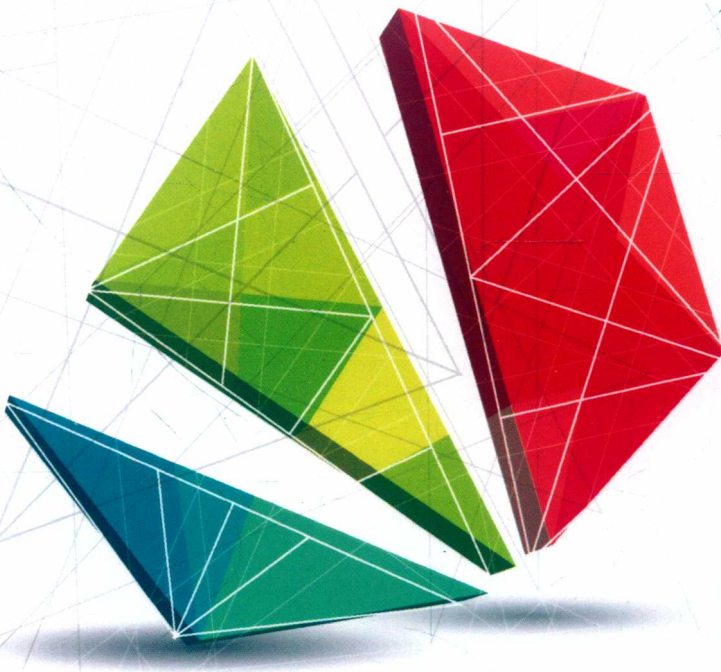
高等学校大数据技术与应用规划教材

Hadoop大数据分析

Hadoop DASHUJU FENXI

高永彬　钱亮宏　方志军　编著



中国铁道出版社有限公司
CHINA RAILWAY PUBLISHING HOUSE CO., LTD.

7

高等学校大数据技术与应用规划教材

Hadoop 大数据分析

高永彬　钱亮宏　方志军　编著

中国铁道出版社有限公司
CHINA RAILWAY PUBLISHING HOUSE CO., LTD.

内 容 简 介

本书从 Hadoop 的原理和使用出发，在重点介绍 Hadoop 生态系统的重要组件 HDFS、MapReduce、YARN、Hive 和 Spark 的同时，注重大数据分析能力的全面提高。

本书共分 13 章，主要内容包括 Hadoop 简介、HDFS 文件系统、YARN 资源管理、MapReduce 计算框架、Hive 简介、Hive 数据定义、Hive 数据操作、Hive 数据查询、Spark 简介、Spark 大数据处理、Spark 机器学习流程、Spark 有监督学习模型和 Spark 无监督学习模型。

本书内容丰富、体系新颖、结构合理、文字精练，适合作为普通高等院校信息类专业 Hadoop 大数据分析课程的教材，也可以作为数据科学行业相关从业人员的自学教材。

图书在版编目（CIP）数据

Hadoop 大数据分析/高永彬，钱亮宏，方志军编著. —北京：中国铁道出版社有限公司，2019.7

高等学校大数据技术与应用规划教材

ISBN 978-7-113-25919-8

Ⅰ.①H… Ⅱ.①高…②钱…③方… Ⅲ.①数据处理软件-高等学校-教材 Ⅳ. ①TP274

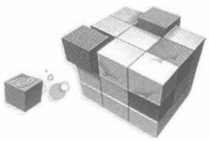
中国版本图书馆 CIP 数据核字（2019）第 119302 号

书　　名：Hadoop 大数据分析
作　　者：高永彬　钱亮宏　方志军

策　　划：曹莉群　　编辑部电话：010-63583215 转 2016
责任编辑：包　宁
封面设计：穆　丽
责任校对：张玉华
责任印制：郭向伟

出版发行：中国铁道出版社有限公司（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.tdpress.com/51eds/
印　　刷：北京柏力行彩印有限公司
版　　次：2019 年 7 月第 1 版　2019 年 7 月第 1 次印刷
开　　本：787 mm×1 092 mm 1/16　插页：2　印张：11.5　字数：287 千
书　　号：ISBN 978-7-113-25919-8
定　　价：38.00 元

前　言



随着信息技术的普及和应用，各行各业产生了大量的数据，人们持续不断地探索处理这些数据的方法，以期最大限度地从中挖掘有用信息。面对如潮水般不断增加的数据，人们不再满足于数据的查询和统计分析，而是期望从数据中提取信息或者知识为决策服务。数据挖掘技术突破数据分析技术的种种局限，结合统计学、数据库、机器学习等技术解决从数据中发现新的信息并辅助决策这一难题，是正在飞速发展的前沿学科。近年来，随着教育部“新工科”建设的不断推进，大数据技术受到广泛关注。数据挖掘作为大数据技术的重要实现手段，能够挖掘数据的关联规则，实现数据的分类、聚类、异常检测和时间序列分析等，解决商务管理、生产控制、市场分析、工程设计和科学探索等各行各业中的数据分析与信息挖掘问题。

