小象科技
ChinaHadoop.cn

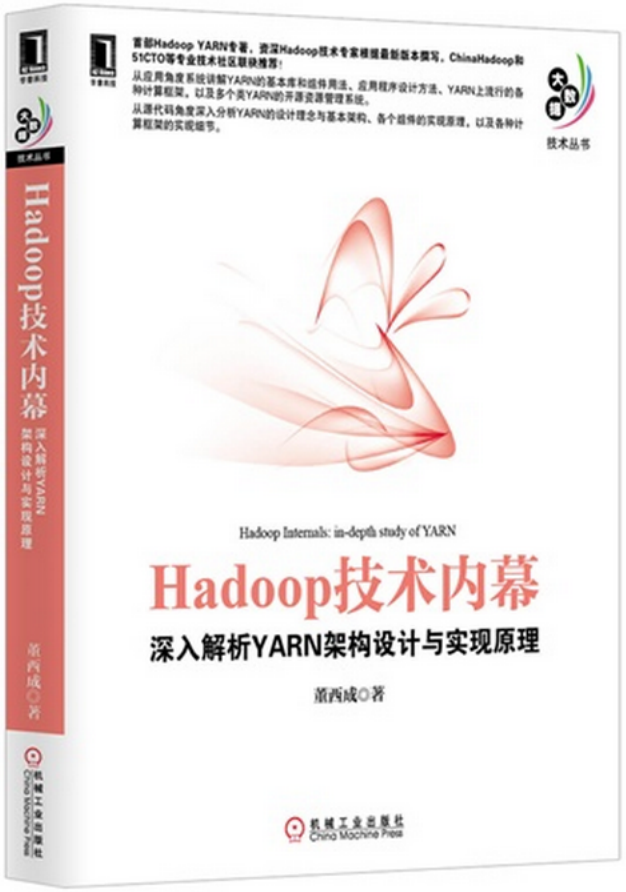
课程安排

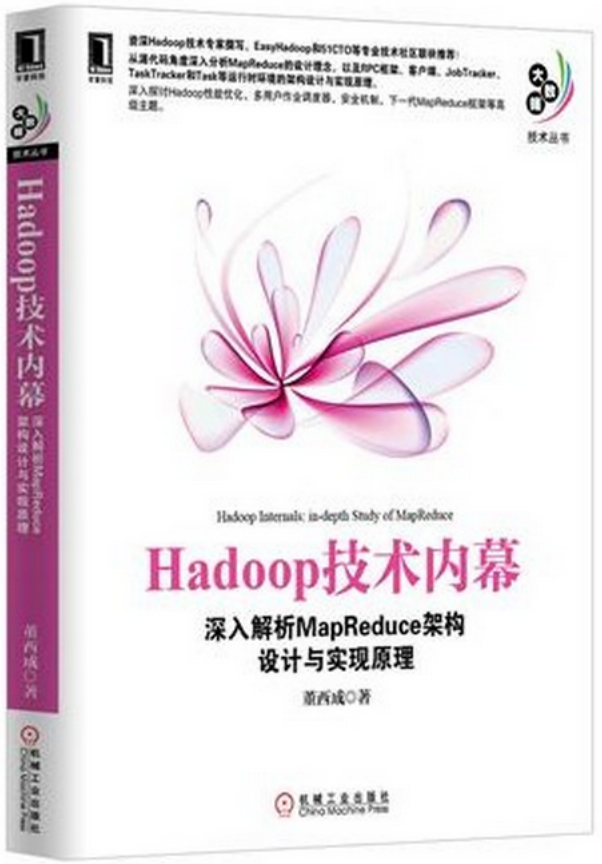
董西成
2017年04月

自我介绍

- 硕士毕业于中国科学院（计算技术研究所）；
- 目前就职于**hulu**（北美著名在线视频公司）；
- **2009**年开始接触**hadoop**，在**hadoop**之上进行了大量定制和二次开发；
- 技术博客：**http://dongxicheng.org/**
- 技术书籍：**http://hadoop123.com**

