**Spark 프로그래밍 기초 강의 스크립트**

## Slide 1: 강의 소개

안녕하세요, 여러분. 저는 이번 학기 Spark 프로그래밍 기초를 담당하게 된 이양원 교수입니다.

제 연락처는 Slide에 나와있는 이메일 주소로 연락 주시면 됩니다. 또한 제 개인 홈페이지도 운영하고 있으니, 추가 자료나 공지사항은 그곳에서도 확인하실 수 있습니다.

오늘부터 우리는 현대 빅데이터 처리의 핵심 기술인 Spark에 대해 깊이 있게 배워보겠습니다. 이 과정은 제0장부터 시작하는데, 이는 본격적인 내용에 앞서 여러분이 꼭 알아야 할 기초 내용들을 다루기 위함입니다.

**Spark编程基础讲义脚本**

## Slide1：课程介绍

大家好，我是本学期担任Spark编程基础课程的李阳源教授。

我的联系方式在幻灯片上显示的电子邮件地址，大家可以通过这个联系我。
另外，我的另一个联系方式是微信。
下面有我的微信二维码。
通过这个我们可以保持联系。

从今天开始，我们将深入学习现代大数据处理的核心技术——Spark。本课程从第0章开始，这是为了在正式内容之前，先讲解大家必须了解的基础内容。

## Slide 2: 선수 과목

본격적으로 Spark를 배우기 전에, 여러분이 이미 가지고 있어야 할 배경 지식에 대해 먼저 말씀드리겠습니다.
첫째, **빅데이터 기술 개론**을 이미 수강하셨어야 합니다. 빅데이터가 무엇인지, 왜 중요한지에 대한 전반적인 이해가 있어야 Spark의 필요성과 가치를 제대로 이해할 수 있습니다.

둘째, **데이터베이스 시스템 원리**를 아셔야 합니다. 특히 관계형 데이터베이스와 SQL 문법에 대한 이해가 필수적입니다. Spark SQL을 배울 때 이 지식이 매우 중요하게 작용합니다. SQL을 아신다면 Spark SQL로의 전환이 훨씬 수월할 겁니다.

셋째, **프로그래밍 능력**입니다. Python이나 Java 중 하나는 어느 정도 다룰 수 있어야 합니다. 특히 객체지향 프로그래밍의 기본 개념들 - 클래스, 객체, 메소드, 상속 같은 개념들을 알고 있어야 합니다.

본 과정에서는 주로 Python을 사용할 예정인데, Python의 기본 문법을 모르시는 분은 미리 복습해 오시기 바랍니다.
혹시 이 중에서 부족한 부분이 있다면, 지금이라도 복습하시거나 제게 개별적으로 상담을 요청하시기 바랍니다.

## Slide2：先修课程

在正式学习Spark之前，我先说明一下大家应该已经具备的背景知识。

第一，应该已经修过大数据技术导论(dǎo lùn, introduction)。需要对大数据是什么、
为什么重要有全面的理解，这样才能正确理解Spark的必要性和价值。

第二，需要了解数据库系统原理。特别是对关系型数据库和SQL语法的理解是必不可少的。在学习Spark SQL时，这个知识非常重要。如果你懂SQL，转换到Spark SQL会容易得多。

第三，编程能力。Python或Java至少要能熟练使用其中一种。特别是要了解面向对象编程的基本概念——类、对象、方法、继承(jìchéng)等概念。

本课程主要使用Python，不熟悉Python基本语法的同学请提前复习。

如果这些内容中有不足的地方，现在也可以复习，或者单独向我咨询(**zī xún**)。

## Slide 3: 과정 특색

자, 여러분이 보시는 이 화려한 워드 클라우드는 빅데이터의 세계를 상징적으로 보여주고 있습니다. VOLUME, PROCESSING, STORAGE, SPEED, VARIETY... 이 모든 키워드들이 오늘날의 데이터 시대를 대표하는 개념들입니다.

이 과정의 특색은 바로 **"빅데이터의 문을 여는 것"**입니다.

