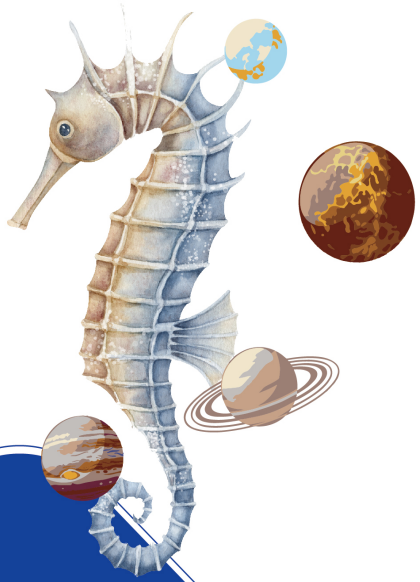
国家一级出版社
全国百佳图书出版单位
China University of Mining and Technology Press

策划编辑：杨　洋
责任编辑：齐　畅
封面设计：刘文东



大数据技术基础

ISBN 978-7-5646-6600-2
9 787564 666002 >

定价：45.00元

免费提供 精品教学资料包
服务热线：400-615-1233
www.xinsijiaocai.com

大数据技术基础

主编　袁　帅　冯明卿

中国矿业大学出版社
China University of Mining and Technology Press

大数据系列精品教材

PPT课件　习题答案
拓展学习资料

- 以工作过程为导向，“教学做评”一体贯通
- 采用六步教学法，确保学习效果可测可控

大数据技术基础

主编　袁　帅　冯明卿

中国矿业大学出版社
China University of Mining and Technology Press

大数据系列精品教材

大数据
技术基础

主　编　袁　帅　冯明卿
副主编　李俊艳　石　艳　韩　丁
参　编　王晓燕　卢　珊　赵　波
　　刘沣啸　李亚栋

中国矿业大学出版社
·徐　州·

图书在版编目（CIP）数据

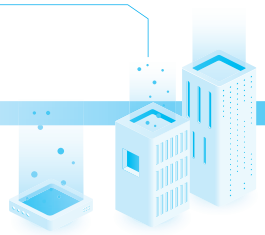
大数据技术基础/袁帅，冯明卿主编. --徐州：中国矿业大学出版社，2024.12. -- ISBN 978-7-5646-6600-2

Ⅰ.TP274

中国国家版本馆CIP数据核字第202459LG85号

书　　名　大数据技术基础（Dashuju Jishu Jichu）
主　　编　袁　帅　冯明卿
责任编辑　齐　畅
出版发行　中国矿业大学出版社有限责任公司
（江苏省徐州市解放南路　邮编221008）
营销热线　（0516）83885370　　83884103
出版服务　（0516）83995789　　83884920
网　　址　http://www.cumtp.com　E-mail: cumtpvip@cumtp.com
印　　刷　三河市龙大印装有限公司
开　　本　787 mm×1092 mm　1/16　**印张**　13.5　**字数**　295千字
版次印次　2024年12月第1版　　2024年12月第1次印刷
定　　价　45.00元
（图书出现印装质量问题，本社负责调换）

前 言



党的二十大报告提出：“推动战略性新兴产业融合集群发展，构建新一代信息技术、人工智能、生物技术、新能源、新材料、高端装备、绿色环保等一批新的增长引擎。”战略性新兴产业代表新一轮科技革命和产业变革方向，是推动经济发展质量变革、效率变革、动力变革的关键力量。党的二十大报告为我国新一代信息技术产业发展指明了方向。新一代信息技术的高速发展，不仅可为加快制造强国、网络强国和数字中国建设提供坚实的支撑，而且也能有力促进百行千业转型升级，为推动我国经济高质量发展增添新动能。

大数据技术代表着当今先进生产力的发展方向，信息技术的广泛应用使信息的重要生产要素和战略资源的作用得以发挥，使人们能更高效地进行资源优化配置，提高社会劳动生产率和社会运行效率，从而推动传统产业不断升级。大数据产业已成为新时期经济增长的重要引擎，有力地促进了可持续发展，深刻地改变着人类的生产、生活方式。增强个体在信息社会的适应力与创造力，提升国民信息素养，对个人的生活、学习和工作，对全面建设社会主义现代化国家具有重大意义。

