高等院校信息类新专业规划教材
大数据和人工智能技术丛书

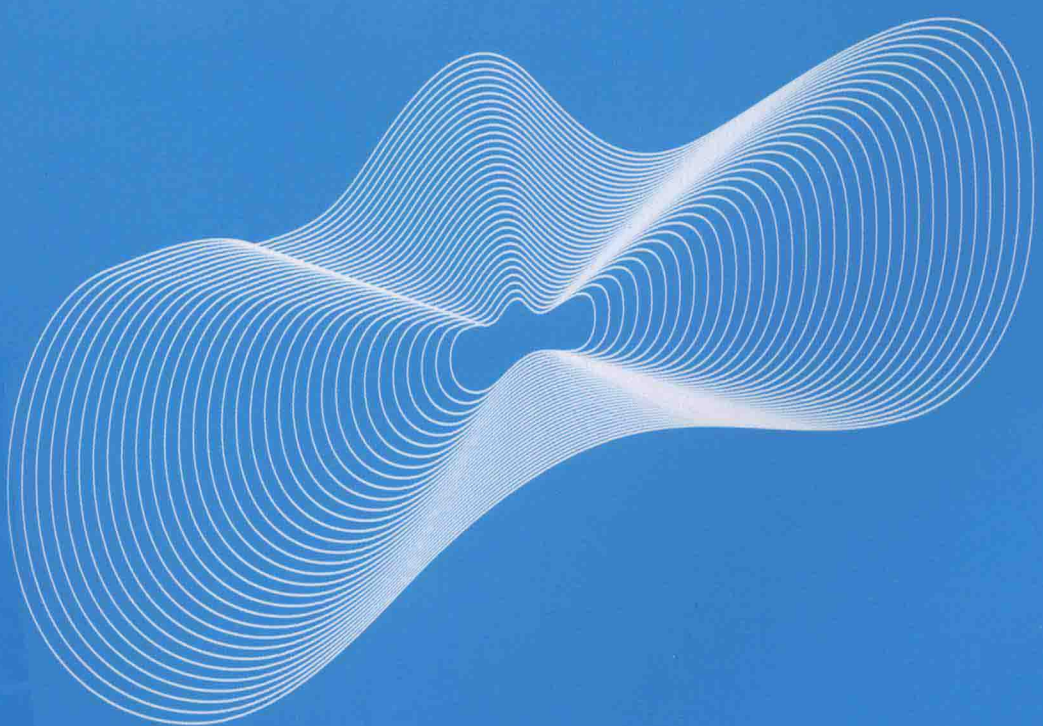
培训推荐教材

华信乾坤

Hadoop DASHUJU PINGTAI
KAIFA YUNWEI SHIXUN

Hadoop大数据平台开发运维实训

主　编◎余　挺
副主编◎张　浩　李　超



北京邮电大学出版社
www.buptpress.com

高等院校信息类新专业规划教材
大数据和人工智能技术丛书

Hadoop 大数据平台开发运维实训

主　编　余　挺
副主编　张　浩　李　超

北京邮电大学出版社
www. buptpress. com

内容简介

大数据时代的到来，迫切需要高校及时建立大数据技术课程体系，为社会培养和输送一大批具备大数据专业素养的高级人才，满足社会对大数据人才日益旺盛的需求。本书定位为大数据技术入门教材，旨在为读者搭建起通向“大数据知识空间”的桥梁。本书将系统地梳理总结 Apache Hadoop 大数据相关技术，介绍大数据存储、并行计算、数据处理等内容，帮助读者形成对大数据知识体系及其应用领域的轮廓性认识，为读者在大数据领域进行更深入的学习和研究奠定基础、指明方向。在本书的基础上，感兴趣的读者可以通过其他诸如《大数据技术原理及应用》《Hadoop 权威指南》等工具书，深入学习和实践大数据相关技术。

本书可作为高等院校计算机、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

图书在版编目(CIP)数据

Hadoop 大数据平台开发运维实训 / 余挺主编. -- 北京：北京邮电大学出版社，2022.1
ISBN 978-7-5635-6584-9

Ⅰ. ①H… Ⅱ. ①余… Ⅲ. ①数据处理软件—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2021)第 261624 号

策划编辑：刘纳新　姚　顺　　**责任编辑**：徐振华　米文秋　　**封面设计**：七星博纳

出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号
邮政编码：100876
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：保定市中画美凯印刷有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：13.5
字　　数：351 千字
版　　次：2022 年 1 月第 1 版
印　　次：2022 年 1 月第 1 次印刷