自我介绍





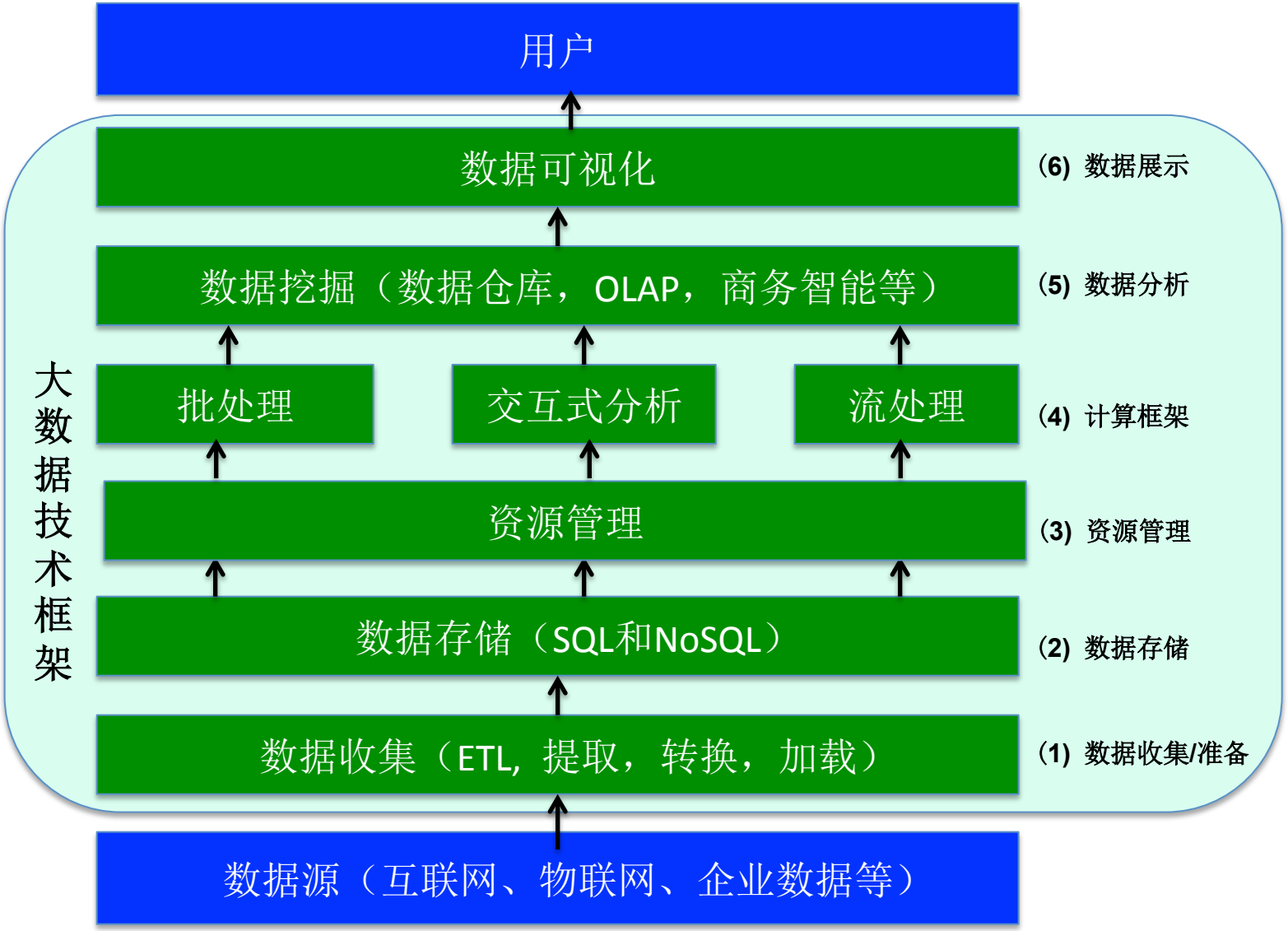
主要内容

1. 大数据技术框架
2. Hadoop与Spark生态系统
3. 课程内容安排

主要内容

1. 大数据技术框架
2. Hadoop与Spark生态系统
3. 课程内容安排

大数据技术框架



改编自：工业和信息化部电信研究院，“2014 大数据白皮书”

