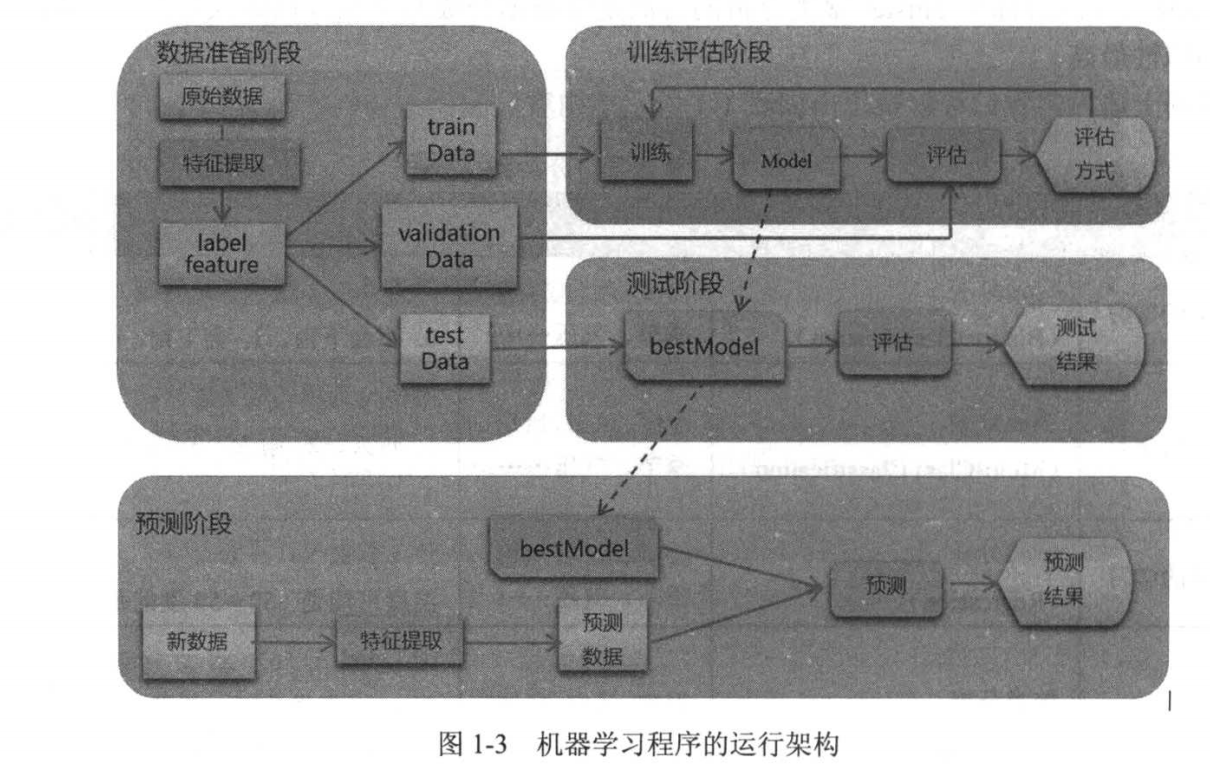
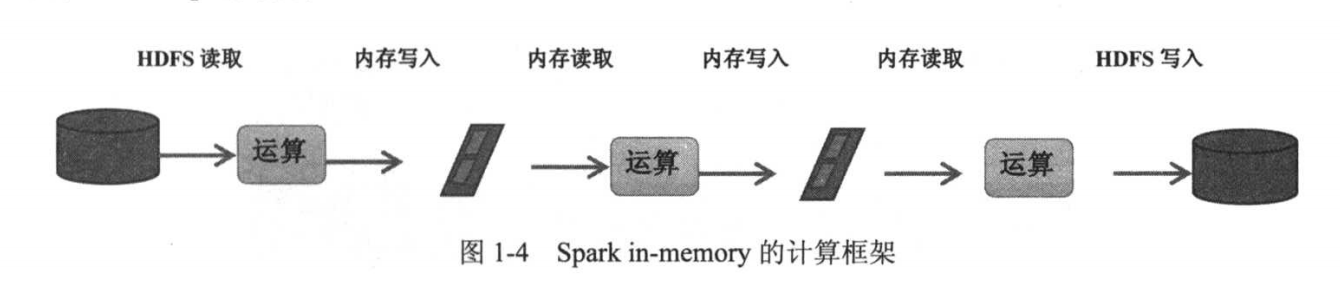