本书主要面向高等职业教育和应用型本科大数据相关专业学生、对大数据感兴趣的技术开发人员。

本书紧紧围绕大数据技术基础讲解以下内容。

模块 1 主要讲解大数据的定义、大数据的特征、大数据的产生及发展、大数据的应用场景、大数据生态系统。通过本模块的学习，学生可以了解大数据的前世今生。

模块 2 主要讲解 Hadoop 的基本构成及生态系统，学生通过学习可以快速了解 Hadoop 的发展及生态系统知识。

模块 3 主要讲解 HDFS 的体系架构、存储原理、读写流程等知识，学生通过学习 HDFS 环境搭建、HDFS 的 Shell 操作、HDFS 的 API 实际操作，能够搭建环境和使用 HDFS 的常用命令及 API。

模块 4 主要讲解 YARN 体系架构、工作流程、调度器、环境搭建以及常用命令

知识。

模块 5 主要讲解分布式并行计算框架 MapReduce，主要内容有 MapReduce 体系架构、工作流程，学生可以结合所学知识进行编程实践。

模块 6 主要讲解分布式服务协调框架 ZooKeeper，主要内容有 ZooKeeper 体系架构、工作流程，学生可利用 ZooKeeper 进行环境的搭建及常用命令的实操。

模块 7 主要讲解常见的数据采集工具，主要包括 Sqoop、Flume、DataX 的相关介绍，学生可据此进行相应工具的环境搭建及常用操作。

模块 8 主要讲解分布式数据仓库 Hive，主要内容有 Hive 体系架构、运行机制、工作原理、重要概念，学生可据此进行相应的环境搭建及应用实践。

模块 9 主要讲解 HBaseNoSQL 数据库，主要内容有 HBase 数据模型、体系架构、运行原理，学生可据此进行相应的环境搭建及应用实践。

模块 10 主要讲解分布式消息队列 Kafka，主要内容有 Kafka 体系架构、运行机制，学生可据此进行相应的环境搭建及应用实践。

模块 11 主要讲解分布式计算框架 Spark，主要内容有 Spark 体系架构、执行过程，学生可据此进行相应的环境搭建及应用实践。

模块 12 主要讲解流计算框架，主要内容有流计算处理过程介绍、Flink 流处理框架、Spark Streaming 流处理框架，学生可据此进行流处理框架的搭建和编程实战。

本书在结构和案例设计、知识和技术讲解、配套资源开发及学习评价方面具有以下特色。

（1）全面融入“岗课赛证”，培养核心技能。教材立足《大数据工程技术人员国家职业技术技能标准》及大数据核心技术课程标准，依据全国职业院校技能大赛“大数据应用开发”赛项、华为认证大数据工程师、大数据应用开发（Java）职业技能等级证书（中级），引入大数据最新技术，培养学生核心技能。

（2）采用“模块 – 任务”驱动教学方式，教学目标明确。本书包含 12 个模块，每个模块按照学习目标、知识讲解、任务实施、知识拓展、任务考评、任务实训、任务自测的结构组织教学内容。

（3）采用“三实教学”，培养实践能力。通过实际场景、实践项目、实战检验，将真实的企业工作模式、操作环境融入教材，将电力数据统计和电商平台消费排名等

项目融入教材，学生可通过实践项目对知识点进行巩固和加强，将零碎的知识糅合在一起，对知识有一个整体性的认识，并进行实战检验。

（4）实现学习资源立体化。本书相关内容均已录制成视频，学生只需扫描书中提供的二维码便可以观看视频掌握相关知识与技能。同时，本书还提供了案例的素材与效果文件，以及丰富的微课视频、PPT 课件、教案、题库、项目案例数据和代码等教学资源。

（5）全面融入课程思政，落实德技并修。教材通过知识拓展，列举中国科技企业自主研发大数据技术、开发大数据应用等事实，增强学生的民族自豪感，培养学生的爱国主义精神、精益求精和创新的工匠精神、艰苦奋斗克服困难的钻研精神等。