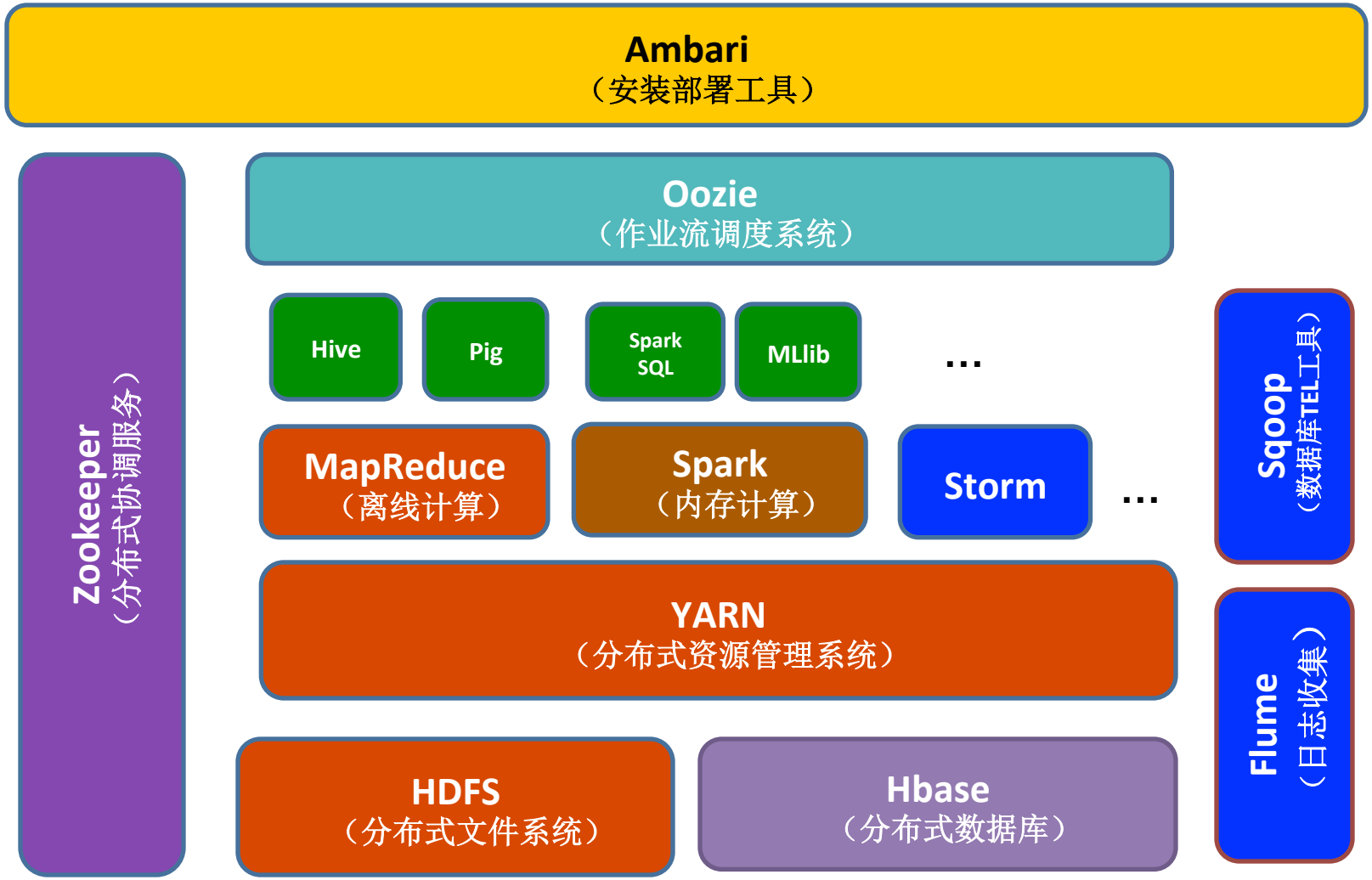
主要内容

1 大数据技术框架

2 Hadoop与Spark生态系统

3 课程内容安排

Hadoop生态系统



计算类型及应用场景

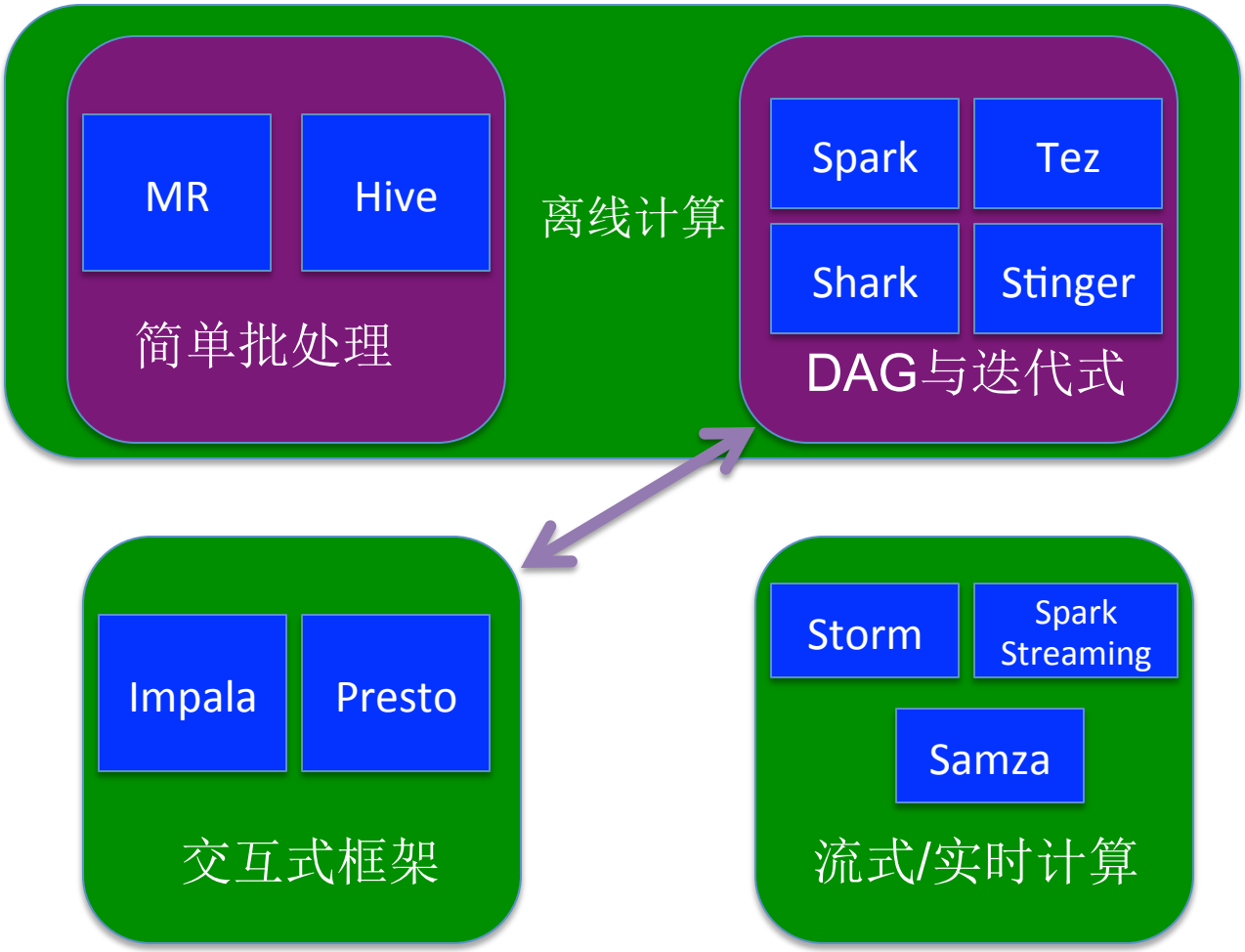
- 批处理计算
  - ✓ 对时间没有严格要求，吞吐率要高
- 迭代式与DAG计算
  - ✓ 机器学习算法
- 交互式计算
  - ✓ 支持类SQL语言，快速进行数据分析
- 流式计算
  - ✓ 数据像流水一样进入系统，需实时对其处理和分析

