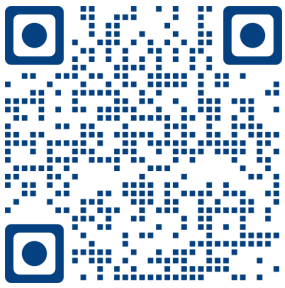
# Spark原理与实践

陈一帅

yschen@bjtu.edu.cn

北京交通大学电子信息工程学院网络智能实验室



课程源自Databrick官方教程，搭配 Piotrszul Spark入门练习代码 （Github），采用 华为网络人工智能引擎在线实验环境，是 Spark 大数据研发的入门课。每个视频几分钟，一路下来，带大家在动手中，走上大数据研发的职业道路。详细课程信息请访问：https://yishuai.github.io/spark

## 目录

- 实验平台
- Jupyter Notebook
- NAIE文件上传技巧

# A. Spark

## 一、优点和特性

Spark是目前最流行的大数据计算平台。我们首先简单了解一下它的各项优点和特性。

B站视频

课程PPT：PDF（2MB）

## 二、基于内存的数据分享

Spark极大地提高了大数据计算的速度，这来自于它基于内存的数据分享机制。本节比较各种存储介质访问的性能差异，获得对Spark基于内存的数据分享机制的理解。

B站视频

## 三、Spark的内核和组件

本节介绍Spark的内核和各组成模块，包括Spark SQL、数据流、机器学习库和图计算库

B站视频

# B. RDD

## 四、弹性分布式数据集（RDD）

弹性分布式数据集（RDD）是Spark的核心数据结构。它是只读的、分布式的、容错的。对它的操作是Lazy（懒）的。理解RDD是理解Spark工作机制的关键。本节介绍它的各项核心设计决策背后的思想，请一定细心领会。

B站视频

课程PPT：PDF（3.4MB）

## 五、分区并行机制

Spark通过将RDD进行分区，然后在各分区上并行计算，实现高性能分布式计算。本节介绍分区的原理、性能和优化方法。本节内容对优化Spark性能非常重要，请细心领会。

B站视频

## 六、RDD操作方法

Spark提供了基于Map-Reduce计算范式的各种RDD操作方法，同时设计了Lazy的操作执行方式。本节介绍它提供的各种操作函数，并通过实例讲解这些函数的使用，以及Lazy方式带来的好处。本节对Spark编程非常重要。

B站视频

## 七、RDD编程练习

本节基于华为NAIE Spark在线练习平台，练习Piotrszul Spark入门代码的第一课：RDD基础。这是一个简单的文本处理例子，帮助你熟悉RDD编程的基本流程和函数。实验手册如下：

B站视频

实验手册1：Spark环境配置和使用，PPT（941KB）

实验手册2：RDD和DataFrame练习，PPT（60KB）

## 八、MIT RDD编程示例

MIT（麻省理工学院）经典课程 6.824 分布式系统中，有一节Spark课程。本节我们一起来看这个课程中讲解的网页PageRank RDD编程实例，进一步提高我们对RDD编程的认识。

B站视频

# C. DataFrame

## 九、DataFrame编程

实际中，Spark提供了SQL和DataFrame编程接口，让我们能够快速上手大数据编程。本节介绍这些编程接口，然后基于华为NAIE Spark在线大数据实验平台，练习Piotrszul Spark入门练习代码的第二课：结构化数据编程。学习了这一课，你就可以开始你的Spark编程之旅了。实验手册如下。Enjoy!

B站视频

PPT（PDF，3.4MB）

## 十、DataFrame入门示例

本节介绍Piotrszul Spark入门练习代码中的DataFrame入门代码。 DataFrame在Spark大数据编程中非常重要，请一定要看看哦。

B站视频

## 十一、Python数据处理

本节介绍用Python进行数据分析和处理时常用到的几个库：Pandas，NumPy，SciPy，Matplotlib。学会了它们，你就可以用Python进行基本的数据分析了哦，快快学习吧。

B站视频

# D. 组件

## 十二、Streaming流式计算

Streaming流式计算能及时给出分析结果，帮助决策、管理和控制，因此应用越来越广。本节介绍Spark流式计算的基本原理，并讲解一个代码。

B站视频

PPT（PDF，1.7MB）

## 十三、机器学习

人工智能时代，机器学习是同学们必须掌握的技能。本节介绍Spark机器学习相关的各种函数，并定位到各个实验中。通过本节及其配套实验的练习，你将打开机器学习的大门。加油！我们还提供了一个基于Python机器学习库SciKit Learn的Python机器学习入门指南，它能帮助你快速上手Python机器学习，快来看看吧。

B站视频

PPT（PDF，1.1MB）

张璇，Python机器学习入门指南：Doc（180KB），PDF（343KB）

# E. 实验

## 十四、实验介绍

本节我们将提供给大家一系列的实验，包括数据和代码，供大家练习。掌握Spark编程的秘诀是什么呢？就是编。好好做好这些实验吧。

B站视频

实验手册3：Spark 大数据练习，PPT（64KB）

## 十五、单词计数示例

本节介绍利用 RDD 和 DataFrame 对文本中的单词进行筛选、频率统计、排序、和结果保存。这项工作在大数据编程中经常遇到，请一定要看看哦。

B站视频

B站视频

Pandas函数速查表（PDF，180KB）

# F. 实验平台

## 十六、实验平台

Spark是目前大数据系统的主流平台。安装Spark非常简单，你也可以利用很多在线的Spark实验环境，快来看看吧。

B站视频

课程PPT：PPT（1MB）

## 十七、Jupyter Notebook

本节以华为NAIE Python/PySpark 在线实验平台为例，讲解如何使用 Jupyter Notebook 这个超级友好的工具，在浏览器中进行Python编程。Jupyter Notebook 太好用了，你一定会喜欢它的。快来看看吧。

B站视频

## 十八、NAIE文件上传技巧

本节介绍在华为NAIE平台上传文件后，找到该文件，确定它的路径的方法。很多同学在操作中遇到这一难题，请一定要看看哦。

B站视频