课工场 kgc.cn

大数据开发工程师系列

北京课工场教育科技有限公司 出品

◉ 视频课程 ◉ 案例素材 ◉ 交流社区 ◉ QQ讨论组

Hadoop & Spark 大数据开发实战

主 编 肖 睿 雷刚跃
副主编 宋丽萍 张 宇
彭 英

中国水利水电出版社
www.waterpub.com.cn

课工场介绍

课工场是专注互联网教育的生态平台，汇聚了中国和北美数百位来自知名互联网企业的行业大咖，向寻求就业和技术提升的人群提供直播、录播、面授等多模式教学场景，并通过遍布全国的线下服务中心提供成熟的学习服务，形成完善的“互联网＋教育”解决方案。同时，课工场也为高校、企业、行业提供教育技术赋能，依托 Transformer 智能教育生态平台，打造智慧校园、企业大学、行业培训的教育场景，提供一站式教育解决方案。

课工场于 2016 年荣膺新浪网“2016 中国影响力科技创新教育机构”，腾讯网“2016 中国影响力教育品牌”，网易“2016 年度最受信赖教育机构”，小米“2016 教育行业突出贡献奖”。

扫一扫关注课工场公众号

关注微信 立送20币

可购买收费课程

课工场APP客户端下载

产品/设计/开发/运维/运营

随时随地随心学

大数据开发工程师系列

Hadoop & Spark 大数据开发实战

主　编　肖　睿　雷刚跃
副主编　宋丽萍　张　宇　彭　英

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，最终形成变革之力。本书围绕 Hadoop 和 Spark 这两个主流技术进行讲解，主要内容包括 Hadoop 环境配置、分布式文件系统 HDFS、分布式计算框架 MapReduce、资源调度框架 YARN 与 Hadoop 新特性、大数据数据仓库 Hive、离线处理辅助系统、Spark Core、Spark SQL、Spark Streaming 等知识。

为保证最优学习效果，本书紧密结合实际应用，利用大量案例说明和实践，提炼含金量十足的开发经验。本书使用 Hadoop 和 Spark 进行大数据开发，并配以完善的学习资源和支持服务，包括视频教程、案例素材下载、学习交流社区、讨论组等终身学习内容，为开发者带来全方位的学习体验，更多技术支持请访问课工场官网：www.kgc.cn。

图书在版编目（CIP）数据

Hadoop & Spark大数据开发实战 / 肖睿，雷刚跃主编. -- 北京 : 中国水利水电出版社，2017.7
（大数据开发工程师系列）
ISBN 978-7-5170-5643-0

Ⅰ. ①H… Ⅱ. ①肖… ②雷… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第164300号

策划编辑：祝智敏　责任编辑：李　炎　加工编辑：祝智敏　封面设计：梁　燕

书　名	大数据开发工程师系列 Hadoop & Spark大数据开发实战 Hadoop & Spark DASHUJU KAIFA SHIZHAN
作　者	主 编 肖 睿 雷刚跃 副主编 宋丽萍 张 宇 彭 英
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座 100038） 网 址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电 话：（010）68367658（营销中心）、82562819（万水）
经　售	全国各地新华书店和相关出版物销售网点
排　版	北京万水电子信息有限公司
印　刷	三河市铭浩彩色印装有限公司
规　格	184mm × 260mm　16 开本　19.25 印张　416 千字
版　次	2017 年 7 月第 1 版　2017 年 7 月第 1 次印刷
印　数	0001—3000 册
定　价	58.00 元

丛书编委会

前　言

丛书设计：

准备好了吗？进入大数据时代！大数据已经并将继续影响人类的方方面面。2015年8月31日，经李克强总理批准，国务院正式下发《关于印发促进大数据发展行动纲要的通知》，这是从国家层面正式宣告大数据时代的到来！企业资本则以BAT互联网公司为首，不断进行大数据创新，从而实现大数据的商业价值。本丛书根据企业人才实际需求，参考历史学习难度曲线，选取“Java + 大数据”技术集作为学习路径，旨在为读者提供一站式实战型大数据开发学习指导，帮助读者踏上由开发入门到大数据实战的互联网 + 大数据开发之旅！

