# Spark安装与启动

环境：

Ubantu 14.04

Hadoop

Spark 2.4.0

## Hadoop安装

### Java安装

http://www.cnblogs.com/liugh/p/6623530.html

### SSH安装

<http://dblab.xmu.edu.cn/blog/install-hadoop/>

### Hadoop安装

http://dblab.xmu.edu.cn/blog/install-hadoop-cluster/

## Spark安装

http://dblab.xmu.edu.cn/blog/1187-2/

## Spark启动

http://dblab.xmu.edu.cn/blog/1187-2/

## 问题

# 数据导入与预处理

## 数据如何导入到分布式系统

sc.textFile()

读取本地文件：

读取分布式系统中的文件：

## 导入后的数据如何进行操作

### 普通文本数据

### Json数据

### Csv数据

## 数据存储

# 复杂操作与特征提取

## 如何对数据进行一系列复杂的运算

## 能否调用封装好的模块对数据进行一系列复杂处理

# 机器学习模型

## 模型数据格式

## 如何构建模型

## 能否使用非Spark的机器学习库

## 模型训练与测试

## 模型存储与加载

# 结果存储与获取

## 结果存储的形式?分布式数据库？分布式文件系统

## 其他程序如何获取分布式系统中的结果