本书采用校企合作的方式，由郑州电力高等专科学校教师与慧科教育科技集团有限公司、国网河南省电力公司经济技术研究院共同编写，项目案例由公司提供。本书由袁帅、冯明卿任主编，李俊艳、石艳、韩丁（国网河南省电力公司经济技术研究院）任副主编，王晓燕、卢珊、赵波、刘沣啸、李亚栋（慧科教育科技集团有限公司）参与编写。

由于编者水平有限，书中难免存在不足之处，欢迎广大读者批评指正。如有问题可发至邮箱 510582939@qq.com，为我们今后出版更好的图书提供帮助。

编　者

课程介绍 – 大数据核心技术

目 录

模块 6

ZooKeeper 分布式服务协调框架

模块 7

数据采集工具

模块 8

Hive 分布式数据仓库

模块 9

HBase NoSQL 数据库

模块 10
Kafka 分布式消息队列

模块 11
Spark 分布式计算框架

模块 12

流计算框架

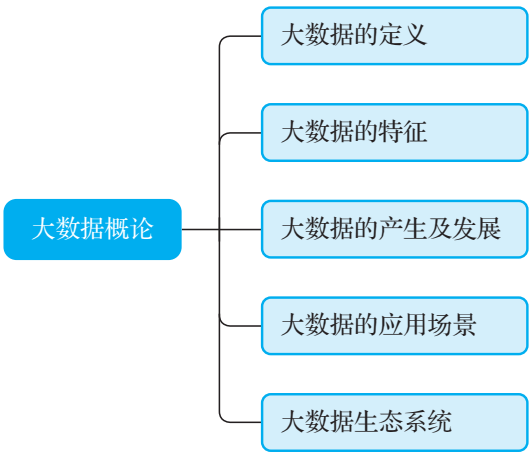
模块 1

大数据概论

大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。大数据技术是指从各种类型的数据中快速获得有价值信息的能力。适用于大数据的技术包括大规模并行处理（massively parallel processing，MPP）数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

本模块主要介绍大数据的定义、大数据的特征、大数据的产生及发展、大数据的应用场景、大数据生态系统。通过学习本模块，可以了解大数据的前世今生。

以下为本模块所要学习的主要内容。



学习目标

知识目标

（1）熟悉大数据的定义。

（2）掌握大数据的特征。

（3）了解大数据的产生及发展。

（4）熟悉大数据的应用场景。

（5）了解大数据生态系统。

能力目标

能够熟练表述大数据的定义、发展及大数据生态系统。

素质目标

（1）掌握国内大数据技术发展及政府在政策方面对大数据技术的支持。

（2）提高系统化思维和问题解决能力，能够适应技术栈变化带来的挑战。

1.1 大数据的定义

大数据（big data）是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要运用新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据从字面上理解就是海量的数据，但在技术上它包括对这些海量数据的采集、过滤、清洗、存储、处理、查看等部分，每一个部分均有相关技术框架。

大数据概述

1.2 大数据的特征

（1）数据体量巨大。百度资料表明，其新首页导航每天需要提供的数据超过1.5 PB（1 PB=1 024 TB），这些数据如果打印出来将超过5 000亿张A4纸。有资料证实，到目前为止，人类生产的所有印刷材料的数据量仅为200 PB。

（2）数据类型多样。现在的数据类型不仅是文本形式，更多的是图片、视频、音频、地理位置信息等多类型的数据，个性化数据占绝对多数。

（3）处理速度快。数据处理遵循“1秒定律”，可从各种类型的数据中快速获得高价值的信息。

（4）价值密度低。以视频为例，一小时的视频，在不间断的监控过程中，可能有用的数据仅仅只有一两秒。

1.3 大数据的产生及发展

1.3.1 大数据的产生

大数据的产生可以追溯到20世纪五六十年代，当时的美国政府和企业开始使用电

子计算机处理数据，但计算机还比较原始，数据的规模和处理能力都非常有限。直到20世纪八九十年代，随着计算机技术的不断发展，数据的规模和处理能力才得到了显著的提高。

1.3.2 大数据的发展

（1）第一阶段。在大数据发展的第一阶段，主要的技术手段是分布式存储和处理技术。Hadoop是其中著名的开源分布式存储和处理框架，它由Apache基金会开发。Hadoop使用Hadoop分布式文件系统（Hadoop distributed file system，HDFS）存储海量数据，并使用MapReduce处理数据。这种技术可以让数据在多个计算节点之间进行分布式存储和处理，从而加快数据处理速度并提高可靠性。