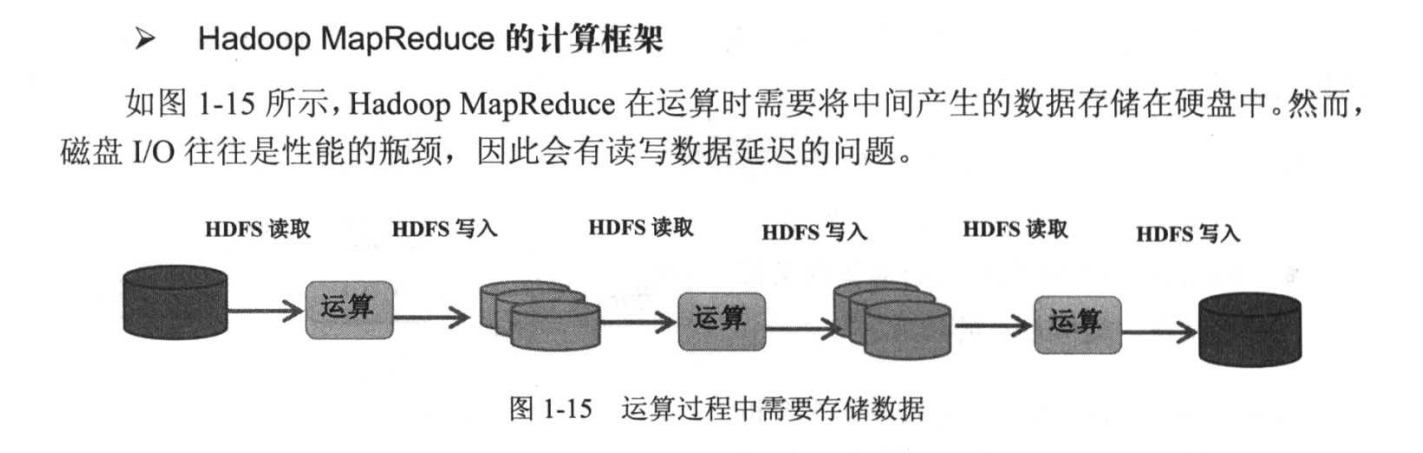
# 第一章 介绍