Hadoop是一系列分布式存储和计算软件，由Doug Cutting创建，能够支持互联网数据量级别的系统。狭义的Hadoop项目仅包含Hadoop Common、HDFS、YARN和MapReduce 4个组件。广义的Hadoop项目还包含了其他一些衍生性的项目组件，它们或多或少依赖以上4个核心组件，如数据存储依赖于HDFS、作业调度和资源管理依赖YARN，同时它们还解决了一些特定领域的问题。常用的包括Spark、HBase、Hive、Sqoop、Oozie、Impala、Hue、Pig等。

截至2019年1月，共有283所高校获批“数据科学与大数据技术”专业，其中985及211高校占比13%。目前国内大数据人才缺口更是达到百万级。由于其开源性、易用性和强大的数据分析能力，Hadoop已成为世界范围内应用最广泛的数据科学工具和语言之一。目前，Hadoop大数据分析与挖掘逐渐成为高校信息类专业的必修课，同时，作为面向各专业的通识课也广受欢迎。

本书作为立足于应用型本科数据科学与大数据教学的Hadoop核心课教材，具有如下特色：

（1）内容安排合理且全面，从Hadoop的安装配置、分布式数据处理、分布式数据仓库到分布式机器学习，循序渐进，深入浅出。

（2）难度适中，适用于本科中高年级的核心课教材，仅需掌握Python基本编程和Linux基本操作就可以学习本书，对Java编程及数学和算法知识不作为必要基础。

（3）理论与案例相结合，理论与实践相结合，包含了泰坦尼克号乘客生存分析、航班准点数据处理、鸢尾花数据建模等实践案例。

本书主要内容分为以下3部分：

第1部分：Hadoop核心基础，包括第1～4章。第1章为Hadoop简介，包括Hadoop的相关背景、基本概念、安装、配置和运行等。第2章为HDFS文件系统，包括HDFS架构、文件库和常用操作等。第3章为YARN资源管理，包括YARN架构、调度策略

和常用操作等。第 4 章为 MapReduce 计算框架，包括各 MapReduce 原理、流程、词频统计和数据连接的实现等。

第 2 部分：Hive 数据仓库，包括第 5～8 章。第 5 章为 Hive 简介，包括 Hive 的相关背景、基本概念、安装、配置和运行等。第 6 章为 Hive 数据定义，包括数据库操作、数据表操作、数据格式、外部表和分区表等。第 7 章为数据操作，包括数据导入、数据插入和数据导出等。第 8 章为 Hive 数据查询，包括基本查询、数据聚合和数据连接等。

第 3 部分：Spark 数据分析，包括第 9～13 章。第 9 章为 Spark 简介，包括 Spark 的相关背景、基本概念、安装、配置和运行等。第 10 章为 Spark 大数据处理，包括大数据的选择、聚合、引用、筛选、连接和变形等。第 11 章为 Spark 机器学习流程，包括数据探索、划分、填充、特征选择、建模调优和测试评估等。第 12 章为 Spark 有监督学习模型，包括线性、决策树、随机森林、神经网络和协同过滤等。第 13 章为 Spark 无监督学习模型，包括 k 均值聚类、主成分分析和关联分析模型等。

本书例子中的所有数据都可在 GitHub 上公开下载，地址为 https://github.com/yepdata/hadoop_textbook。

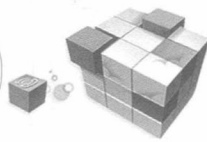
本书由高永彬、钱亮宏和方志军编著。具体分工如下：高永彬编写第 1～4 章；方志军编写第 5～8 章，钱亮宏编写第 9～13 章。全书由范磊和许华根主审。同时感谢戴仁月、严娟和刘敏对本书的贡献。