丛书特点：

1．以企业需求为设计导向

满足企业对人才的技能需求是本丛书的核心设计原则，为此课工场大数据开发教研团队，通过对数百位BAT一线技术专家进行访谈、对上千家企业人力资源情况进行调研、对上万个企业招聘岗位进行需求分析，从而实现技术的准确定位，达到课程与企业需求的高契合度。

2．以任务驱动为讲解方式

丛书中的技能点和知识点都由任务驱动，读者在学习知识时不仅可以知其然，而且可以知其所以然，帮助读者融会贯通、举一反三。

3．以实战项目来提升技术

本丛书均设置项目实战环节，该环节综合运用书中的知识点，帮助读者提升项目开发能力。每个实战项目都设有相应的项目思路指导、重难点讲解、实现步骤总结和知识点梳理。

4．以互联网 + 实现终身学习

本丛书可通过使用课工场APP进行二维码扫描来观看配套视频的理论讲解和案例操作，同时课工场（www.kgc.cn）开辟教材配套版块，提供案例代码及案例素材下载。此外，课工场还为读者提供了体系化的学习路径、丰富的在线学习资源和活跃的学习社区，方便读者随时学习。

读者对象：

1．大中专院校的老师和学生

2．编程爱好者

3．初中级程序开发人员

4．相关培训机构的老师和学员

读者服务：

为解决本丛书中存在的疑难问题，读者可以访问课工场官方网站（www.kgc.cn），也可以发送邮件到 ke@kgc.cn，我们的客服专员将竭诚为您服务。

致谢：

本丛书是由课工场大数据开发教研团队研发编写的，课工场（kgc.cn）是北京大学旗下专注于互联网人才培养的高端教育品牌。作为国内互联网人才教育生态系统的构建者，课工场依托北京大学优质的教育资源，重构职业教育生态体系，以学员为本、以企业为基，构建教学大咖、技术大咖、行业大咖三咖一体的教学矩阵，为学员提供高端、靠谱、炫酷的学习内容！