（2）第二阶段。在大数据发展的第二阶段，出现了更多的开源分布式存储和处理框架，如Spark和Storm。Spark是一种内存计算框架，它使用弹性分布式数据集（resilient distributed datasets，RDD）作为基本数据结构，具有较快的计算速度和高效的内存管理能力。Storm是一个实时数据处理框架，可以实时处理流式数据。

（3）第三阶段。在大数据发展的第三阶段，随着机器学习和人工智能的发展，大数据处理技术也开始融合这些技术。这个阶段的主要技术手段包括深度学习、自然语言处理、图像处理等。这些技术可以让大数据处理变得更加智能化和自动化，从而可以为企业和个人提供更多的价值。

1.4 大数据的应用场景

大数据已经被广泛应用于各个领域，包括商业、医疗、金融、政府等。

（1）商业。大数据已经成为商业领域中不可或缺的一部分。大数据可以帮助企业进行市场分析、客户行为分析、产品开发等。通过对大数据的深度挖掘和分析，企业可以更好地了解消费者的需求和喜好，并可以根据这些信息调整自己的商业战略。

（2）医疗。大数据在医疗领域也有着广泛的应用。医疗数据的规模很大、复杂度很高，通过大数据分析可以挖掘潜在的医学知识、寻找治疗方案，从而提高医疗效率和治疗效果。例如，大数据可以用于医学图像分析、基因组学研究、病例分析等。

（3）金融。金融领域是大数据应用的另一个重要领域。大数据可以用于风险管理、信用评估、投资决策等。例如，金融机构可以通过大数据分析来预测市场趋势和风险，从而进行更高效的投资。

（4）政府。政府也是大数据应用的一个重要领域。政府可以通过大数据分析来更好地管理公共资源、优化公共服务和预测社会需求。例如，政府可以通过大数据分析来优化城市交通、预测自然灾害、加强环境保护等。

