

# 《图解Spark：核心技术与案例实战》

## 图书基本信息

书名：《图解Spark：核心技术与案例实战》

13位ISBN编号：9787121302365

出版时间：2017-1

作者：郭景瞻

页数：480

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《图解Spark：核心技术与案例实战》

## 内容概要

《图解Spark：核心技术与案例实战》以Spark 2.0版本为基础进行编写，全面介绍了Spark核心及其生态圈组件技术。主要内容包括Spark生态圈、实战环境搭建、编程模型和内部重要模块的分析，重点介绍了消息通信框架、作业调度、容错执行、监控管理、存储管理以及运行框架，同时还介绍了Spark生态圈相关组件，包括Spark SQL的即席查询、Spark Streaming的实时流处理应用、MLbase/MLlib的机器学习、GraphX的图处理、SparkR的数学计算和Alluxio的分布式内存文件系统等。

《图解Spark：核心技术与案例实战》从Spark核心技术进行深入分析，重要章节会结合源代码解读其实现原理，围绕着技术原理介绍了相关典型实例，读者通过这些实例可以更加深入地理解Spark的运行机制。另外《图解Spark：核心技术与案例实战》还应用了大量的图表进行说明，让读者能够更加直观地理解Spark相关原理。通过《图解Spark：核心技术与案例实战》，读者将能够很快地熟悉和掌握Spark大数据分析计算的利器，在生产中解决实际问题。