感谢您购买本丛书，希望本丛书能成为您大数据开发之旅的好伙伴！

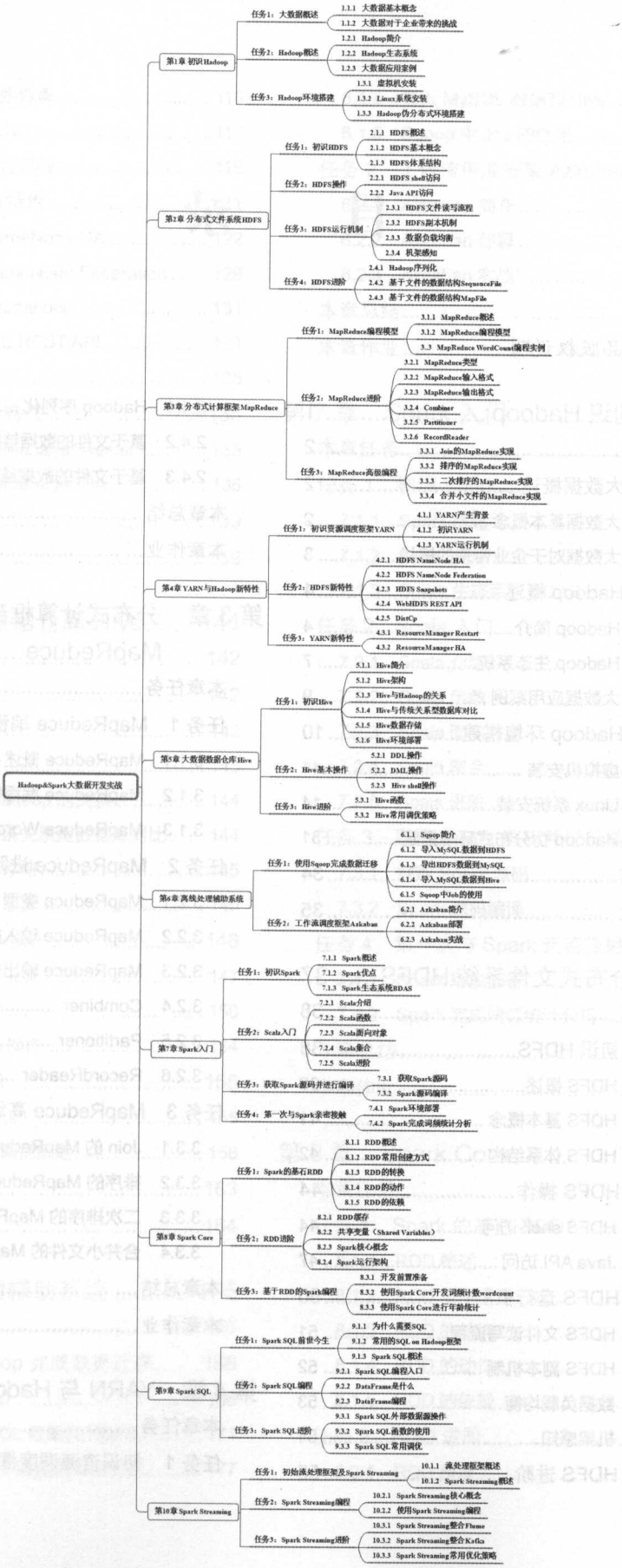
关于引用作品版权说明

为了方便读者学习，促进知识传播，本书选用了一些知名网站的相关内容作为学习案例。为了尊重这些内容所有者的权利，特此声明，凡在书中涉及的版权、著作权、商标权等权益均属于原作品版权人、著作权人、商标权人。

为了维护原作品相关权益人的权益，现对本书选用的主要作品的出处给予说明（排名不分先后）。

序号	选用的网络作品	版权归属
1	MapReduce	hadoop.apache.org
2	YARN	IBM
3	Hive	hive.apache.org
4	Sqoop	sqoop.apache.org
5	Spark	spark.apache.org
6	Spark Streaming	storm.apache.org