计算框架分类

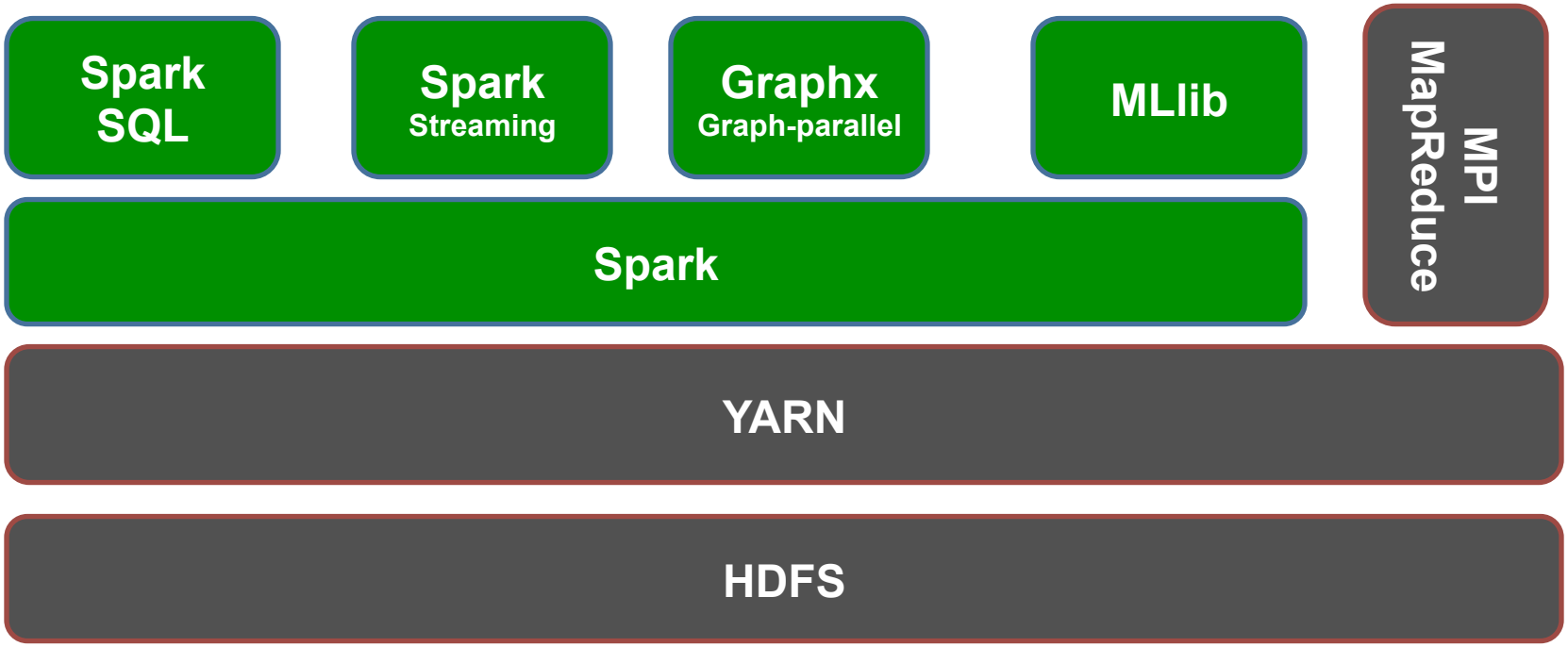
Real-Time	Interactive	Non-Interactive	Batch
✓ Online systems ✓ R-T analytics ✓ CEP	✓ Parameterized Report ✓ Drilldown ✓ Visualization ✓ Exploration	✓ Data preparation ✓ Incremental batch Processing ✓ Dashboards /Scorecards	✓ Operational batch processing ✓ Enterprise Reports ✓ Data Mining
0~5s	5s~1m	1m~1h	1h+

【注】摘自Hortonworks PPT：“Stinger Initiative: Deep Dive”

计算框架分类



Spark生态系统



常见的错误观点举例

Baidu 百度

spark取代 spark取代hadoop

百度一下

spark取代**hadoop**

spark取代**mapreduce**

呼之欲出!比Spark快10倍的Hadoop3.0有哪些实用新特性? - OPEN ...