由于编者水平有限，加之时间仓促，书中难免存在疏漏和不足之处，敬请老师和同学批评指正。

编　者

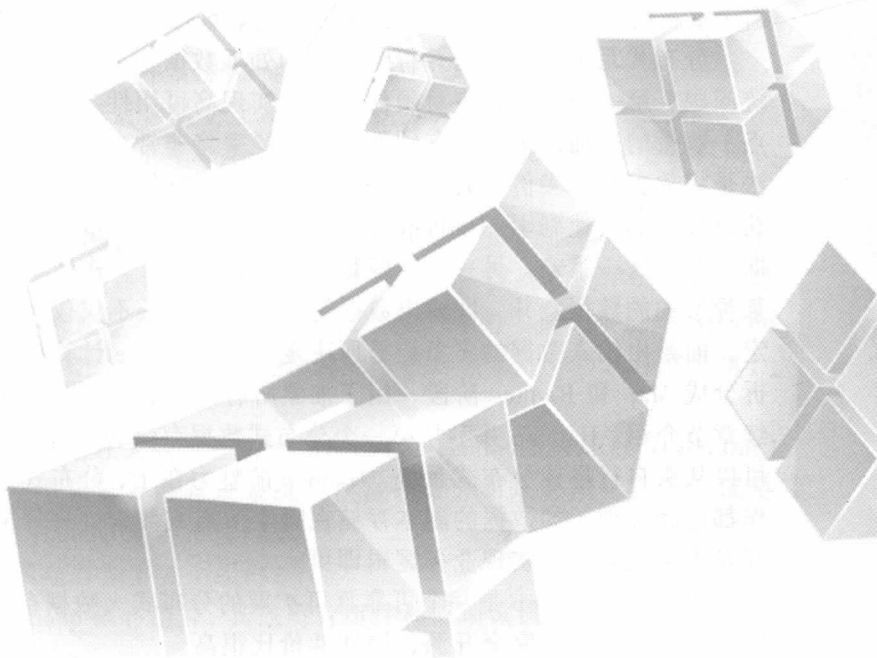
2019 年 5 月

目　录



第1部分

Hadoop 核心基础



第1章 Hadoop简介

1.1 Hadoop产生背景

随着互联网的爆炸式发展，全球的数据量正以指数级增长。据工信部表示，我国数据量正以每年50%的速度增长，预计到2020年中国数据总量将达到8.8 ZB（1 ZB = 1024 EB，1 EB = 1024 PB），而世界数据总量将达到40 ZB。据IBM称，整个人类文明所获得的全部数据中，有90%是过去两年内产生的。随着数据体量的增加，大数据时代悄然来临。

全球范围内数以亿计的人们每天与互联网接触，而这些线上行为也产生了丰富的数据轨迹。互联网公司能够捕捉到每一次与用户的交互，包括线上搜索、网页点击、停留时间和达成交易。数据不仅仅是存储量上的增长，同时还更加丰富和多样化。产生了大量图片、视频、文本、语音和传感器信号等不同于传统的能用行和列表示的结构化数据。

随着数据量的增加，采用单台高性能服务器已无法处理和分析全量数据。即使单台高性能服务器能够满足需要，其硬件成本投入与计算能力并不线性相关，即计算能力提升一倍时硬件成本投入通常需要提升远超一倍。而基于多台商用服务器的分布式集群则能够较好地处理和分析大数据。所谓商用服务器，指的并不是已淘汰的服务器或家用PC，而是相对于定制化大型机而言的企业级通用服务器，硬件成本相比于定制化大型机要低很多。

分布式系统并不只是并行的存储和处理数据这么简单，还需要考虑许多复杂的系统问题。第一个问题是应对硬件故障以实现高可用性。随着服务器（在集群中又称节点）数量的增加，集群中有硬件故障的概率将非常高。为避免数据丢失，最常用的方法是将数据复制多份，放在不同的服务器上，这样即使一台服务器出现故障，仍能够获取数据的其他副本。Hadoop分布式文件系统（Hadoop Distributed File System，HDFS）即采用了这一策略。另一个问题是数据的合并方式。多数数据处理和分析过程都需要某种形式的数据合并，即输出数据的某一条记录并不仅仅由输入数据的某一条记录决定，而是由输入数据的多条记录所决定。MapReduce计算框架将数据处理的每一阶段拆分成Map和Reduce阶段，用于抽象所有的数据处理和分析过程，具体细节将在后续章节介绍。Hadoop并不是第一个分布式数据存储和处理平台，甚至任何开发人员都可以从头自行搭建分布式系统。Hadoop的魅力在于，分布式系统中那些琐碎的后勤工作都已经实现（如数据的副本存储在哪个节点或数据处理任务分配给哪个节点），而开发人员只需要去实现业务逻辑即可。

简言之，Hadoop是一个可靠且可扩展的分布式大数据存储和处理平台，可以在商用服务器上运行，完全开源，因此性价比很高。

1.2 Hadoop 简要历史

Hadoop 由 Doug Cutting 创建，他同时也是文本搜索库 Apache Lucene 项目的创建人。Hadoop 原先是开源网页搜索引擎 Apache Nutch 项目的一部分，而 Nutch 同时又是 Lucene 的一部分。