由于篇幅有限，以上列表中可能并未全部列出本书所选用的作品。在此，我们衷心感谢所有原作品的相关版权权益人及所属公司对职业教育的大力支持！


Hadoop&Spark大数据开发实战
第1章 初识Hadoop
任务1：大数据概述
1.1.1 大数据基本概念
1.1.2 大数据对于企业带来的挑战
任务2：Hadoop概述
1.2.1 Hadoop简介
1.2.2 Hadoop生态系统
1.2.3 大数据应用案例
任务3：Hadoop环境搭建
1.3.1 虚拟机安装
1.3.2 Linux系统安装
1.3.3 Hadoop伪分布式环境搭建
第2章 分布式文件系统HDFS
任务1：初识HDFS
2.1.1 HDFS概述
2.1.2 HDFS基本概念
2.1.3 HDFS体系结构
任务2：HDFS操作
2.2.1 HDFS shell访问
2.2.2 Java API访问
任务3：HDFS运行机制
2.3.1 HDFS文件读写流程
2.3.2 HDFS副本机制
2.3.3 数据负载均衡
2.3.4 机架感知
任务4：HDFS进阶
2.4.1 Hadoop序列化
2.4.2 基于文件的数据结构SequenceFile
2.4.3 基于文件的数据结构MapFile
第3章 分布式计算框架MapReduce
任务1：MapReduce编程模型
3.1.1 MapReduce概述
3.1.2 MapReduce编程模型
3.3 MapReduce WordCount编程实例
任务2：MapReduce进阶
3.2.1 MapReduce类型
3.2.2 MapReduce输入格式
3.2.3 MapReduce输出格式
3.2.4 Combiner
3.2.5 Partitioner
3.2.6 RecordReader
任务3：MapReduce高级编程
3.3.1 Join的MapReduce实现
3.3.2 排序的MapReduce实现
3.3.3 二次排序的MapReduce实现
3.3.4 合并小文件的MapReduce实现
第4章 YARN与Hadoop新特性
任务1：初识资源调度框架YARN
4.1.1 YARN产生背景
4.1.2 初识YARN
4.1.3 YARN运行机制
任务2：HDFS新特性
4.2.1 HDFS NameNode HA
4.2.2 HDFS NameNode Federation
4.2.3 HDFS Snapshots
4.2.4 WebHDFS REST API
4.2.5 DistCp
任务3：YARN新特性
4.3.1 ResourceManager Restart
4.3.2 ResourceManager HA
第5章 大数据数据仓库Hive
任务1：初识Hive
5.1.1 Hive简介
5.1.2 Hive架构
5.1.3 Hive与Hadoop的关系
5.1.4 Hive与传统关系型数据库对比
5.1.5 Hive数据存储
5.1.6 Hive环境部署
任务2：Hive基本操作
5.2.1 DDL操作
5.2.2 DML操作
5.2.3 Hive shell操作
任务3：Hive进阶
5.3.1 Hive函数
5.3.2 Hive常用调优策略
第6章 离线处理辅助系统
任务1：使用Sqoop完成数据迁移
6.1.1 Sqoop简介
6.1.2 导入MySQL数据到HDFS
6.1.3 导出HDFS数据到MySQL
6.1.4 导入MySQL数据到Hive
6.1.5 Sqoop中Job的使用
任务2：工作流调度框架Azkaban
6.2.1 Azkaban简介
6.2.2 Azkaban部署
6.2.3 Azkaban实战
第7章 Spark入门
任务1：初识Spark
7.1.1 Spark概述
7.1.2 Spark优点
7.1.3 Spark生态系统BDAS
任务2：Scala入门
7.2.1 Scala介绍
7.2.2 Scala函数
7.2.3 Scala面向对象
7.2.4 Scala集合
7.2.5 Scala进阶
任务3：获取Spark源码并进行编译
7.3.1 获取Spark源码
7.3.2 Spark源码编译
任务4：第一次与Spark亲密接触
7.4.1 Spark环境部署
7.4.2 Spark完成词频统计分析
第8章 Spark Core
任务1：Spark的基石RDD
8.1.1 RDD概述
8.1.2 RDD常用创建方式
8.1.3 RDD的转换
8.1.4 RDD的动作
8.1.5 RDD的依赖
任务2：RDD进阶
8.2.1 RDD缓存
8.2.2 共享变量（Shared Variables）
8.2.3 Spark核心概念
8.2.4 Spark运行架构
任务3：基于RDD的Spark编程
8.3.1 开发前置准备
8.3.2 使用Spark Core开发词频计数wordcount
8.3.3 使用Spark Core进行年龄统计
第9章 Spark SQL
任务1：Spark SQL前世今生
9.1.1 为什么需要SQL
9.1.2 常用的SQL on Hadoop框架
9.1.3 Spark SQL概述
任务2：Spark SQL编程
9.2.1 Spark SQL编程入口
9.2.2 DataFrame是什么
9.2.3 DataFrame编程
任务3：Spark SQL进阶
9.3.1 Spark SQL外部数据源操作
9.3.2 Spark SQL函数的使用
9.3.3 Spark SQL常用调优
第10章 Spark Streaming
任务1：初始流处理框架及Spark Streaming
10.1.1 流处理框架概述
10.1.2 Spark Streaming概述
任务2：Spark Streaming编程
10.2.1 Spark Streaming核心概念
10.2.2 使用Spark Streaming编程
任务3：Spark Streaming进阶
10.3.1 Spark Streaming整合Flume
10.3.2 Spark Streaming整合Kafka
10.3.3 Spark Streaming常用优化策略