1.2 Spark的核心是RDD(Resilient Distributed Dataset)弹性分布式数据集



数据处理三种方式：RDD DataFrame Spark SQL 三者区别(是否定义schema-定义字段名和数据类型)



# 第二章 VirtualBox虚拟机安装

下载虚拟机

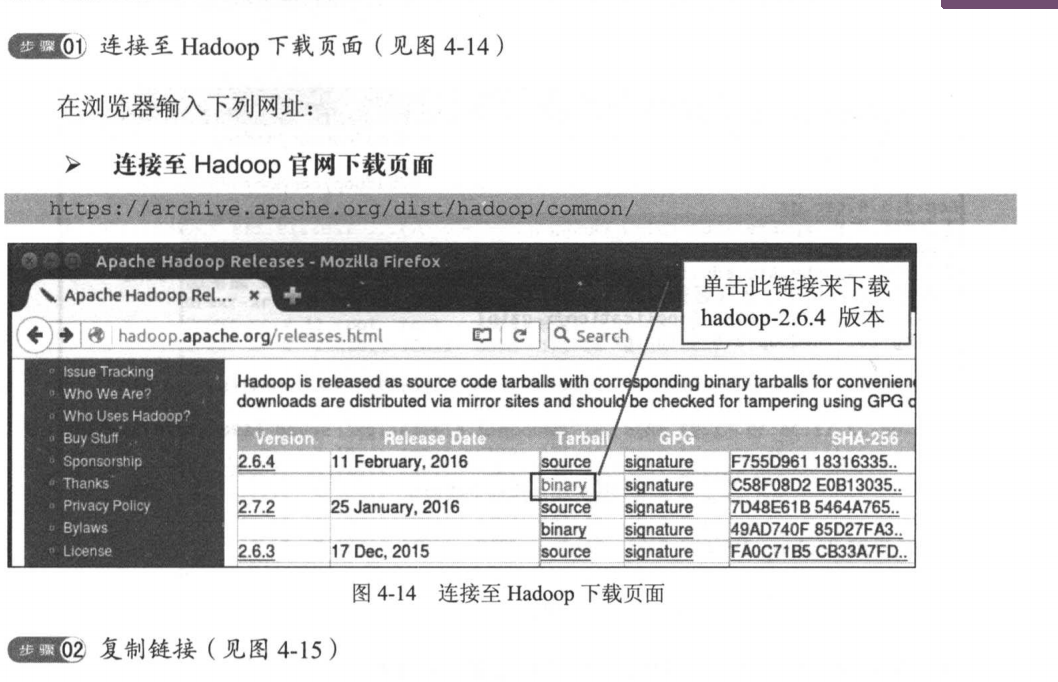
https://www.virtualbox.org/wiki/Downloads

# 第三章 VirtualBox虚拟机安装Ubuntu Linux

下载Ubuntu系统并在虚拟机上面安装

<https://mirrors.tuna.tsinghua.edu.cn/ubuntu-releases/14.04/>

# 第四章 Hadoop Signal Node Cluster 安装



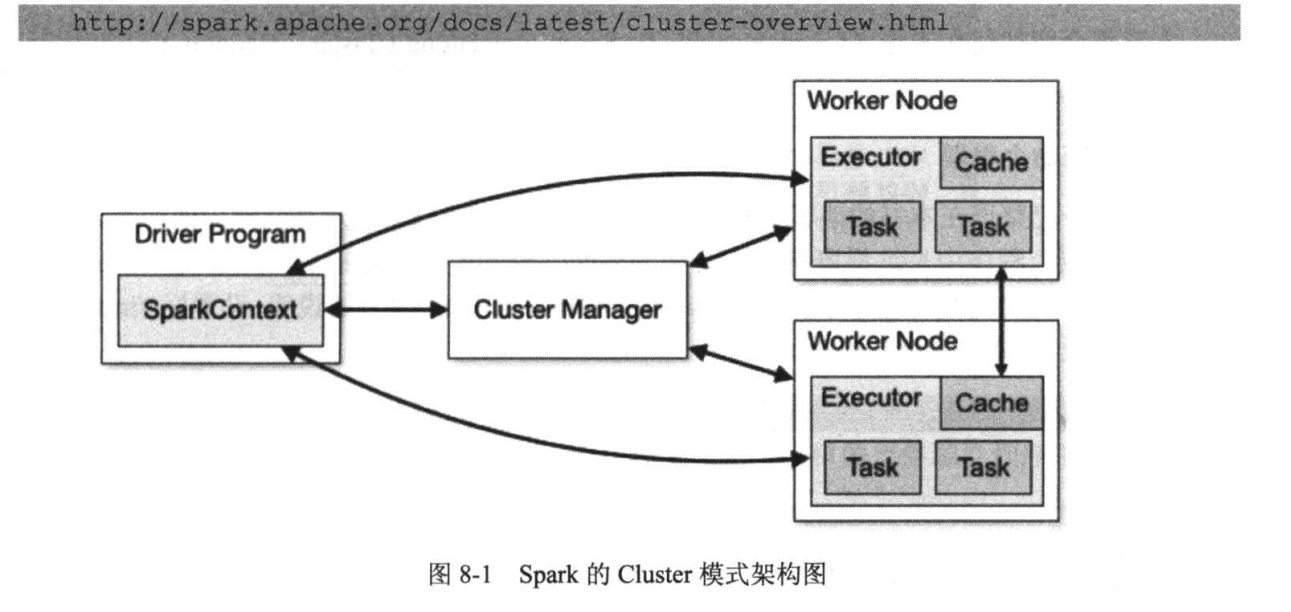
https://archive.apache.org/dist/hadoop/common/