Nutch 创建于 2002 年，然而该项目的创建人意识到他们原先的系统架构无法扩展到数以亿计的网页。而就在 2003 年 10 月，谷歌发表了一篇描述谷歌分布式文件系统（Google File System，GFS）架构的论文。2004 年，Nutch 的开发人员实现了该论文的思想，并命名为 Nutch 分布式文件系统（Nutch Distributed File System，NDFS）。

2004 年 12 月，谷歌又发表了一篇介绍 MapReduce 计算框架的论文。2005 年，Nutch 的开发人员同样实现了 MapReduce，并将 Nutch 中的所有算法迁移到了 MapReduce 计算框架和 NDFS 文件系统中。

Nutch 中 NDFS 和 MapReduce 的实现完全可以应用于搜索引擎以外的领域。因此，2006 年 2 月，这部分从 Nutch 中独立成了 Lucene 下面的一个子项目，称为 Hadoop。与此同时，Doug Cutting 加入了雅虎，带领一个团队全身心投入，使得 Hadoop 成为能够支持互联网数据量级别的系统。2006 年 4 月，Hadoop 0.1 版本正式发布。

2008 年 1 月，Hadoop 称为 Apache 顶级项目，不再作为 Lucene 的子项目存在，证明了自己的成功。与此同时，许多雅虎以外的公司（如脸书等）也开始使用 Hadoop 并贡献代码。2008 年 4 月，Hadoop 用 910 个节点在 209 s 内排序了 1 TB 数据，创造了世界纪录，2007 年的纪录是 297 s。2009 年 4 月，雅虎宣布 Hadoop 用 62 s 排序了 1 TB 数据。

1.3 Hadoop 生态系统组件

最初的 Hadoop 是由 HDFS 文件系统和 MapReduce 计算框架组成的。MapReduce 本质上是一个批处理系统，并不适合做交互式分析。这意味着用户提交一个查询后，无法在几秒内得到结果。多数在 MapReduce 中运行的查询需要几分钟甚至更长时间，因此比较适合做离线数据处理。

正是由于 MapReduce 应用场景的局限性，在 Hadoop 诞生后的几年时间里，衍生出了许多其他项目组件，这些项目组件都可以理解为 Hadoop 生态系统中的项目，同时 Hadoop 这个词从广义上来说也都包含了这些项目组件。这些项目大多也是 Apache 软件基金会的项目。例如，HBase 就是一个能快速响应查询结果的分布式存储系统，适用于需要快速读/写的数据应用。

Hadoop 生态系统组件不断壮大的过程中，YARN（Yet Another Resource Negotiator）的引入功不可没。它是一个集群资源管理系统，使得任意分布式应用（而不仅仅是 MapReduce 程序）可以在 Hadoop 集群中运行。

狭义的 Hadoop 项目仅包含如下 4 个组件：

- Hadoop Common：Hadoop 核心组件，其他所有组件都依赖该核心组件，为 Hadoop

的其他组件提供了一些常用工具，主要包括系统配置工具、序列化机制和 Hadoop 抽象文件系统等。

- HDFS：Hadoop 分布式文件系统，为 Hadoop 的其他组件提供高吞吐量的数据访问、管理文件的存储位置和副本情况。
- YARN：Hadoop 作业调度和资源管理系统，即用户提交作业的排队方式、分配资源（主要为 CPU 核心和内存）的数量和调度具体执行任务的节点。
- MapReduce：基于 YARN 的并行处理大数据的计算框架，用户只需要实现 Map 和 Reduce 阶段，其他后勤工作由框架完成。

广义的 Hadoop 项目还包含了其他一些衍生性的项目组件，它们或多或少依赖以上 4 个核心组件，如数据存储依赖 HDFS、作业调度和资源管理依赖 YARN，同时它们还解决了一些特定领域的问题。Hadoop 生态系统的项目组件如图 1-1 所示。