2016年6月3日 - Apache hadoop 项目组最新消息,hadoop3.x以后将会调整方案架构,将Mapreduce 基于内存+io+磁盘,共同处理数据。 其实最大改变的是hdfs,hdfs 通过最近bla...

www.open-open.com/lib/... - 百度快照 - 73%好评

hadoop3.0x 后要比spark快10倍!,hadoop3.0xspark_云计算 | 帮客之家

2015年2月6日 - hadoop3.0x 后要比spark快10倍!,hadoop3.0xspark Apache hadoop 项目组最新消息,hadoop3.x以后将会调整方案架构,将Mapreduce 基于内存+io+磁盘,共同...

www.bkjia.com/yjs/9534... - 百度快照 - 105条评价

Spark核心开发者:性能超Hadoop百倍&Spark:大数据的“电光石火” -...

2014年8月22日 - Spark核心开发者:性能超Hadoop百倍&Spark:大数据的“电光石火” ...我个人希望我的研究想法可以超越论文的阶段,所以Berkeley这几点十分吸引我。 最后...

blog.csdn.net/likika20... - 百度快照

主要内容

1 大数据技术框架

2 Hadoop与Spark生态系统

3 课程内容安排

课程内容简介

➢ 课程特色

- ✓ 以目前主流的、最新的Spark稳定版2.1.x为基础；
- ✓ 深入浅出地介绍Spark生态系统原理及应用，内容包括Spark各组件（Spark Core/SQL/Streaming/MLlib）基本原理、使用方法、实战经验以及在线演示；
- ✓ 本课程精心设计了若干实验案例，帮助大家在理解理论的基础上，亲手实践Spark。

➢ 基础要求

- ✓ 了解Linux基础知识，掌握Java或Scala语言基础，了解HDFS、YARN
- ✓ 项目构建工具maven
- ✓ 集成开发环境intellij idea(不要使用eclipse)
- ✓ 代码管理工具git

➢ 时间安排

- ✓ 在线直播，共9次
- ✓ 每周2次（周二、周四晚上20:00-22:00），个别课程时间变动会提前通知

课程内容：第一部分 Spark概述（1课时）

- Spark产生背景
- Spark 基本特点
- Spark版本演化
- Spark核心概念
  - ✓包括RDD, transformation, action, cache等
- Spark生态系统
  - ✓包括Spark生态系统构成，以及与Hadoop生态系统关系
- Spark在互联网公司中的地位与应用
  - ✓介绍当前互联网公司的Spark应用案例
- 本课程与Spark 2.0的关系
- Spark集群搭建
  - ✓包括测试集群搭建和生产环境中集群搭建方法，并亲手演示整个过程

课程内容：第二部分 Spark Core（共3课时）

2.1 Spark 程序设计与实战

- Spark运行模式介绍
- Spark开发环境构建
- 常见transformation与action用法
- 常见控制函数介绍
- 在线演示：简易电影受众分析系统

2.2 Spark 内部原理剖析与源码阅读

- Spark运行模式剖析
- Spark运行流程剖析
- Spark shuffle剖析
- Spark 源码阅读

2.3 Spark 程序调优

- 数据存储格式调优
- 资源调优
- 程序参数调优
- 程序实现调优

课程内容：第三部分 Spark SQL 2.0(共2课时)

3.1 Spark SQL基本原理

- Spark SQL是什么
- Spark SQL基本原理
- Spark Dataframe与DataSets
- Spark SQL与Spark Core的关系

3.2 Spark SQL程序设计与应用案例

- Spark SQL程序设计
- 如何访问MySQL、HDFS等数据源，如何处理parquet格式数据
- 常用的DSL语法有哪些，如何使用
- Spark SQL应用案例：篮球运动员评估系统
  - ✓ 背景介绍
  - ✓ 数据导入
  - ✓ 数据分析
  - ✓ 结论

课程内容：第四部分 Spark Streaming(共1课时)

4.1 Spark Streaming基本原理

- Spark Streaming是什么
- Spark Streaming基本原理
- Structured Streaming
- Spark Streaming 编程接口介绍
- Spark Streaming应用案例

4.2 Spark Streaming程序设计

- 常见流式数据处理模式
- Spark Streaming与Kafka 交互
- Spark Streaming与Redis交互
- Spark Streaming部署与运行