ISBN 978-7-5635-6584-9　　**定　价**：38.00 元

前　言

进入 2012 年后，大数据（Big Data）一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。

随着云时代的来临，大数据吸引了越来越多的关注。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能并没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。大数据通常用来形容一个公司创造的大量非结构化和半结构化数据，这些数据在下载到关系数据库时用于分析会花费过多的时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据分析需要像 MapReduce 一样的计算框架来向数十、数百甚至数千台计算机分配工作。

本书主要向读者介绍一种大规模数据处理的开源框架——Hadoop 生态系统。在深入探讨 Hadoop 的技术细节和应用之前，有必要花时间来了解 Hadoop 及其取得巨大成功的历史背景。Hadoop 并不是凭空想象出来的，它的出现源于人们创建和使用的数据量的爆炸性增长。在此背景下，不仅庞大的跨国公司面临着海量数据处理的困难，小型创业公司同样如此。与此同时，一些变革改变了软件和系统的部署方式，除了传统的基础设施，人们开始使用甚至偏好于分布式资源处理框架。

本书揭开了 Apache Hadoop 的神秘面纱，着重讲解了如何应用 Hadoop 和相关技术搭建工作系统并完成任务。本书共分为 9 章：第 1 章讲解 Hadoop 的生态系统，以及在行业中的应用场景；第 2 章讲解 Hadoop 分布式文件系统，包括 NameNode 和 DataNode 节点、机架感知策略、HDFS Shell 命令等；第 3 章讲解 MapReduce 并行计算框架，让读者了解 MapReduce的工作原理；第 4 章讲解 HBase 分布式数据库，讲述了 HBase 如何实现数据存储、HBase 的节点类型、HBase API 开发；第 5 章讲解 Hive 数据仓库，介绍了 Hive 的架构、HQL 语法结构、Hive 数据查询案例；第 6 章讲解 Kafka 消息系统，介绍了 Kafka 消息系统的工作原理、Kafka 消息细节处理等；第 7 章讲解 Flume 日志处理系统，介绍了 Flume

的日志处理技术、Flume 如何进行流计算技术处理；第 8 章讲解 ZooKeeper 分布式协调系统，帮助读者理解如何实现 Hadoop 组件之间的协调控制；第 9 章讲解 Sqoop 数据迁移工具，涵盖了有效使用 Sqoop 处理实际场景中的数据迁移工作。

通过阅读本书，读者将迅速掌握编程概念，打下坚实的基础，并养成良好的习惯。此后，读者就可以开始了解其他大数据平台技术，如 Spark 内存计算框架、Flink 流批一体化处理平台，并能够更轻松地掌握大数据技术。

目　录

第1章 Hadoop大数据平台概述

从大数据自身的技术体系来说，大数据所有的技术都紧紧围绕数据价值化来展开，企业对大数据的利用当前也逐渐从传统的数据采集和分析向数据生产转变，相信在工业互联网时代这一趋势会越发明显。

对于企业来说，借助于大数据来降低运营成本是一个重要的诉求，而通过大数据技术来降低运营成本的出发点非常多，不同行业企业要结合自身的实际情况来进行方案规划。当前很多企业利用大数据来构建自己的价值化考核体系，这是降耗提效的好方式。

大数据时代，数据的应用已经渗透到各行各业，但是传统的数据挖掘和分析已经不能满足行业发展的需求，大数据技术为企业业务分析和行业发展带来了新的思维角度，将会充分激发数据对社会发展的影响和推动。如何有效利用大数据平台？接下来我们就一起来了解Apache Hadoop大数据生态系统。

1.1 Hadoop大数据平台起源

Hadoop是Hadoop项目创建者Doug Cutting儿子的一只玩具的名字。他的儿子一直称呼一只黄色的大象玩具为Hadoop，这刚好满足Cutting的命名需求——简短、容易拼写和发音、毫无意义、不会在别处被使用，于是Hadoop就诞生了。Hadoop的发行版本有很多，有华为发行版、星环发行版、Intel发行版、Cloudera发行版(CDH)、MapR版本以及HortonWorks版本等。所有发行版本都是基于Apache Hadoop衍生出来的，产生这些版本的原因可归结为Apache Hadoop的开源协议：任何人都可以对其进行修改，并作为开源或商业产品发布和销售。

1.1.1 Hadoop发展历程

1. Hadoop大数据平台的起源

① Hadoop最早起源于Nutch项目，Nutch的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，其遇到了严重的可扩展性问题——如何解决数十亿网页的存储和索引问题。