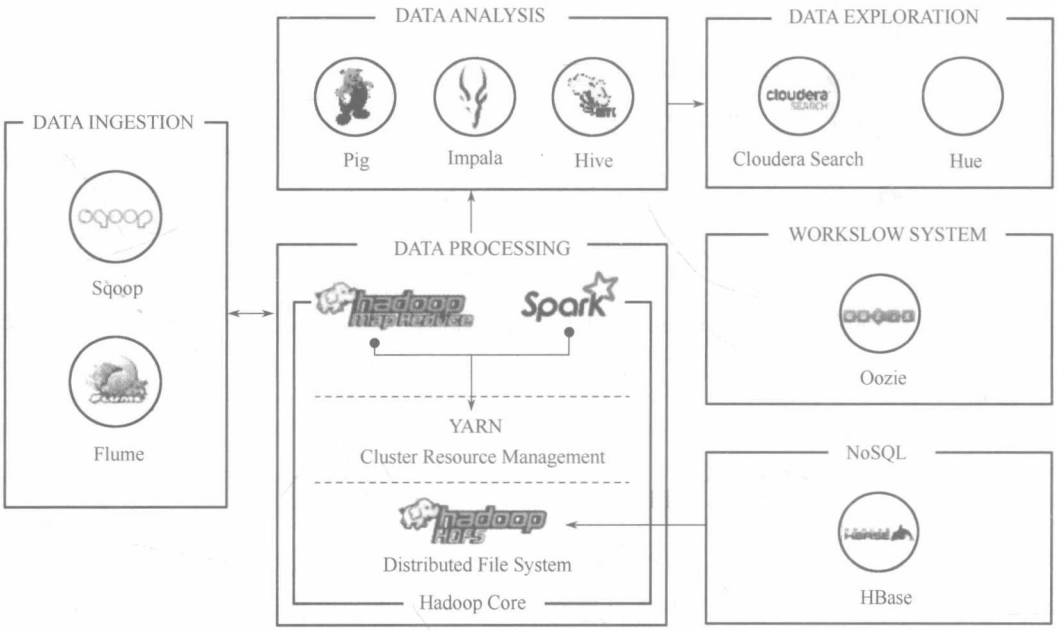


图 1-1　Hadoop 生态系统的项目组件

- Spark：大数据快速处理和分析的通用引擎，编程模型简单，支持多种应用，包括 ETL（extraction, transformation & loading）、机器学习、流处理和图计算等。
- HBase：非 SQL 型数据库，用于存储结构化的大数据表，由于查询不用通过 YARN 做资源调度，因此响应速度很快。
- Hive：SQL 型数据仓库，用于做数据的批处理，吞吐量高但响应速度较慢。
- Sqoop：数据导入/导出工具，用于从传统关系型数据库（如 MySQL 和 Oracle 等）导入/导出数据到 HDFS 或 Hive。
- Oozie：工作流调度系统，用于定义一系列工作流程以及执行路径，并按特定频率执行或由特定事件触发。
- Impala：SQL 型数据仓库，对 Hive 做了较大优化，响应速度上做了较大提升，主要由 Cloudera 公司主导开发。

- Hue：Hadoop 平台的网页接口，使得分析人员能够通过网页界面提交作业并查看结果，提升了用户体验。
- Pig：数据处理工具，用简单的脚本语言做复杂的变换、聚合和分析操作。

1.4 Hadoop 版本和商用支持

自从 2006 年 4 月 Hadoop 0.1 版本正式发布，Hadoop 生态系统中的项目持续不断演进，开源社区一直保持活跃，几次重大的版本升级和增加的特性如下。

- 2011 年 11 月，Hadoop 1.0 版本发布，意味着 Hadoop 已经完全适用于企业生产环境，包括了许多企业安全策略方面的支持。
- 2012 年 5 月，Hadoop 2.0 版本发布，增加了 HDFS 多命名节点的机制，即可以配置多个命名节点，则单个命名节点的故障不会引发整个 HDFS 文件系统无法访问；另外，引入了 YARN 作为作业调度和资源管理系统。
- 2016 年 9 月，Hadoop 3.0 版本发布，支持 2 个以上命名节点，HDFS 支持纠删码。

同时，随着 Hadoop 生态系统的不断丰富，大数据软件市场也充满了机遇。Hadoop 虽然已经大大简化了用户搭建分布式大数据系统的流程，但仍然需要较高的技术能力，尤其是应对突发平台故障的能力。另外，Hadoop 生态系统中的各项目组件存在多种版本，不同版本之间的兼容性和稳定性对于企业用户而言至关重要。因此，一些提供 Hadoop 技术服务的公司应运而生，它们会基于开源 Hadoop 生态系统各项目组件，经过严格的版本兼容性和稳定性测试，打包并发布稳定的发行版，并在此基础上提供管理工具、技术支持和咨询服务。其中，最具有代表性的是 Cloudera 和 Hortonworks。