## 书籍目录

### 第一篇 基础篇

#### 第1章 Spark及其生态圈概述

##### 1.1 Spark简介

###### 1.1.1 什么是Spark

###### 1.1.2 Spark与MapReduce比较

###### 1.1.3 Spark的演进路线图

##### 1.2 Spark生态系统

###### 1.2.1 Spark Core

###### 1.2.2 Spark Streaming

###### 1.2.3 Spark SQL

###### 1.2.4 BlinkDB

###### 1.2.5 MLBase/MLlib

###### 1.2.6 GraphX

###### 1.2.7 SparkR

###### 1.2.8 Alluxio

##### 1.3 小结

#### 第2章 搭建Spark实战环境

##### 2.1 基础环境搭建

###### 2.1.1 搭建集群样板机

###### 2.1.2 配置集群环境

##### 2.2 编译Spark源代码

###### 2.2.1 配置Spark编译环境

###### 2.2.2 使用Maven编译Spark

###### 2.2.3 使用SBT编译Spark

###### 2.2.4 生成Spark部署包

##### 2.3 搭建Spark运行集群

###### 2.3.1 修改配置文件

###### 2.3.2 启动Spark

###### 2.3.3 验证启动

###### 2.3.4 第一个实例

##### 2.4 搭建Spark实战开发环境

###### 2.4.1 CentOS中部署IDEA

###### 2.4.2 使用IDEA开发程序

###### 2.4.3 使用IDEA阅读源代码

##### 2.5 小结

### 第二篇 核心篇

#### 第3章 Spark编程模型

##### 3.1 RDD概述

###### 3.1.1 背景

###### 3.1.2 RDD简介

###### 3.1.3 RDD的类型

##### 3.2 RDD的实现

###### 3.2.1 作业调度

###### 3.2.2 解析器集成

###### 3.2.3 内存管理

###### 3.2.4 检查点支持

###### 3.2.5 多用户管理

## 3.3 编程接口

### 3.3.1 RDD分区 (Partitions)

### 3.3.2 RDD首选位置 (PreferredLocations)

### 3.3.3 RDD依赖关系 (Dependencies)

### 3.3.4 RDD分区计算 (Iterator)

### 3.3.5 RDD分区函数 (Partitioner)

## 3.4 创建操作

### 3.4.1 并行化集合创建操作

### 3.4.2 外部存储创建操作

## 3.5 转换操作

### 3.5.1 基础转换操作

### 3.5.2 键值转换操作

## 3.6 控制操作

## 3.7 行动操作

### 3.7.1 集合标量行动操作

### 3.7.2 存储行动操作

## 3.8 小结

## 第4章 Spark核心原理

### 4.1 消息通信原理

#### 4.1.1 Spark消息通信架构

#### 4.1.2 Spark启动消息通信

#### 4.1.3 Spark运行时消息通信

### 4.2 作业执行原理

#### 4.2.1 概述

#### 4.2.2 提交作业

#### 4.2.3 划分调度阶段

#### 4.2.4 提交调度阶段

#### 4.2.5 提交任务

#### 4.2.6 执行任务

#### 4.2.7 获取执行结果

### 4.3 调度算法

#### 4.3.1 应用程序之间

#### 4.3.2 作业及调度阶段之间

#### 4.3.3 任务之间

### 4.4 容错及HA

#### 4.4.1 Executor异常

#### 4.4.2 Worker异常

#### 4.4.3 Master异常

### 4.5 监控管理

#### 4.5.1 UI监控

#### 4.5.2 Metrics

#### 4.5.3 REST

### 4.6 实例演示

#### 4.6.1 计算年降水实例

#### 4.6.2 HA配置实例

## 4.7 小结

## 第5章 Spark存储原理

### 5.1 存储分析

#### 5.1.1 整体架构

- 5.1.2 存储级别
- 5.1.3 RDD存储调用
- 5.1.4 读数据过程
- 5.1.5 写数据过程
- 5.2 Shuffle分析
  - 5.2.1 Shuffle简介
  - 5.2.2 Shuffle的写操作
  - 5.2.3 Shuffle的读操作
- 5.3 序列化和压缩
  - 5.3.1 序列化
  - 5.3.2 压缩
- 5.4 共享变量
  - 5.4.1 广播变量
  - 5.4.2 累加器
- 5.5 实例演示
- 5.6 小结
- 第6章 Spark运行架构
  - 6.1 运行架构总体介绍
    - 6.1.1 总体介绍
    - 6.1.2 重要类介绍
  - 6.2 本地（Local）运行模式
    - 6.2.1 运行模式介绍
    - 6.2.2 实现原理
  - 6.3 伪分布（Local-Cluster）运行模式
    - 6.3.1 运行模式介绍
    - 6.3.2 实现原理
  - 6.4 独立（Standalone）运行模式
    - 6.4.1 运行模式介绍
    - 6.4.2 实现原理
  - 6.5 YARN运行模式
    - 6.5.1 YARN运行框架
    - 6.5.2 YARN-Client运行模式介绍
    - 6.5.3 YARN-Client 运行模式实现原理
    - 6.5.4 YARN-Cluster运行模式介绍
    - 6.5.5 YARN-Cluster 运行模式实现原理
    - 6.5.6 YARN-Client与YARN-Cluster对比
  - 6.6 Mesos运行模式
    - 6.6.1 Mesos介绍
    - 6.6.2 粗粒度运行模式介绍
    - 6.6.3 粗粒度实现原理
    - 6.6.4 细粒度运行模式介绍
    - 6.6.5 细粒度实现原理
    - 6.6.6 Mesos粗粒度和Mesos细粒度对比
  - 6.7 实例演示
    - 6.7.1 独立运行模式实例
    - 6.7.2 YARN-Client实例
    - 6.7.3 YARN-Cluster实例
  - 6.8 小结