많은 학생들이 빅데이터, 특히 Spark를 어려워합니다. 왜일까요? 개념도 복잡하고, 설치도 까다롭고, 실습 환경 구축도 쉽지 않기 때문입니다.

하지만 걱정하지 마세요. 이 과정은 여러분이 Spark 학습의 문턱을 넘는 데 필요한 모든 장애물을 제거하고, 학습의 길을 최대한 평탄하게 만드는 것을 목표로 설계되었습니다.

실습 환경 구축부터 시작해서, 기본 개념, 프로그래밍, 실전 응용까지 단계적으로 차근차근 배워나갈 것입니다.

여러분이 이 과정을 마칠 때쯤이면, Spark는 더 이상 어려운 기술이 아닌, 여러분의 강력한 도구가 되어 있을 겁니다.

## Slide3：课程特色

好，大家看到的这个华丽的词云，象征性地展示了大数据的世界。VOLUME、PROCESSING、STORAGE、SPEED、VARIETY……
这些关键词都代表了当今数据时代的核心概念。

本课程的特色就是**"开启(kāi qǐ)大数据之门"**。

很多学生觉得大数据，特别是Spark很难。为什么呢？因为概念复杂、安装繁琐、实验环境搭建也不容易。

但是不用担心。本课程的目标就是消除**(xiāochú)**大家跨越**(kuàyuè)**Spark学习门槛**(ménkǎn)**所需的所有障碍，尽可能让学习之路平坦顺畅**(shùnchàng)**。

从实验环境搭建开始，到基本概念、编程、实战应用，我们会循序渐进**(xúnxù jiànjìn)**、逐步**(zhúbù)**学习。

等大家完成这门课程时，Spark就不再是一项困难的技术，而是成为你们强大的工具。

## Slide 4: 과정 내용 개요

이제 구체적으로 무엇을 배울 것인지 말씀드리겠습니다.
먼저, **Spark는 현재 가장 인기 있는 빅데이터 처리 기술**입니다. Hadoop의 뒤를 이어 빅데이터 생태계의 사실상 표준이 되었죠. 왜 그럴까요? 속도가 빠르고, 사용하기 쉽고, 다양한 용도로 활용할 수 있기 때문입니다.

본 과정에서는 **Spark의 기술 원리와 프로그래밍 방법**을 중점적으로 다룹니다. 단순히 사용법만 배우는 것이 아니라, 내부적으로 어떻게 작동하는지 이해하는 것이 중요합니다. 그래야 문제가 생겼을 때 해결할 수 있고, 성능을 최적화할 수 있습니다.
Spark는 여러 언어를 지원합니다 - Scala, Java, Python, R. 이 중에서 **우리는 Python을 사용**할 것입니다. 왜 Python일까요? 문법이 간결하고, 데이터 과학 분야에서 가장 널리 쓰이며, 배우기가 상대적으로 쉽기 때문입니다.
과정 내용을 보면:

- **Spark 소개와 설계 원리** - Spark가 무엇이고, 왜 빠른지
- **실험 환경 구축** - 여러분의 컴퓨터에서 Spark를 실행할 수 있도록
- **RDD 프로그래밍** - Spark의 가장 기본이 되는 데이터 구조
- **Spark SQL** - SQL처럼 데이터를 쿼리하는 방법
- **Structured Streaming** - 실시간 데이터 처리
- **Spark MLlib** - 머신러닝 라이브러리

이 과정을 마치면 여러분은 **기초적인 Spark 애플리케이션을 독자적으로 개발할 수 있는 능력**을 갖추게 됩니다. 이는 여러분이 실제 연구나 업무에서 빅데이터 문제를 해결하는 데 직접 활용할 수 있는 실용적인 능력입니다.