Cloudera 公司发布的 Hadoop 发行版简称为 CDH（Cloudera Distribution of Hadoop），也是世界范围内应用最广泛的 Hadoop 发行版，本书所有例子都将在 CDH 上运行。截至本书撰稿时，CDH 的最新版本为 6.1，其对应的 Hadoop 生态系统项目组件的版本为：Hadoop 3.0.0、Hive 2.1.1、HBase 2.1.1、Spark 2.4、Sqoop 1.4.7、Oozie 5.0.0，以及其他组件。

类似的，Hortonworks 公司发布的 Hadoop 发行版简称为 HDP（Hortonworks Data Platform）。截至本书撰稿时，HDP 的最新版本为 3.1，其对应的 Hadoop 生态系统项目组件的版本如图 1-2 所示。

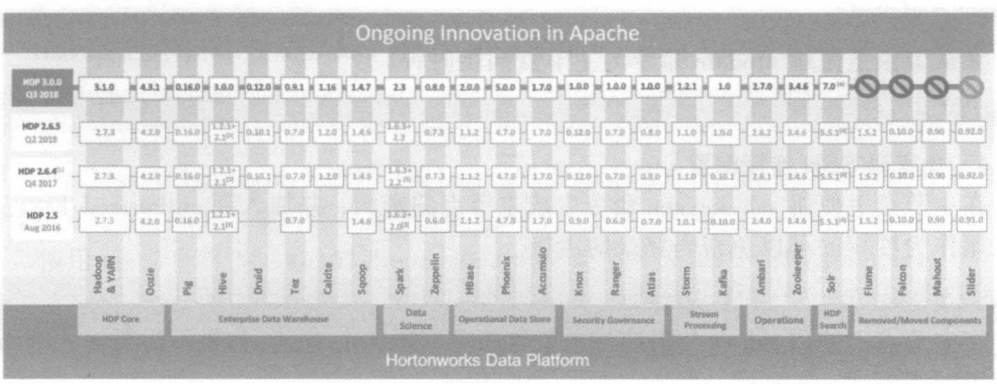


图 1-2 Hadoop 生态系统项目组件的版本

2018 年 10 月，Cloudera 和 Hortonworks 公司正式合并，意味着 Hadoop 的标准将更加统一。

1.5 Hadoop 的基础环境配置

搭建分布式 Hadoop 涉及 3 个节点，其中，主机名为 master 的节点为主节点，运行 HDFS 命名节点、YARN 资源管理等服务，同时也运行从节点的 HDFS 数据节点、YARN 节点管理等服务；主机名为 slave1 和 slave2 的节点为从节点，运行 HDFS 数据节点、YARN 节点管理等服务。

以下步骤中，主机名和 IP 地址以真实环境为准。以下每个步骤都尽可能注明执行位置为主节点、从节点或所有节点。

1. 关闭防火墙（所有节点）

CentOS 的防火墙服务 firewalld 用于限制特定端口的外部访问。Hadoop 包含了多种服务，使用到了不同端口，因此需要将所有节点的防火墙关闭。

```
sudo service firewalld stop
```

2. 主机名映射配置（所有节点）

配置集群中各组成节点的主机名和 IP 地址的映射，打开文件/etc/hosts。

```
vim /etc/hosts
```

在文件末尾添加如下内容，以实际的主机名和 IP 地址为准。

```
192.168.95.128 master192.168.95.129 slave1192.168.95.130 slave2
```

3. 时钟同步（所有节点）

启动时钟同步服务。

使用 systemctl enable 命令配置时钟同步服务在系统启动时自动启动。

```
systemctl enable ntpd
```

使用 systemctl start 命令启动时钟同步服务。

```
systemctl start ntpd
```

在从节点上使用 ntpdate 命令与主节点同步时钟。

```
ntpdate master
```

4. 免密码 SSH 连接配置

在主节点上使用 ssh-keygen 命令生成主节点的密钥，并加入到 authorized_keys 文件末尾作为授权的可以免密码登录的主机。

```
ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsacat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
```

在主节点上使用 ssh 命令测试与主节点的免密码 SSH 连接。如果不需要输入密码，则表示配置成功。

```
ssh master
```

在从节点上使用 scp 命令从主节点复制主节点的密钥到从节点，并加入到文件 authorized_keys 末尾作为授权的可以免密码登录的主机。

```
scp master:~/.ssh/id_dsa.pub ~/master_dsa.pubcat ~/master_ das.pub >> ~/.ssh/authorized_keys
```

在主节点上使用 ssh 命令测试与从节点的免密码 SSH 连接。如果不需要输入密码，则表示配置成功。

```
ssh slave1
ssh slave2
```

1.6 Hadoop 的安装

本节以在 CentOS 7.5 操作系统上搭建分布式 CDH 6.1 为例（该版本于 2018 年 12 月发布），搭建其他版本的过程类似。整个安装过程要求保持互联网连接。