② 从2003年开始，Google陆续发表的3篇论文为该问题提供了可行的解决方案。

- 分布式文件系统(DFS)：可用于处理海量网页的存储问题。

- 分布式计算框架 MapReduce：可用于处理海量网页的索引计算问题。
- BigTable 分布式数据库：OLTP(联机事务处理，On-Line Transaction Processing)用于执行增、删、改操作，OLAP(联机分析处理，On-Line Analysis Processing)用于执行查询操作。

③ Nutch 的开发人员完成了相应的开源实现 HDFS 和 MapReduce，并将其从 Nutch 中剥离出来，成为独立项目 Hadoop。直到 2008 年 1 月，Hadoop 成为 Apache 顶级项目，迎来了快速发展期。

2. Hadoop 官网

我们可以通过 Hadoop 官网 http://hadoop.apache.org/来学习 Hadoop 的核心技术。Hadoop 大数据平台的处理主要就是存储和计算，我们安装 Hadoop 集群，目的是实现两个核心功能：一个操作系统 YARN 和一个分布式文件系统 HDFS，其实 MapReduce 就是运行在 YARN 之上的应用。

1.1.2 Hadoop 核心组件

Hadoop 是 Apache 旗下的一套开源软件平台，Hadoop 主要提供的功能是：利用服务器集群，根据用户自定义的逻辑对海量数据进行分布式处理。

1. Hadoop 的概念

① 狭义上：属于 Apache 基金会的一个顶级项目 Apache Hadoop。

② 广义上：以 Hadoop 为核心的整个大数据处理体系，包括计算和存储能力。

2. Hadoop 的核心组件

① Hadoop Common：支持其他 Hadoop 模块的常用工具。

② Hadoop 分布式文件系统(HDFS)：一种分布式文件系统，可提供对应用程序数据的高吞吐量访问。

③ Hadoop YARN：作业调度和集群资源管理的框架。

④ Hadoop MapReduce：一种用于并行处理大型数据集的基于 YARN 的系统。

3. Apache 的其他 Hadoop 相关项目

① Ambari：一种用于供应、管理和监控 Apache Hadoop 集群的基于 Web 的工具，其中包括对 HDFS、Hadoop MapReduce、Hive、HCatalog、HBase、ZooKeeper、Oozie、Pig 和 Sqoop 的支持。Ambari 还提供了一个用于查看集群运行状况的仪表板，如数据热图和可以直观地查看 MapReduce、Pig 和 Hive 应用程序的功能，以及以用户友好的方式诊断其性能特征的功能。

② Avro：数据序列化系统。

③ Cassandra：无单点故障的可扩展多主数据库。

④ Chukwa：管理大型分布式系统的数据收集系统。

⑤ HBase：可扩展的分布式数据库，支持大型表格的结构化数据存储。

⑥ Hive：提供数据汇总和即席查询的数据仓库基础架构。

⑦ Mahout：可扩展的机器学习和数据挖掘库。

⑧ Pig：用于并行计算的高级数据流语言和执行框架。

⑨ Spark：用于 Hadoop 数据的快速和通用计算引擎。Spark 提供了一个简单而富有表现力的编程模型，它支持广泛的应用程序，包括数据抽取、转换、加载(ETL)，机器学习，流计算处

理和图计算。

⑩ Tez:一种基于 Hadoop YARN 的通用数据流编程框架,它提供了一个强大且灵活的引擎,可执行任意有向无环图(DAG)任务来处理批处理和交互式用例的数据。Hadoop、Pig 和 Hadoop 生态系统中的其他框架以及其他商业软件(如 ETL 工具)正在采用 Tez 来替代 Hadoop MapReduce 作为底层执行引擎。

⑪ ZooKeeper:分布式应用程序的高性能协调服务。

1.1.3 Hadoop 与云计算的关系

云计算是分布式计算、并行计算、网格计算、多核计算、网络存储、虚拟化、负载均衡等传统计算机技术和互联网技术融合发展的产物,其借助于基础设施即服务(IaaS)、平台即服务(PaaS)、软件即服务(SaaS)等业务模式,把强大的计算能力提供给终端用户。现阶段云计算的两大底层支撑技术为"虚拟化"和"大数据技术",而 Hadoop 则是云计算的 PaaS 层的解决方案之一,并不等同于 PaaS,更不等同于云计算本身。