## 第三篇 组件篇

### 第7章 Spark SQL

#### 7.1 Spark SQL简介

##### 7.1.1 Spark SQL发展历史

##### 7.1.2 DataFrame/Dataset介绍

#### 7.2 Spark SQL运行原理

##### 7.2.1 通用SQL执行原理

##### 7.2.2 SparkSQL运行架构

##### 7.2.3 SQLContext运行原理分析

##### 7.2.4 HiveContext介绍

#### 7.3 使用Hive-Console

##### 7.3.1 编译Hive-Console

##### 7.3.2 查看执行计划

##### 7.3.3 应用Hive-Console

#### 7.4 使用SQLConsole

##### 7.4.1 启动HDFS和Spark Shell

##### 7.4.2 与RDD交互操作

##### 7.4.3 读取JSON格式数据

##### 7.4.4 读取Parquet格式数据

##### 7.4.5 缓存演示

##### 7.4.6 DSL演示

#### 7.5 使用Spark SQL CLI

##### 7.5.1 配置并启动Spark SQL CLI

##### 7.5.2 实战Spark SQL CLI

#### 7.6 使用Thrift Server

##### 7.6.1 配置并启动Thrift Server

##### 7.6.2 基本操作

##### 7.6.3 交易数据实例

##### 7.6.4 使用IDEA开发实例

#### 7.7 实例演示

##### 7.7.1 销售数据分类实例

##### 7.7.2 网店销售数据统计

#### 7.8 小结

### 第8章 Spark Streaming

#### 8.1 Spark Streaming简介

##### 8.1.1 术语定义

##### 8.1.2 Spark Streaming特点

#### 8.2 Spark Streaming编程模型

##### 8.2.1 DStream的输入源

##### 8.2.2 DStream的操作

#### 8.3 Spark Streaming运行架构

##### 8.3.1 运行架构

##### 8.3.2 消息通信

##### 8.3.3 Receiver分发

##### 8.3.4 容错性

#### 8.4 Spark Streaming运行原理

##### 8.4.1 启动流处理引擎

##### 8.4.2 接收及存储流数据

##### 8.4.3 数据处理

## 8.5 实例演示

### 8.5.1 流数据模拟器

### 8.5.2 销售数据统计实例

### 8.5.3 Spark Streaming+Kafka实例

## 8.6 小结

## 第9章 Spark MLlib

### 9.1 Spark MLlib简介

#### 9.1.1 Spark MLlib介绍

#### 9.1.2 Spark MLlib数据类型

#### 9.1.3 Spark MLlib基本统计方法

#### 9.1.4 预言模型标记语言

### 9.2 线性模型

#### 9.2.1 数学公式

#### 9.2.2 线性回归

#### 9.2.3 线性支持向量机

#### 9.2.4 逻辑回归

#### 9.2.5 线性最小二乘法、Lasso和岭回归

#### 9.2.6 流式线性回归

### 9.3 决策树

### 9.4 决策模型组合

#### 9.4.1 随机森林

#### 9.4.2 梯度提升决策树

### 9.5 朴素贝叶斯

### 9.6 协同过滤

### 9.7 聚类

#### 9.7.1 K-means

#### 9.7.2 高斯混合

#### 9.7.3 快速迭代聚类

#### 9.7.4 LDA

#### 9.7.5 二分K-means

#### 9.7.6 流式K-means

### 9.8 降维

#### 9.8.1 奇异值分解降维

#### 9.8.2 主成分分析降维

### 9.9 特征提取和变换

#### 9.9.1 词频—逆文档频率

#### 9.9.2 词向量化工具

#### 9.9.3 标准化

#### 9.9.4 范数化

### 9.10 频繁模式挖掘

#### 9.10.1 频繁模式增长

#### 9.10.2 关联规则挖掘

#### 9.10.3 PrefixSpan

### 9.11 实例演示

#### 9.11.1 K-means聚类算法实例

#### 9.11.2 手机短信分类实例

## 9.12 小结

## 第10章 Spark GraphX

### 10.1 GraphX介绍

- 10.1.1 图计算
- 10.1.2 GraphX介绍
- 10.1.3 发展历程
- 10.2 GraphX实现分析
  - 10.2.1 GraphX图数据模型
  - 10.2.2 GraphX图数据存储
  - 10.2.3 GraphX图切分策略
  - 10.2.4 GraphX图操作
- 10.3 实例演示
  - 10.3.1 图例演示
  - 10.3.2 社区发现演示
- 10.4 小结
- 第11章 SparkR
  - 11.1 概述
    - 11.1.1 R语言介绍
    - 11.1.2 SparkR介绍
  - 11.2 SparkR与DataFrame
    - 11.2.1 DataFrames介绍
    - 11.2.2 与DataFrame的相关操作
  - 11.3 编译安装SparkR
    - 11.3.1 编译安装R语言
    - 11.3.2 安装SparkR运行环境
    - 11.3.3 安装SparkR
    - 11.3.4 启动并验证安装
  - 11.4 实例演示
  - 11.5 小结
- 第12章 Alluxio
  - 12.1 Alluxio简介
    - 12.1.1 Alluxio介绍
    - 12.1.2 Alluxio系统架构
    - 12.1.3 HDFS与Alluxio
  - 12.2 Alluxio编译部署
    - 12.2.1 编译Alluxio
    - 12.2.2 单机部署Alluxio
    - 12.2.3 集群模式部署Alluxio
  - 12.3 Alluxio命令行使用
    - 12.3.1 接口说明
    - 12.3.2 接口操作示例
  - 12.4 实例演示
    - 12.4.1 启动环境
    - 12.4.2 Alluxio上运行Spark
    - 12.4.3 Alluxio上运行MapReduce
  - 12.5 小结



# 《图解Spark：核心技术与案例实战》

## 精彩短评

- 1、核心篇较为丰富，组件篇内容略泛泛
- 2、我想看的写得太多，不想看的写得太多
- 3、排版和语言组织都让人迷糊 中国人写的东西看着像机器翻译的外国论文 也没说清楚知识

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)