1. 配置软件库（主节点）

CentOS 属于 RHEL 型操作系统，使用 yum 作为包管理工具，依赖于软件库来安装软件。

配置 CDH 6.1 的软件库，创建并打开软件库文件/etc/yum.repos.d/cloudera-cdh6.repo。

```
vim /etc/yum.repos.d/cloudera-cdh6.repo
```

使其看起来如下所示。

```
[cloudera-cdh6]
# Packages for Cloudera's Distribution for Hadoop, Version 6, on RedHat or CentOS 7 x86_64
name=Cloudera's Distribution for Hadoop, Version 6
baseurl=https://archive.cloudera.com/cdh6/6.1.0/redhat7/yum/
gpgkey=https://archive.cloudera.com/cdh6/6.1.0/redhat7/
yum/RPM-GPG-KEY-cloudera gpgcheck = 0
```

使用 rpm 命令导入软件库 GPG 密钥。

```
sudo rpm --import https://archive.cloudera.com/cdh6/6.1.0/redhat7/ yum/RPM-GPG-KEY-cloudera
```

2. 安装 Java 环境（主节点）

Hadoop 的运行和开发需要 Java 环境，即 Java Development Kit（JDK）。从 Java 官方网站（http://www.oracle.com/technetwork/java/javase/downloads/java-archive-javase8- 2177648.html）下载最新的 Java 版本，本节以 Java SE Development Kit 8u192 为例（见图 1-3）。选择接受授权条款，即 Accept License Agreement，并选择下载 Linux x64 的 tar.gz 版本，即 jdk-8u192-linux-x64.tar.gz。需要注意的是，Hadoop 暂时还不支持 JDK 9。

使用 tar 命令将 jdk-8u162-linux-x64.tar.gz 文件解压缩到文件夹/usr/java。

```
sudo tar zxf jdk-8u192-linux-x64.tar.gz -C /usr/java/
```

在主节点上使用 scp 命令从主节点复制 Java 路径到从节点的相同路径。

```
scp -r /usr/java/jdk-8u192-linux-x64 slave1:/usr/java/s
cp -r /usr/java/jdk-8u192-linux-x64 slave2:/usr/java/
```

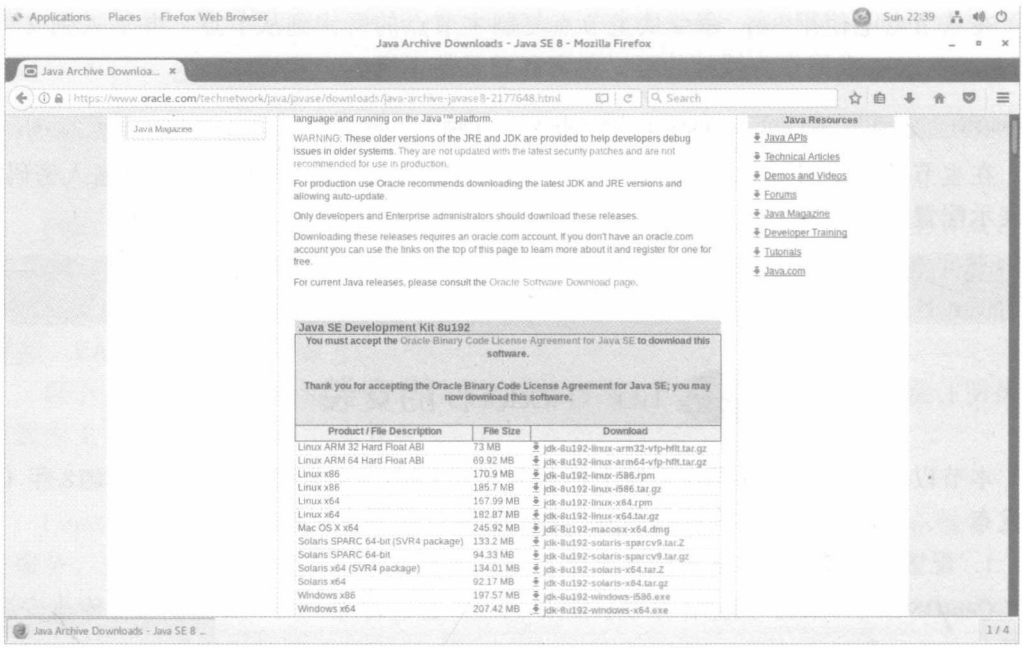


图 1-3　Java 官方网站

3. 安装关系型数据库 MySQL（主节点）

Hadoop 中的一些组件（如 Hive、Sqoop、Hue 或 Oozie）需要关系型数据库存储一些辅助信息才能正常运行，如 Hive 需要将元数据（metadata）存储在关系型数据库中。本节以 MySQL 为例。