大数据与云计算密不可分。大数据必然无法用单台计算机进行处理,必须采用分布式计算架构。大数据的特色在于对海量数据的挖掘,但它必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。它们之间的关系可以这样来理解:云计算技术就是一个容器,大数据正是存放在这个容器中的水,大数据是要依靠云计算技术来进行存储和计算的。大数据发展具有如下趋势。

趋势一:数据的资源化。

资源化是指大数据成为企业和社会关注的重要战略资源,并已成为大家争相抢夺的新焦点。因而,企业必须要提前制订大数据营销战略计划,抢占市场先机。

趋势二:与云计算的深度结合。

大数据离不开云计算技术,云计算为大数据提供了弹性可拓展的基础设备,是产生大数据的平台之一。自 2013 年起,大数据技术已开始和云计算技术紧密结合,未来两者的关系将更为密切。除此之外,物联网、移动互联网等新兴计算形态也将一齐助力大数据革命,让大数据营销发挥出更大的影响力。

趋势三:科学理论的突破。

随着大数据的快速发展,就像计算机和互联网一样,大数据很有可能是新一轮的技术革命。随之兴起的数据挖掘、机器学习和人工智能等相关技术,可能会改变数据世界里的很多算法和基础理论,实现科学技术上的突破。

1.2 Hadoop 集群搭建和简单应用

1.2.1 集群服务器规划

1. 节点规划

本教材中,读者可以使用 4 台 CentOS 6.7 虚拟机进行集群搭建。为了方便呈现每台主机

的功能，主机角色和 IP 地址设置等参考表 1-1。

表 1-1　节点规划

服务器	IP	用户	HDFS	YARN
hadoop1	192.168.123.102	hadoop	NameNode,DataNode	NodeManager
hadoop2	192.168.123.103	hadoop	DataNode	NodeManager
hadoop3	192.168.123.104	hadoop	DataNode,SecondaryNameNode	NodeManager
hadoop4	192.168.123.105	hadoop	DataNode	ResourceManager,NodeManager

2. Hadoop 集群的部署模式

Hadoop 的运行模式分为 3 种：本地运行模式、伪分布运行模式、集群运行模式。

(1) 独立模式(即本地运行模式)

无须运行任何守护进程，所有程序都在单个 Java 虚拟机(JVM)上执行。由于在本机模式下测试和调试 MapReduce 程序较为方便，因此这种模式适合用在开发阶段。独立模式无须配置任何文件。

(2) 伪分布运行模式

如果 Hadoop 对应的 Java 进程都运行在一个物理机器上，则称为伪分布运行模式。以 Windows 为例，在其他系统下，需要修改路径。

(3) 集群运行模式

集群中每一个节点都可以独立运行 Hadoop 的相关进程，防止单点故障，适合应用于生产环境。

1.2.2　Hadoop 软件安装

1. 实验环境规划

- 规划安装用户：hadoop 用户。
- 规划安装目录：/home/hadoop/apps。
- 规划数据目录：/home/hadoop/data。

注意：apps 和 data 两个目录需要自己单独创建。

2. 上传安装软件，并实现软件解压缩

使用 hadoop 用户，尽量不使用 CentOS 操作系统 root 用户登录。

```
[hadoop@hadoop1 apps]$ ls
hadoop-2.7.5-centos-6.7.tar.gz
[hadoop@hadoop1 apps]$ tar -zxvf hadoop-2.7.5-centos-6.7.tar.gz
```

3. 修改配置文件

配置文件目录：/home/hadoop/apps/hadoop-2.7.5/etc/hadoop。

(1) hadoop-env.sh 配置文件

```
[hadoop@hadoop1 hadoop]$ vi hadoop-env.sh
```

Hadoop 使用 Java 开发环境，需要修改操作系统 JAVA_HOME 环境变量：

```
export JAVA_HOME = /usr/local/jdk1.8.0_73
```

(2) core-site.xml 配置文件

```
[hadoop@hadoop1 hadoop] $ vi core-site.xml
```

- fs.defaultFS：这个属性用来指定 NameNode 的 HDFS 协议的文件系统通信地址，可以指定为一个主机＋端口，也可以指定为一个 NameNode 服务〔这个服务内部可以有多台 NameNode 实现高可用性双机集群(HA)的 NameNode 服务〕。
- hadoop.tmp.dir：Hadoop 集群在工作时存储的一些临时文件的目录。

参考配置如下：

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://hadoop1:9000</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/home/hadoop/data/hadoopdata</value>
        </property>
</configuration>
```

(3) hdfs-site.xml 配置文件