目 录

前言

关于引用作品版权说明

第 1 章 初识 Hadoop 1

本章任务 2

任务 1 大数据概述 2

1.1.1 大数据基本概念 2

1.1.2 大数据对于企业带来的挑战 3

任务 2 Hadoop 概述 4

1.2.1 Hadoop 简介 4

1.2.2 Hadoop 生态系统 7

1.2.3 大数据应用案例 9

任务 3 Hadoop 环境搭建 10

1.3.1 虚拟机安装 11

1.3.2 Linux 系统安装 14

1.3.3 Hadoop 伪分布式环境搭建 31

本章总结 34

本章作业 35

第 2 章 分布式文件系统 HDFS 37

本章任务 38

任务 1 初识 HDFS 38

2.1.1 HDFS 概述 38

2.1.2 HDFS 基本概念 41

2.1.3 HDFS 体系结构 42

任务 2 HDFS 操作 44

2.2.1 HDFS shell 访问 44

2.2.2 Java API 访问 47

任务 3 HDFS 运行机制 50

2.3.1 HDFS 文件读写流程 51

2.3.2 HDFS 副本机制 52

2.3.3 数据负载均衡 53

2.3.4 机架感知 54

任务 4 HDFS 进阶 55

2.4.1 Hadoop 序列化 55

2.4.2 基于文件的数据结构 SequenceFile 60

2.4.3 基于文件的数据结构 MapFile 65

本章总结 67

本章作业 68

第 3 章 分布式计算框架 MapReduce 69

本章任务 70

任务 1 MapReduce 编程模型 70

3.1.1 MapReduce 概述 70

3.1.2 MapReduce 编程模型 71

3.1.3 MapReduce WordCount 编程实例 72

任务 2 MapReduce 进阶 77

3.2.1 MapReduce 类型 77

3.2.2 MapReduce 输入格式 78

3.2.3 MapReduce 输出格式 80

3.2.4 Combiner 81

3.2.5 Partitioner 84

3.2.6 RecordReader 87

任务 3 MapReduce 高级编程 94

3.3.1 Join 的 MapReduce 实现 94

3.3.2 排序的 MapReduce 实现 101

3.3.3 二次排序的 MapReduce 实现 103

3.3.4 合并小文件的 MapReduce 实现 109

本章总结 113

本章作业 114

第 4 章 YARN 与 Hadoop 新特性 115

本章任务 116

任务 1 初识资源调度框架 YARN 116

4.1.1 YARN 产生背景 116
4.1.2 初识 YARN 117
4.1.3 YARN 运行机制 119
任务 2 HDFS 新特性 121
4.2.1 HDFS NameNode HA 122
4.2.2 HDFS NameNode Federation 129
4.2.3 HDFS Snapshots 131
4.2.4 WebHDFS REST API 134
4.2.5 DistCp 135
任务 3 YARN 新特性 135
4.3.1 ResourceManager Restart 135
4.3.2 ResourceManager HA 136
本章总结 139
本章作业 139
第 5 章 大数据数据仓库 Hive 141
本章任务 142
任务 1 初识 Hive 142
5.1.1 Hive 简介 142
5.1.2 Hive 架构 143
5.1.3 Hive 与 Hadoop 的关系 144
5.1.4 Hive 与传统关系型数据库对比 144
5.1.5 Hive 数据存储 145
5.1.6 Hive 环境部署 145
任务 2 Hive 基本操作 146
5.2.1 DDL 操作 147
5.2.2 DML 操作 150
5.2.3 Hive shell 操作 154
任务 3 Hive 进阶 155
5.3.1 Hive 函数 155
5.3.2 Hive 常用调优策略 158
本章总结 163
本章作业 164
第 6 章 离线处理辅助系统 165
本章任务 166
任务 1 使用 Sqoop 完成数据迁移 166
6.1.1 Sqoop 简介 166
6.1.2 导入 MySQL 数据到 HDFS 171
6.1.3 导出 HDFS 数据到 MySQL 177
6.1.4 导入 MySQL 数据到 Hive 179
6.1.5 Sqoop 中 Job 的使用 180
任务 2 工作流调度框架 Azkaban 180
6.2.1 Azkaban 简介 181
6.2.2 Azkaban 部署 182
6.2.3 Azkaban 实战 186
本章总结 189
本章作业 189
第 7 章 Spark 入门 191
本章任务 192
任务 1 初识 Spark 192
7.1.1 Spark 概述 192
7.1.2 Spark 优点 193
7.1.3 Spark 生态系统 BDAS 195
任务 2 Scala 入门 198
7.2.1 Scala 介绍 199
7.2.2 Scala 函数 202
7.2.3 Scala 面向对象 203
7.2.4 Scala 集合 206
7.2.5 Scala 进阶 209
任务 3 获取 Spark 源码并进行编译 211
7.3.1 获取 Spark 源码 211
7.3.2 Spark 源码编译 212
任务 4 第一次与 Spark 亲密接触 214
7.4.1 Spark 环境部署 214
7.4.2 Spark 完成词频统计分析 215
本章总结 216
本章作业 217
第 8 章 Spark Core 219
本章任务 220
任务 1 Spark 的基石 RDD 220
8.1.1 RDD 概述 220
8.1.2 RDD 常用创建方式 221
8.1.3 RDD 的转换 223
8.1.4 RDD 的动作 225
8.1.5 RDD 的依赖 227
任务 2 RDD 进阶 230
8.2.1 RDD 缓存 230