1.5 大数据生态系统

大数据生态系统如图 1-1 所示。

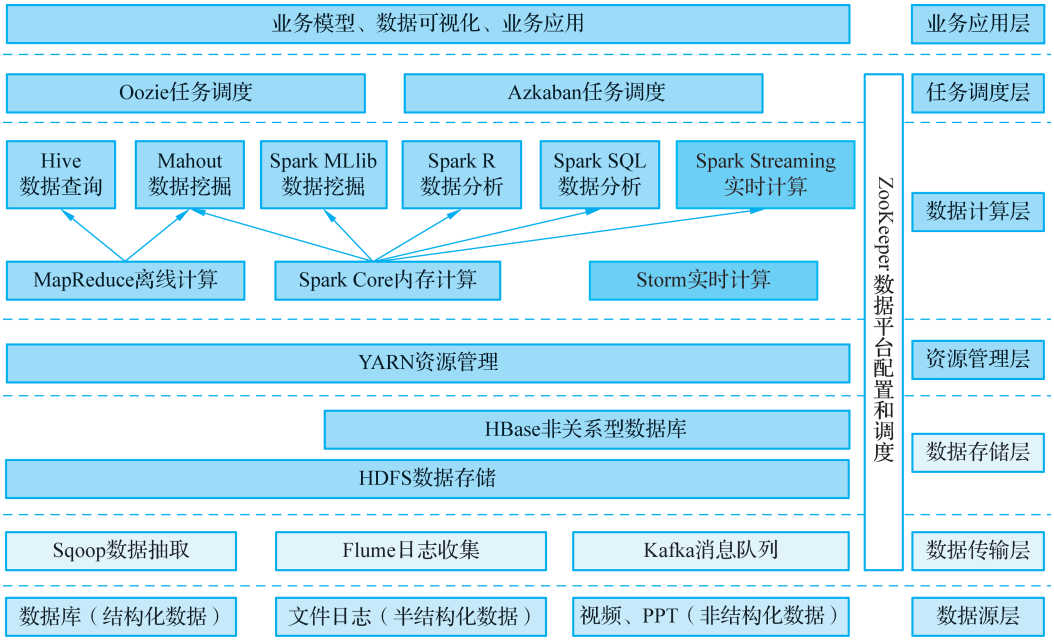


图 1-1 大数据生态系统

图 1-1 中涉及的技术名词解释如下。


大数据生态系统

（1）Sqoop。Sqoop 是一款开源工具，主要用于在 Hadoop（Hive）与传统的数据库（MySQL）之间进行数据传递，可以将一个关系型数据库（如 MySQL、Oracle 等）中的数据导入 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入关系型数据库中。

（2）Flume。Flume 是 Cloudera 提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输系统，Flume 支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume 提供对数据进行简单处理并写到各种数据接收方（可定制）的能力。

（3）Kafka。Kafka 是一种高吞吐量的分布式消息系统，有以下特性。

① 通过时间复杂度为 O（1）的磁盘数据结构提供消息的持久化，这种结构对于即使数以太字节（TB）的消息存储也能够保持长时间的稳定性能。

② 高吞吐量。即使是非常普通的硬件，Kafka 也可以支持每秒数百万的消息。

③ 支持通过 Kafka 服务器和消费机集群来实现消息的分区处理。

④ 支持 Hadoop 并行数据加载。

（4）Storm。Storm 为分布式实时计算提供了一组通用原语，可被用于“流处理”

之中，实时处理消息并更新数据库。这是管理队列及工作者集群的另一种方式。Storm也可被用于连续计算（continuous computation），对数据流做连续查询，在计算时就将结果以流的形式输出给用户。

（5）Spark。Spark是当前最流行的开源大数据内存计算框架，可以基于Hadoop上存储的大数据进行计算。

（6）Oozie。Oozie是一个管理Hadoop作业（job）的工作流程调度管理系统。Oozie协调作业就是通过时间（频率）和有效数据触发当前的Oozie工作流程的。

（7）HBase。HBase是一个分布式的、面向列的开源数据库。HBase不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。

（8）Hive。Hive是一个基于Hadoop的数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的SQL查询功能，可以将SQL语句转换为MapReduce任务来运行。它的优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

（9）R语言。R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，是一个用于统计计算和统计制图的优秀工具。

（10）Mahout。Mahout是一个可扩展的机器学习和数据挖掘库。当前Mahout支持的4个主要用例如下。

① 推荐挖掘：收集用户动作并以此给用户推荐可能喜欢的事物。

② 聚集：收集文件并进行相关文件分组。

③ 分类：从现有的分类文档中学习，寻找文档中的相似特征，并为无标签的文档进行正确的归类。

④ 频繁项集挖掘：将一组项分组，并识别哪些个别项会经常一起出现。

（11）ZooKeeper。ZooKeeper是Google的Chubby锁服务的一个开源实现。它是一个针对大型分布式系统的可靠协调系统，提供的功能包括配置维护、名字服务、分布式同步、组服务等。ZooKeeper的目标就是封装好复杂易出错的关键服务，将简单、易用的接口提供给用户，用户可以使用ZooKeeper提供的接口编写出高质量的分布式应用。

1.6 知识拓展——华为云MRS：国云大数据引擎

随着互联网技术的飞速发展，大数据已成为企业和社会面临的重要挑战。Hadoop作为开源的分布式计算平台，为解决大数据问题提供了强大的工具。然而，企业自行部署Hadoop系统面临成本高、周期长、运维复杂等难题。针对这些问题，华为云推出了大数据MapReduce服务（MRS），为企业提供了一个高性能、低成本、安全可靠、易于运维的全栈大数据平台。