```
[hadoop@hadoop1 hadoop] $ vi hdfs-site.xml
```

- dfs.namenode.name.dir：NameNode 元数据的存放目录，记录了 HDFS 中文件的元数据。
- dfs.datanode.data.dir：DataNode 数据的存放目录，也就是数据块(Block)的存放目录。
- dfs.replication：HDFS 的副本数设置。文件写入时被分割为 Block 后，每个 Block 的冗余副本个数，默认配置是 3。
- dfs.secondary.http.address：SecondaryNameNode 运行节点的信息，可以和 NameNode 不同节点，也可以和 NameNode 同一节点。

参考配置如下：

```
<configuration>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>/home/hadoop/data/hadoopdata/name</value>
                <description>可以配置多个不同目录</description>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/home/hadoop/data/hadoopdata/data</value>
                <description>DataNode 的数据存储目录</description>
        </property>
        <property>
```

```
        <name>dfs.replication</name>
        <value>2</value>
        <description>HDFS 数据块的副本存储个数，默认是 3</description>
    </property>
    <property>
        <name>dfs.secondary.http.address</name>
        <value>hadoop3:50090</value>
        <description>SecondaryNameNode 运行节点的信息，可以和 NameNode 不同节点，也
可以和 NameNode 相同节点</description>
    </property>
</configuration>
```

(4) mapred-site.xml 配置文件

```
[hadoop@hadoop1 hadoop]$ cp mapred-site.xml.template mapred-site.xml
[hadoop@hadoop1 hadoop]$ vi mapred-site.xml
```

MapReduce.framework.name:指定 MapReduce 框架为 YARN 方式，Hadoop 二代 MP 也基于资源管理系统 YARN 来运行。

参考配置如下：

```
<configuration>
    <property>
        <name>MapReduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

(5) yarn-site.xml 配置文件

```
[hadoop@hadoop1 hadoop]$ vi yarn-site.xml
```

- yarn.resourcemanager.hostname:YARN 总管理器的进程间通信(IPC)地址。
- yarn.nodemanager.aux-services:NodeManager 上的附属服务，需配置成 MapReduce_shuffle，才可运行 MapReduce 程序。

参考配置如下：

```
<configuration>
<!-- Site specific YARN configuration properties -->
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>hadoop4</value>
    </property>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>MapReduce_shuffle</value>
        <description>shuffle service</description>
    </property>
</configuration>
```

(6) slaves 配置文件

```
[hadoop@hadoop1 hadoop]$ vi slaves
```

参考配置如下：

```
hadoop1
hadoop2
hadoop3
hadoop4
```

4. 把安装包分发给其他的节点

每台服务器中的 Hadoop 安装包的目录必须一致，安装包的配置信息还必须保持一致：

```
[hadoop@hadoop1 hadoop]$ scp -r ~/apps/hadoop-2.7.5/ hadoop2:~/apps/
[hadoop@hadoop1 hadoop]$ scp -r ~/apps/hadoop-2.7.5/ hadoop3:~/apps/
[hadoop@hadoop1 hadoop]$ scp -r ~/apps/hadoop-2.7.5/ hadoop4:~/apps/
```

5. 配置 Hadoop 环境变量

因为我们使用 hadoop 用户进行安装，所以编辑个人目录下的.bashrc 文件，如下：

```
[hadoop@hadoop1 ~]$ vi .bashrc
export HADOOP_HOME = /home/hadoop/apps/hadoop-2.7.5
export PATH = $PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:
```

使环境变量生效：

```
[hadoop@hadoop1 bin]$ source ~/.bashrc
```

6. 查看 Hadoop 版本

命令及输出内容参考如下：

```
[hadoop@hadoop1 bin]$ hadoop version
hadoop 2.7.5
Subversion Unknown -r Unknown
Compiled by root on 2020-3-2T05:30Z
Compiled with protoc 2.5.0
From source with checksum 9f118f95f47043332d51891e37f736e9
This command was run using /home/hadoop/apps/hadoop-2.7.5/share/hadoop/common/hadoop-common-
2.7.5.jar
```

7. Hadoop 集群初始化

HDFS 初始化只能在主节点上进行，执行过程中会产生大量控制台输出提示，命令及最后输出的内容参考如下：

```
[hadoop@hadoop1 ~]$ hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
20/03/03 11:13:24 INFO namenode.namenode: STARTUP_MSG:
/************************************************************
20/03/03 11:13:26 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/data/
hadoopdata/name/current/fsimage.ckpt_0000000000000000000 of size 323 bytes saved in 0 seconds.
20/03/03 11:13:26 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
```