## Slide4：课程内容概要

现在我来具体说明一下我们要学什么。
首先，Spark是目前最热门的大数据处理技术。它继Hadoop之后，成为了大数据生态系统的事实标准。为什么呢？因为速度快、易于使用、可以应用于多种用途。

本课程重点讲解Spark的技术原理和编程方法。不仅仅是学习使用方法，理解其内部运作机制也很重要。这样遇到问题时才能解决，也能优化性能。

Spark支持多种语言——Scala、Java、Python、R。其中我们使用Python。为什么用Python呢？因为语法简洁**(jiǎn jié)**，在数据科学领域应用最广泛，相对容易学习。

课程内容包括：

- Spark介绍和设计原理 - Spark是什么，为什么快
- 实验环境搭建 - 让大家能在自己的电脑上运行Spark
- RDD编程 - Spark最基本的数据结构
- Spark SQL - 像SQL一样查询数据的方法
- Structured Streaming - 实时数据处理
- Spark MLlib - 机器学习库

完成本课程后，大家将具备独立开发基础Spark应用程序的能力。这是可以直接应用于实际研究或工作中解决大数据问题的实用能力。

**Slide 5: 교학 대강 (강의 계획)**
자, 이제 학기 전체 일정을 한번 살펴보겠습니다. 총 32시간의 수업으로 구성되어 있습니다.

**제1장 - 빅데이터 기술 개술 (2시간)**
빅데이터의 기본 개념과 핵심 기술들을 복습합니다. Hadoop, Spark, NoSQL 같은 대표적인 기술들을 개괄적으로 살펴봅니다.

**제2장 - Spark의 설계와 실행 원리 (18시간)**
Spark의 내부 아키텍처를 이해합니다. 클러스터 매니저, 드라이버, 익스큐터가 어떻게 협력하는지, RDD가 어떻게 설계되고 실행되는지 배웁니다. 이론적이지만 매우 중요한 부분입니다.

**제3장 - 빅데이터 실험 환경 구축 (12시간)**
여기서 좀 시간을 투자합니다. Linux 시스템, Hadoop, MySQL, Kafka, Anaconda까지 하나씩 설치해봅니다. 처음엔 복잡해 보이지만, 한 번 제대로 설치해두면 이후 실습이 훨씬 수월합니다. 실습 시간에 제가 직접 도와드릴 겁니다.

**제4장 - Spark 환경 구축과 사용 방법 (4시간)** Spark를 설치하고, spark-shell에서 간단한 코드를 실행해봅니다. 그리고 독립적인 Spark 애플리케이션을 만드는 방법도 배웁니다.

**Slide5：教学大纲（授课计划）**
好，现在我们来看一下整个学期的日程安排(**ān pái)**。总共由96个学时的课程构成。

**第1章 - 大数据技术概述（2学时）**
复习大数据的基本概念和核心技术。概括性地了解Hadoop、Spark、NoSQL等代表性技术。

**第2章 - Spark的设计与运行原理（18学时）**
理解Spark的内部架构。学习集群管理器、驱动程序、执行器如何协作，RDD如何设计和执行。虽然是理论内容，但非常重要。

**第3章 - 大数据实验环境搭建（12学时）**
这里我们要投入一些时间。逐一(**zhú yī)**安装Linux系统、Hadoop、MySQL、Kafka、Anaconda。刚开始看起来复杂，但一旦 **(yí dàn)**正确安装好，之后的实验就会顺利得多。实验课上我会直接帮助大家。

**第4章 - Spark环境搭建和使用方法（8学时）**
安装Spark，在spark-shell中运行简单的代码。还要学习创建独立Spark应用程序的方法。

**제5장 - RDD 프로그래밍 (4시간)** Spark의 핵심입니다.
RDD가 무엇인지, transformation과 action의 차이는 무엇인지, 키-값 쌍은 어떻게 다루는지 배웁니다. 많은 실습 예제를 통해 익숙해지도록 하겠습니다.

**제6장 - Spark SQL (4시간)** DataFrame을 사용해서 SQL처럼 데이터를 다룹니다. Parquet 파일 읽고 쓰기, 데이터베이스 연결, pandas와의 통합까지 다룹니다. 데이터 분석가들이 가장 많이 사용하는 부분입니다.