MRS的优势如下。

（1）高性能：MRS 充分利用华为云计算和存储优势，结合优化的 Hadoop 内核，实现数据的高效处理和存储。

（2）低成本：企业无须自行搭建和维护 Hadoop 集群，降低了硬件和软件投入，同时减少了运维成本。

（3）高安全性：华为云提供多层次的安全保障措施，确保数据的安全性和隐私性。

（4）易运维：MRS 提供一站式的大数据集群云服务，简化了运维过程，降低了运维难度和复杂度。

（5）高可靠性：基于华为 FusionInsight 大数据企业级平台，MRS 具备高可用性和容错能力，确保系统稳定运行。

（6）海量数据分析与存储：MRS 支持 Hadoop、Spark、HBase、Kafka、Storm 等大数据组件，满足各种海量数据分析和存储需求。

（7）实时数据处理：MRS 具备实时数据处理能力，帮助企业快速响应市场变化和业务需求。

（8）定制开发能力：MRS 提供灵活的定制开发接口，企业可以根据业务需求进行定制开发，构建符合自身需求的大数据系统。

华为云 MRS 为企业提供了一个高性能、低成本、安全可靠、易于运维的全栈大数据平台，帮助企业快速构建海量数据信息处理系统，并通过实时与非实时的分析挖掘，发现全新价值点和企业商机。

华为云 MRS 凭借其强大的技术实力和卓越的服务能力，已经成为推动企业智慧转型的重要引擎。随着数字化转型的不断深入和大数据技术的不断发展，MRS 将继续发挥其在企业智慧转型中的关键作用，为中国企业的未来发展注入新的活力和动力。

1.7 任务自测

1. 选择题

（1）下列（　　）不属于大数据生态系统的组件。

A. HDFS　　B. ZooKeeper　　C. SmartBI　　D. Spark

（2）Spark 出现在大数据发展的（　　）阶段。

A. 第一　　B. 第二　　C. 第三　　D. 第四

2. 简答题

（1）大数据的特征有哪些？

（2）大数据的应用场景有哪些？

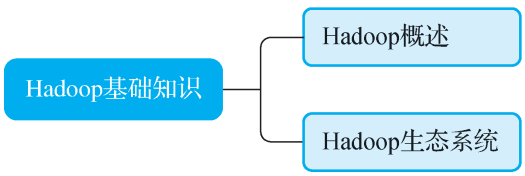
（3）什么是 Hive ？

模块 2
Hadoop 基础知识

Hadoop 是一个开源的分布式计算框架，用于解决大数据存储和处理的问题。它基于 Google 的 MapReduce 论文和 Google 文件系统（GFS）的思想，由 Apache 软件基金会开发和维护。Hadoop 的设计目标是作为一个能够处理大规模数据集的应用程序，提供高可靠性、高性能和高可扩展性的数据处理服务。

本模块主要介绍 Hadoop 的基本构成及生态系统，使学生能够快速了解 Hadoop 的发展及生态系统。

以下为本模块所要学习的主要内容。



学习目标

知识目标

（1）了解 Hadoop 的起源及历史。

（2）了解 Hadoop 的技术优势。

（3）熟悉 Hadoop 生态系统。

能力目标

（1）能够清晰表述 Hadoop 的起源及发展历史。

（2）能够结合已学知识自主学习国内厂商的相关技术。

素质目标

了解国内厂商对 Hadoop 的优化及政府在政策方面对大数据技术的支持。

2.1 Hadoop 概述

以 Hadoop 分布式文件系统（HDFS）和 MapReduce（Google MapReduce 的开源实现）为核心的 Hadoop 为用户提供了系统底层细节透明的分布式基础架构。HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上，形成分布式系统；MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。因此，用户可以利用 Hadoop 轻松地组织计算机资源，从而搭建自己的分布式计算平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。

2.2 Hadoop 生态系统

2.2.1 Hadoop 的基本模块

Hadoop 包括以下 4 个基本模块。

（1）Hadoop 基础功能库：支持其他 Hadoop 模块的通用程序包。

（2）HDFS：一个分布式文件系统，能够以高吞吐量访问应用中的数据。

（3）YARN：一个作业调度和资源管理框架。

（4）MapReduce：一个基于 YARN 的大数据并行处理程序。

2.2.2 Hadoop 的其他项目

除了基本模块，Hadoop 还包括以下项目。

（1）Ambari：基于 Web，用于配置、管理和监控 Hadoop 集群。支持 HDFS、MapReduce、Hive、HCatalog、HBase、ZooKeeper、Oozie、Pig 和 Sqoop。Ambari 还提供显示集群健康状况的仪表盘，如热点图等。Ambari 以图形化的方式查看 MapReduce、Pig 和 Hive 应用程序的运行情况，因此可以通过对用户友好的方式诊断应用的性能问题。