课程内容：第五部分 Spark MLlib(共1课时)

- Spark MLlib简介
- 数据表示方式
- MLlib中的聚类、分类和推荐算法
- 如何使用MLlib的算法
- MLlib 2.0实践

课程内容：第六部分 综合实例(共1课时)

- 背景介绍
- 什么是Lambda architecture
- 利用HDFS+Spark Core+MLlib+Redis构建批处理线
- 利用Kafka+Spark Streaming+Redis构建实时处理线

课程难度预警

课程	内容	难度
1	Spark 2.1概述	
2	Spark程序设计与实战	
3	Spark内部原理剖析与源码阅读	
4	Spark程序调优	
5	Spark SQL基本原理与程序设计	
6	Spark SQL高级程序设计与应用案例	
7	Spark Streaming基本原理与程序设计	
8	Spark MLlib	
9	综合案例分析	

课程互动

- 技术论坛
  - ✓ http://wenda.chinahadoop.cn/topic/spark
  - ✓ 每节课对应一个问答帖子（课前公布链接），大家跟帖提问，每人可问多个问题，别人可点赞投票
- 调查问卷
  - ✓ 每节课结束，会有关于本节课的调查问卷（课后公布链接）
- 在线互动
  - ✓ 每节课最后三十分钟为互动时间，主要回答大家的问题
- 推荐阅读资料
  - ✓ 每节课会针对本节课内容推荐课后学习资料

预期收益

- 利用Spark解决大数据处理问题
- 具备一定的Spark程序调优技能
- 学习一些通用的大数据学习思路与方法

Keep In Mind

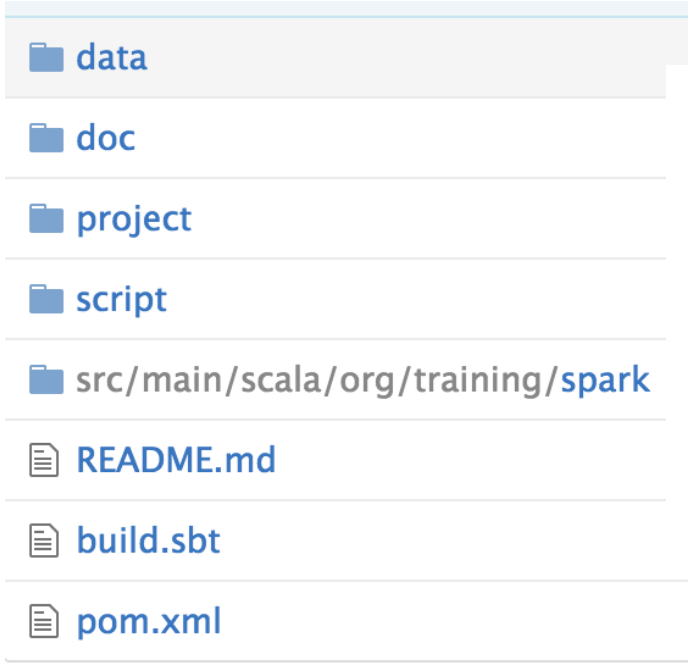
- Spark不是一门孤立的技术
  - ✓经常与Hadoop（HDFS/YARN/HBase等）一起使用
  - ✓Scala/Java/Python语言
  - ✓Maven/SBT项目构建工具
- Spark在不断发展过程中
  - ✓变化最多的是spark内核
  - ✓API层很少变动，即程序设计者不需要不断修改程序
- 持续学习
  - ✓这门课只是你你学习Spark的一个开始

Keep In Mind

适当调整你对本课程最终收益的预期！

本课程所有代码示例下载

- **git clone https://github.com/XichengDong/sparktraining**
  - ✓ 包含20+个实例，涉及Spark Core, Streaming, SQL以及MLlib等
  - ✓ 包含数据集、说明文档、运行脚本等
  - ✓ 所有程序均可直接在IDE中运行



hadoop123：董西成的微信公众号

专注于Hadoop/spark等大数据相关技术的分享

小象科技
ChinaHadoop.cn

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop

让你的数据产生价值！