使用 wget 命令下载 MySQL 的软件库。

```
sudo wget http://repo.mysql.com/mysql-community-release-el7-5.noarch.rpm
```

使用 rpm 命令导入 MySQL 的软件库。

```
sudo rpm -ivh mysql-community-release-el7-5.noarch.rpm
```

使用 yum 命令安装 MySQL 软件包。

```
sudo yum install mysql-server
```

打开 MySQL 配置文件。

```
vim /etc/my.cnf
```

修改文件，使其看起来如下所示。

```
[mysqld]
datadir=/var/lib/mysql
socket=/var/lib/mysql/mysql.sock
transaction-isolation = READ-COMMITTED
# Disabling symbolic-links is recommended to prevent assorted security risks;
# to do so, uncomment this line:
symbolic-links = 0

key_buffer_size = 32M
```

```
max_allowed_packet = 32M
thread_stack = 256K
thread_cache_size = 64
query_cache_limit = 8M
query_cache_size = 64M
query_cache_type = 1

max_connections = 550
#expire_logs_days = 10
#max_binlog_size = 100M

#log_bin should be on a disk with enough free space.
#Replace '/var/lib/mysql/mysql_binary_log' with an appropriate path for your
#system and chown the specified folder to the mysql user.
log_bin=/var/lib/mysql/mysql_binary_log

#In later versions of MySQL, if you enable the binary log and do not set
#a server_id, MySQL will not start. The server_id must be unique within
#the replicating group.
server_id=1

binlog_format = mixed

read_buffer_size = 2M
read_rnd_buffer_size = 16M
sort_buffer_size = 8M
join_buffer_size = 8M

# InnoDB settings
innodb_file_per_table = 1
innodb_flush_log_at_trx_commit = 2
innodb_log_buffer_size = 64M
innodb_buffer_pool_size = 1G
innodb_thread_concurrency = 8
innodb_flush_method = O_DIRECT
innodb_log_file_size = 512M

[mysqld_safe]
log-error=/var/log/mysqld.log
pid-file=/var/run/mysqld/mysqld.pid

sql_mode=STRICT_ALL_TABLES
```

使用 systemctl enable 命令配置 MySQL 服务在系统启动时自动启动。

```
sudo systemctl enable mysqld
```

使用 systemctl start 命令启动 MySQL 服务。

```
sudo systemctl start mysqld
```

执行安装脚本/usr/bin/mysql_secure_installation 配置 MySQL 中 root 用户的密码以及其他安全选项，这里将 root 用户密码设置为 123456。其他提示则按照如下示例输入。

```
sudo /usr/bin/mysql_secure_installation
[...]
```

```
Enter current password for root (enter for none):
OK, successfully used password, movimng on...
[...]
Set root password? [Y/n] y
New password: 123456
Re-enter new password: 123456
Password updated successfully!
Reloading privimlege tables..
 ... Success!
[...]
Remove anonymous users? [Y/n] y
 ... Success!
[...]
Disallow root login remotely? [Y/n] n
 ... Skipping.
[...]
Remove test database and access to it? [Y/n] y
 - Dropping test database...
[...]
Reload privimlege tables now? [Y/n] y
 ... Success!
[...]
```

4. 安装关系型数据库 MySQL JDBC 驱动器（主节点）

使用 rpm 命令安装 MySQL JDBC 驱动器。

```
sudo yum install mysql-connector-java
```

5. 安装 Hadoop 核心组件（所有节点）

使用 yum install 命令在线安装 Hadoop 核心组件，包括：

- HDFS 命令节点服务：hadoop-hdfs-namenode.x86_64；
- HDFS 数据节点服务：hadoop-hdfs-datanode.x86_64；
- YARN 资源管理器服务：hadoop-yarn-resourcemanager.x86_64；
- YARN 节点管理器服务：hadoop-yarn-nodemanager.x86_64。

在主节点上安装所有 4 个服务。

```
sudo yum install hadoop-hdfs-namenode.x86_64
sudo yum install hadoop-hdfs-datanode.x86_64
sudo yum install hadoop-yarn-resourcemanager.x86_64
sudo yum install hadoop-yarn-nodemanager.x86_64
```

在从节点上仅安装 HDFS 数据节点和 YARN 节点管理器服务。

```
sudo yum install hadoop-hdfs-datanode.x86_64
sudo yum install hadoop-yarn-nodemanager.x86_64
```