8.2.2 共享变量（Shared Variables）..... 233
8.2.3 Spark 核心概念..... 235
8.2.4 Spark 运行架构..... 236
任务 3 基于 RDD 的 Spark 编程..... 237
8.3.1 开发前置准备..... 237
8.3.2 使用 Spark Core 开发词频计数 WordCount..... 238
8.3.3 使用 Spark Core 进行年龄统计..... 242
本章总结..... 243
本章作业..... 243
第 9 章 Spark SQL..... 245
本章任务..... 246
任务 1 Spark SQL 前世今生..... 246
9.1.1 为什么需要 SQL..... 246
9.1.2 常用的 SQL on Hadoop 框架..... 247
9.1.3 Spark SQL 概述..... 248
任务 2 Spark SQL 编程..... 250
9.2.1 Spark SQL 编程入口..... 250
9.2.2 DataFrame 是什么..... 251
9.2.3 DataFrame 编程..... 252
任务 3 Spark SQL 进阶..... 259
9.3.1 Spark SQL 外部数据源操作..... 259
9.3.2 Spark SQL 函数的使用..... 263
9.3.3 Spark SQL 常用调优..... 266
本章总结..... 269
本章作业..... 269
第 10 章 Spark Streaming..... 271
本章任务..... 272
任务 1 初始流处理框架及 Spark Streaming..... 272
10.1.1 流处理框架概述..... 272
10.1.2 Spark Streaming 概述..... 274
任务 2 Spark Streaming 编程..... 277
10.2.1 Spark Streaming 核心概念..... 278
10.2.2 使用 Spark Streaming 编程..... 282
任务 3 Spark Streaming 进阶..... 286
10.3.1 Spark Streaming 整合 Flume..... 287
10.3.2 Spark Streaming 整合 Kafka..... 290
10.3.3 Spark Streaming 常用优化策略..... 294
本章总结..... 297
本章作业..... 297

第1章

初识 Hadoop

本章重点

※ Hadoop 环境部署

本章目标

※ 了解大数据和 Hadoop 是什么
※ 掌握 Hadoop 的核心构成
※ 了解 Hadoop 生态系统
※ 掌握虚拟机、CentOS 和 Hadoop 的安装

本章任务

学习本章，需要完成以下 3 个工作任务。请记录下来学习过程中所遇到的问题，可以通过自己的努力或访问 kgc.cn 解决。

任务 1：大数据概述

了解大数据的基本概念和基本特征，大数据对企业带来的挑战有哪些。

任务 2：Hadoop 概述

了解 Hadoop 是什么，掌握 Hadoop 的核心构成，了解 Hadoop 生态系统中各个组件的功能。

任务 3：Hadoop 环境搭建

掌握虚拟机、CentOS、Hadoop 的安装。

任务 1　大数据概述

关键步骤如下：

- 了解大数据是什么。
- 了解大数据的特征。
- 了解大数据给企业带来了哪些方面的挑战。

1.1.1　大数据基本概念

1. 大数据概述

相信大家会在各种场合经常听到“大数据”这个词，被誉为数据仓库之父的 Bill Inmon 早在 20 世纪 90 年代就经常将大数据挂在嘴边了。那么到底什么是大数据呢？这是我们本任务中要了解的。

我们现在生活的时代是一个数据时代，近年来随着互联网的的高速发展，每分每秒都在产生数据，那么产生的这些数据如何进行存储和相应的分析处理呢？在这种情况下，各大公司纷纷研发和采用一批新技术，主要包括分布式文件系统、分布式计算框架等等，这些是我们需要学习和掌握的。

互联网周刊对大数据的定义为：“大数据”的概念远不止大量的数据（TB）和处理大量数据的技术，或者所谓的“4 个 V”之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，而这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨

大价值的产品和服务，或深刻的洞见，最终形成变革之力。

2. 大数据特征

（1）数据量大（Volume）

随着网络技术的发展和普及，每时每刻都会产生大量的数据。在我们的日常生活中，比如说你在电商网站上购物、在直播平台看直播、阅读新闻等等操作，都会产生很多的日志，每分每秒产生的数据量是非常巨大的。

（2）类型繁多（Variety）

大数据中最常见的类型是日志，除了日志之外常见的还有音频、视频、图片等等。由于不同类型的数据没有明显的模式，具有多样性的特点，这对于数据的处理要求也会更高。

（3）价值密度低（Value）

现阶段每时每刻产生的数据量已经很大了，如何从大量的日志中提取出来我们所需要的、对我们有价值的东西是最重要的。数据量越来越大，那么里面必然会存在着大量与我们所需要的不相干的信息，如何更迅速地完成数据的价值提炼，是大数据时代有待解决的问题。

（4）处理速度快（Velocity）

传统的离线处理的时效性不高，换句话说时延是非常高的。随着时代的发展，对时效性要求越来越高，需要实时对产生的数据进行分析处理，而不是采用原来的批处理方式。

1.1.2　大数据对于企业带来的挑战

1. 对现有数据库的挑战

随着互联网时代的到来，现在产生的数据如果想存储在传统数据库里面是不太现实的，即便传统的数据库有集群的概念，但是传统的数据库不能处理数 TB 级别的数据分析。而且现阶段产生的数据类型有很多，有些类型的数据是没办法使用结构化数据查询语言（SQL）来处理的。

2. 实时性的技术挑战

我们知道数据所产生的价值是随着时间的流逝而大大降低的，所以当数据产生后我们要尽可能快的进行处理。最典型的就是电商网站的推荐系统，早些年的推荐系统都是基于批处理来进行的，比如每隔半天或者一天进行计算然后再进行推荐，这样就会有很大的延时，对于订单的转换而言有效果但不很好。如果能做到实时推荐，那么肯定能大大提高公司的营收。

传统的离线批处理对处理时间的要求并不高。实时处理的要求是区别大数据应用和传统数据库技术、或者离线技术的关键差别之一。

3. 对数据中心、运维的挑战

每天创建的数据量正呈爆炸式增长，那么这么多数据如何进行高效的收集、存储、计算都是数据中心要面临的一个非常棘手的问题。要处理快速增长的数据量所需要的机器日益增多，那么对于运维团队来说压力也会增加。

至此，在掌握以上相关知识后，任务 1 就可以完成了。

任务 2　Hadoop 概述

关键步骤如下：

- 认知 Hadoop 是什么。
- 了解 Hadoop 的发展史。
- 掌握 Hadoop 中的核心组件及功能。
- 了解 Hadoop 常用的发行版本。
- 了解 Hadoop 生态系统中常用的处理框架。
- 了解大数据在企业中的应用案例。

1.2.1　Hadoop 简介

1. 什么是 Hadoop

Hadoop 是 Apache 基金会旗下的一个分布式系统基础架构。主要包括：分布式文件系统 HDFS（Hadoop Distributed File System）、分布式计算系统 MapReduce 和分布式资源管理系统 YARN。可以使得用户在不了解分布式底层细节的情况下，开发分布式程序、充分利用集群的分布式能力进行运算和存储。以 Apache Hadoop 为生态系统的框架是目前分析海量数据的首选。

针对第一节中描述的大数据，我们如何对这些数据进行分析或者提取出我们所需要的有价值的信息呢？我们可以采用 Hadoop 以及生态圈提供的分布式存储和分布式计算的功能来完成。

2. Hadoop 发展史

（1）2002 年，Doug Cutting 团队开发了网络搜索引擎 Nutch，这就是 Hadoop 的前身；

（2）2003—2004 年，Google 两篇论文诞生：GFS 和 MapReduce；

（3）2006 年，为致力于 Hadoop 技术的发展，Doug Cutting 加入 Yahoo!；

（4）2008 年 1 月，Hadoop 成为 Apache 顶级项目，并在同年 7 月打破最快排序 1TB 数据的世界纪录；

（5）2008 年 9 月，Hive 成为 Hadoop 子项目；