**제7장 - Spark Streaming (4시간)** Spark Streaming 이것은 원래 실시간으로 들어오는 데이터 처리를 지원했었으나 현재는 Sturctured Streaming이 이것을 대신하므로 이번 수업 과정에서는 생략하도록 하겠습니다.

**제8장 - Structured Streaming (4시간)** 더 최신 API인 Structured Streaming을 배웁니다. DataFrame 기반이라 좀 더 쉽고 강력합니다. 장애 허용, 지연 데이터 처리 같은 실전 주제들도 다룹니다.

**제9장 - Spark MLlib (4시간)** 마지막으로 머신러닝입니다. spark.mllib와 spark.ml 두 라이브러리를 배우고, 실제 머신러닝 파이프라인을 구축해봅니다.

이와 같이 계획하고 있으나 주어진 시간 동안 최선을 다하겠으나 여러분이 협조하지 않으면 아마 중간에서 그만 둘수도 있어요. 우리 모두 힘내서 열심히 해 봅시다.



**第5章 - RDD编程（20学时）**这是Spark的核心。
学习RDD是什么，transformation和action的区别是什么，如何处理键值对。通过大量实验例子让大家熟悉。

**第6章 - Spark SQL（20学时）**
使用DataFrame像SQL一样处理数据。学习Parquet文件的读写、数据库连接、与pandas的整合。这是数据分析师最常用的部分。

**第7章 - Spark Streaming（4学时）**
Spark Streaming 最初支持实时数据流处理，但由于目前已被 Structured Streaming 取代，因此本课程中将不再涵盖这部分内容。

**第8章 - Structured Streaming（10学时）**
学习更新的API——Structured Streaming。基于DataFrame，所以更简单更强大。还会讲容错处理、延迟数据处理等实战主题。

**第9章 - Spark MLlib（6学时）**
最后是机器学习。学习spark.mllib和spark.ml两个库，构建实际的机器学习管道。

虽然计划如此，我也会在给定的时间内全力以赴(**yǐ fù**)，但如果大家不配合的话，可能中途就不得不放弃了。让我们一起加油努力吧！

## Slide 6: 과정 배치 교학 자원 - 실험

이론만으로는 부족합니다. 실제로 손으로 코드를 작성하고 실행해봐야 진짜 실력이 됩니다.

그래서 우리는 **총 7개의 실습 과제**를 준비했습니다.

**실험1 - Linux, Hadoop, MySQL 설치 및 사용** 기본 환경을 구축합니다. 이게 제대로 안 되면 나머지를 할 수 없으니 꼭 완료하세요.

**실험2 - Spark 설치 및 사용** Spark를 설치하고 기본 명령어들을 실행해봅니다.

**실험3 - RDD 프로그래밍 초급 실습** RDD의 기본 연산들을 직접 코딩해봅니다. map, filter, reduce 같은 핵심 연산들을 익힙니다.

**실험4 - Spark SQL 프로그래밍 초급 실습** DataFrame을 만들고 SQL 쿼리를 실행해봅니다.

**실험5 - Spark Streaming 프로그래밍 초급 실습** 간단한 스트리밍 애플리케이션을 만들어봅니다.

**실험6 - Structured Streaming 프로그래밍 실습** 더 현대적인 스트리밍 API를 사용해봅니다.

**실험7 - Spark MLlib 프로그래밍 초급 실습** 간단한 머신러닝 모델을 학습시키고 예측해봅니다.

각 실습마다 상세한 가이드가 있으니 차근차근 따라하시면 됩니다. 물론 실습 시간에 막히는 부분이 있으면 언제든 질문하세요.

