级开发动手：实战世界上第一个Spark 1.0.0 课程，涵盖Spark 1.0.0 所有的企业级开发技术”

第四阶级:掌握基于Spark上的核心框架的使用

Spark 作为云计算大数据时代的集大成者，在实时流处理、图技术、机器学习、NoSQL 查询等方面具有显著的优势，我们使用 Spark 的时候大部分时间都是在使用其上的框架例如 Shark、Spark Streaming 等：

1， Spark Streaming 是非常出色的实时流处理框架，要掌握其 DStream、transformation 和 checkpoint 等；

2， Spark的离线统计分析功能，Spark 1.0.0版本在Shark的基础上推出了Spark SQL，离线统计分析的功能的效率有显著的提升，需要重点掌握；

3， 对于 Spark 的机器学习和 GraphX 等要掌握其原理和用法；

推荐课程：“Spark企业级开发最佳实践”

第五阶级:做商业级别的Spark项目

通过一个完整的具有代表性的 Spark 项目来贯穿 Spark 的方方面面，包括项目的架构设计、用到的技术的剖析、开发实现、运维等，完整掌握其中的每一个阶段和细节，这样就可以让您以后可以从容面对绝大多数 Spark 项目。

推荐课程：“Spark架构案例鉴赏：Conviva、Yahoo！、优酷土豆、网易、腾讯、淘宝等公司的实际Spark案例”

第六阶级：提供Spark解决方案

1， 彻底掌握 Spark 框架源码的每一个细节；

2， 根据不同的业务场景的需要提供 Spark 在不同场景的下的解决方案；

3， 根据实际需要，在 Spark 框架基础上进行二次开发，打造自己的 Spark 框架；

推荐课程：“精通Spark：Spark内核剖析、源码解读、性能优化和商业案例实战”

前面所述的成为 Spark 高手的六个阶段中的第一和第二个阶段可以通过自学逐步完成，随后的三个阶段最好是由高手或者专家的指引下一步步完成，最后一个阶段，基本上就是到”无招胜有招”的时期，很多东西要用心领悟才能完成。

Spark 亚太研究院

Spark 亚太研究院，提供 Spark、Hadoop、Android、Html5、云计算和移动互联网一站式解决方案。以帮助企业规划、部署、开发、培训和使用为核心，并规划和实施人才培训完整路径，提供源码研究和应用技术训练。

近期活动及相关课程

决战云计算大数据时代 Spark 亚太研究院 100 期公益大奖堂
每周四晚上 20：00—21:00
课程介绍：http://edu.51cto.com/course/course_id-1659.html#showDesc
报名参与：http://ke.qq.com/cgi-bin/courseDetail?course_id=6167

近期公开课：

《决胜大数据时代：Hadoop、Yarn、Spark 企业级最佳实践》
集大数据领域最核心三大技术：Hadoop 方向 50%：掌握生产环境下、源码级别下的 Hadoop 经验，解决性能、集群难点问题；Yarn 方向 20%：掌握最佳的分布式集群资源管理框架，能够轻松使用 Yarn 管理 Hadoop、Spark 等；Spark 方向 30%：未来统一的大数据框架平台，剖析 Spark 架构、内核等核心技术，对未来转向 SPARK 技术，做好技术储备。课程内容落地性强，即解决当下问题，又有助于驾驭未来。
开课时间：9 月 26—28 日 上海、10 月 26—28 日北京、11 月 1—3 日深圳
咨询电话：4006-998-758