编 程 学 习 笔 记

PySpark

作 者 姓 名： 颜佳

目 录

[1 第一章 1](#_Toc76209995)

[1.1 第一节 1](#_Toc76209996)

[1.1.1 第一段 1](#_Toc76209997)

[1.2 1](#_Toc76209998)

[2 1](#_Toc76209999)

Spark

Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎。Spark是UC Berkeley AMP lab (加州大学伯克利分校的AMP实验室)所开源的类Hadoop MapReduce的通用并行框架

# PySpark

## 简介

PySpark 是 Spark 为 Python 开发者提供的 API。PySpark 支持 Spark 的大部分功能，例如 Spark SQL、DataFrame、Streaming、MLlib 和 Spark Core。

## DataFrame

### 创建

df = spark.createDataFrame([

Row(a=1, b=2., c='string1', d=date(2000, 1, 1), e=datetime(2000, 1, 1, 12, 0)),

Row(a=2, b=3., c='string2', d=date(2000, 2, 1), e=datetime(2000, 1, 2, 12, 0)),

Row(a=4, b=5., c='string3', d=date(2000, 3, 1), e=datetime(2000, 1, 3, 12, 0))

])