（2）Avro：数据序列化系统。

（3）Cassandra：可扩展的、无单点故障的 NoSQL 多主数据库。

（4）Chukwa：用于大型分布式系统的数据采集系统。

（5）HBase：可扩展的分布式数据库，支持大表的结构化数据存储。

（6）Hive：数据仓库基础架构，提供数据汇总和命令行即时查询功能。

（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：可高速处理 Hadoop 数据的通用计算引擎。Spark 提供了一种简单而富有表达能力的编程模式，支持 ETL、机器学习、数据流处理、图像计算等多种应用。

（10）Tez：完整的数据流编程框架，基于 YARN 建立，提供强大而灵活的引擎，可执行任意有向无环图（directed acyclic graph，DAG）数据处理任务，既支持批处理又支持交互式的用户场景。Tez 已经被 Hive、Pig 等 Hadoop 生态圈的组件所采用，用来代替 MapReduce 作为底层执行引擎。

（11）ZooKeeper：用于分布式应用的高性能协调服务。

2.3 任务实施

2.3.1 安装虚拟机软件与虚拟机


安装虚拟机软件


安装虚拟机系统

VMware Workstation 是一款功能强大的桌面虚拟计算机软件，能够为用户提供可在单一的桌面上同时运行不同的操作系统和进行开发、测试、部署新的应用程序的最佳解决方案。

本书使用 VMware Workstation-pro-17.5，其操作系统为 Windows，需要下载 Windows 版软件并按照安装要求完成安装。

Linux 是一个供用户免费使用和自由传播的类 UNIX 操作系统，是一个基于 POSIX 和 UNIX 的多用户、多任务、支持多线程和多 CPU 的操作系统。本书采用基础的 Linux 发行版 CentOS 作为虚拟机环境部署相关软件，下载链接为：https://mirrors.aliyun.com/centos/7.9.2009/isos/x86_64/。打开网站后，选择 CentOS-7-x86_64-DVD-2207-02.iso 下载安装。本书项目实践需要采用 3 个虚拟机实现，主机名分别定义为 hadoop001、hadoop002、hadoop003。

在苹果计算机上，可以安装免费虚拟机软件 VirtualBox 来实现与 VMware Workstation 同样的功能。访问 VirtualBox 官网（https://www.virtualbox.org/），下载 dmg 格式的安装文件进行安装即可，建议选择较新版本 VirtualBox V7.0.14，考虑到网上资料丰富，本书不再赘述。虚拟机软件安装成功后，可以按照以下步骤继续安装虚拟机。