## Slide6：课程配套教学资源 - 实验

光有理论是不够的。必须亲手编写代码并运行，才能真正掌握。

因此我们准备了**总共7个实验任务**。

**实验1 - Linux、Hadoop、MySQL安装和使用** 搭建基本环境。这个没做好，其他都做不了，所以一定要完成。

**实验2 - Spark安装和使用** 安装Spark并执行基本命令。

**实验3 - RDD编程初级实践** 直接编写RDD的基本操作。熟悉map、filter、reduce等核心操作。

**实验4 - Spark SQL编程初级实践** 创建DataFrame并执行SQL查询。

**实验5 - Structured Streaming编程实践** 使用更现代的流处理API。

**实验6 - Spark MLlib编程初级实践** 训练简单的机器学习模型并进行预测。

每个实验都有详细的指南，大家**按部就班(ànbùjiùbān)**地做就可以了。当然，实验时间遇到困难随时可以提问。

## Slide 7: 종합 실험 사례 1 - 타오바오 더블 11 데이터 분석

기본 실습 외에도, 우리는 **대형 종합 프로젝트**를 준비했습니다.

첫 번째는 **"타오바오 더블11 데이터 분석 및 예측"** 프로젝트입니다.
타오바오를 아시나요? 중국의 최대 전자상거래 플랫폼입니다. 그리고 더블11, 즉 11월 11일은 중국 최대의 쇼핑 축제일입니다. 우리나라의 블랙프라이데이보다 훨씬 큰 규모죠.
이 프로젝트는 **실제 타오바오 사용자 거래 데이터**를 사용합니다. 물론 익명화된 데이터입니다.
여러분이 배울 내용을 보면:

- **데이터 전처리** - 원시 데이터를 정제하고 변환하기
- **데이터 저장** - Hadoop HDFS에 저장하고 Hive로 관리하기
- **데이터 쿼리** - SQL로 필요한 정보 추출하기
- **데이터 시각화** - ECharts로 멋진 차트 만들기

오른쪽 다이어그램을 보시면 전체 데이터 파이프라인이 나와 있습니다:

1. 원시 거래 데이터를 수집
2. Hadoop에 저장하고 Hive로 구조화
3. Spark로 분석
4. MySQL에 결과 저장
5. ECharts로 시각화

이 과정에서 Linux, MySQL, Hadoop, Hive, Sqoop, Eclipse, ECharts, Spark 등 **빅데이터 생태계의 주요 도구들을 모두 사용**해봅니다.
이 프로젝트 하나만 완료해도 여러분의 이력서에 당당히 쓸 수 있는 실전 프로젝트가 됩니다.

## Slide7：综合实验案例1 -淘宝双11数据分析

除了基本实验，我们还准备了大型综合项目。

第一个是**"淘宝双11数据分析与预测"**项目。
大家知道淘宝吗？
中国最大的电子商务平台。
而双11，即11月11日，是中国最大的购物节。
这个项目使用真实的淘宝用户交易数据。
当然是匿名化的数据。
大家要学习的内容：

- 数据预处理 - 清洗和转换原始数据
- 数据存储 - 存储到Hadoop HDFS并用Hive管理
- 数据查询 - 用SQL提取需要的信息
- 数据可视化 - 用ECharts制作精美的图表

右边的图表展示了整个数据管道：

1. 收集原始交易数据
2. 存储到Hadoop并用Hive结构化
3. 用Spark分析
4. 将结果存储到MySQL
5. 用ECharts可视化

在这个过程中，会使用Linux、MySQL、Hadoop、Hive、Sqoop、Eclipse、ECharts、Spark等大数据生态系统的主要工具。
仅完成这一个项目，就可以在简历上堂堂正正地写上实战项目。

Spark는 정말 강력한 도구입니다.

제대로 배워두면 데이터 엔지니어, 데이터 과학자, 머신러닝 엔지니어 등 다양한 진로에서 핵심 역량이 됩니다.

이번 학기 함께 열심히 해봅시다. 여러분의 성장을 지켜보는 것이 저에게도 큰 기쁨입니다.

다음 시간에는 제1장, 빅데이터 기술 개술부터 시작하겠습니다. 미리 관련 자료를 읽어오시면 좋겠습니다.

수고하셨습니다!

Spark是非常强大的工具。

学好了，在数据工程师、数据科学家、机器学习工程师等多种职业发展方向上都会成为核心能力。

这学期让我们一起努力吧。看到大家的成长对我来说也是很大的快乐。

下次课从第1章大数据技术概述开始。希望大家提前阅读相关资料。

辛苦了！