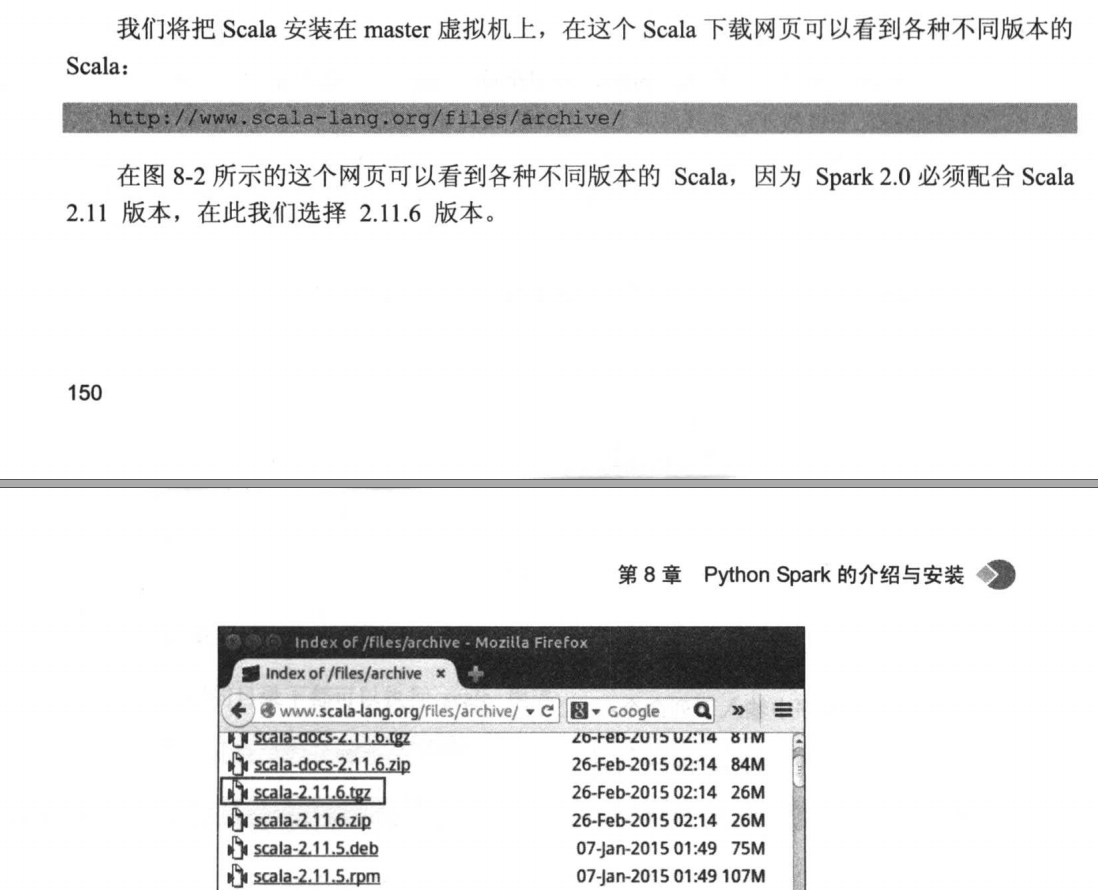
# 第五章 Hadoop Multi Node Cluster 安装

# 第六章 Hadoop HDFS命令

# 第七章 Hadoop MapReduce

# 第八章 Python-Spark安装





# 第九章 Python运行Spark