2.3.1.1 安装 CentOS 虚拟机

（1）打开 VMware，选择 hadoop001 服务器，打开虚拟机启动界面，单击“编辑虚拟机设置”链接，如图 2-1 所示。

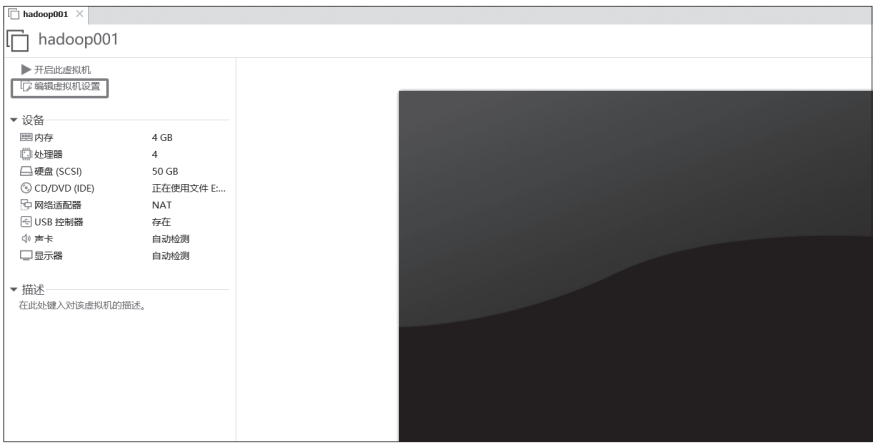


图 2-1　单击“编辑虚拟机设置”链接

（2）进入虚拟机编辑界面，选择“使用 ISO 映像文件”，如图 2-2 所示。

图 2-2　选择使用 ISO 映像文件

（3）单击“开启此虚拟机”链接，进入操作系统安装界面，选择“Install CentOS 7”选项，如图 2-3 所示。

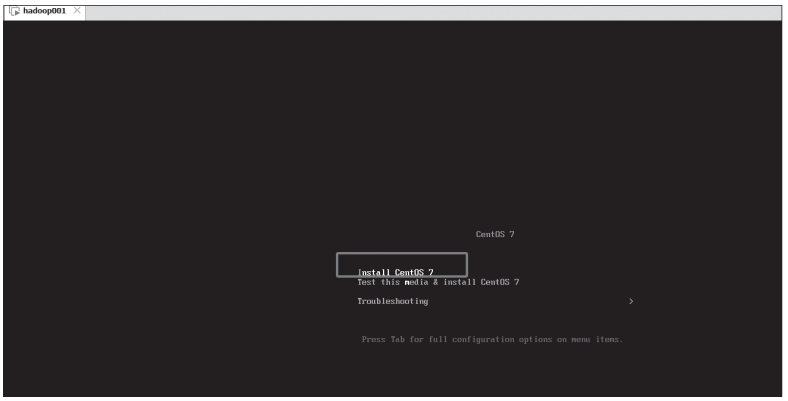


图 2-3　选择“Install CentOS 7”选项

（4）进入系统安装界面，配置 CentOS 7 的语言为“简体中文（中国）”，如图 2-4 所示。

图 2-4　选择安装语言

（5）配置操作系统界面如图 2-5 所示。

图 2-5　配置操作系统界面

（6）进入“日期和时间”界面，配置时区，单击“完成”按钮。

（7）进入“软件选择”界面，选中“带 GUI 的服务器”单选按钮，单击“完成”按钮，如图 2-6 所示。

图 2-6　选中“带 GUI 的服务器”单选按钮

（8）进入“安装目标位置”界面，保持默认即可，单击“完成”按钮，如图 2-7 所示。

（9）进入“网络和主机名”界面，开启网络并修改主机名为“hadoop001”，单击“完成”按钮，如图 2-8 所示。

图 2-7 “安装目标位置”界面

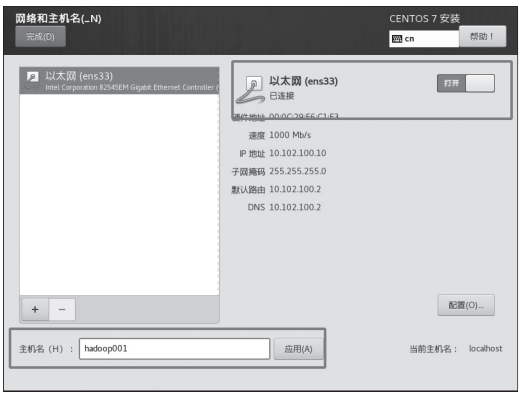


图 2-8 开启网络并修改主机名

（10）进入“ROOT 密码”界面，配置 Root 账户密码，如图 2-9 所示。选择“创建用户”选项，进入“创建用户”界面，创建用户“hadoop”，如图 2-10 和图 2-11 所示。

（11）等待操作系统安装完成后单击“重启”按钮，如图 2-12 所示。

图 2-9 配置 Root 账户密码

图 2-10 选择“创建用户”选项

图 2-11 创建用户“hadoop”

图 2-12 单击“重启”按钮

2.3.1.2 复制虚拟机

（1）打开 VMware 软件，选中创建的虚拟机，右击并选择“管理”→“克隆”选项，复制虚拟机，如图 2-13 所示。

图 2-13 复制虚拟机

（2）打开“克隆虚拟机向导”对话框，单击“下一页”按钮进入下一步操作，如图 2-14 所示。

（3）默认选中“克隆自”选项组中的“虚拟机中的当前状态”单选按钮，单击“下一页”按钮进入下一步操作，如图 2-15 所示。

图 2-14 “克隆虚拟机向导”对话框

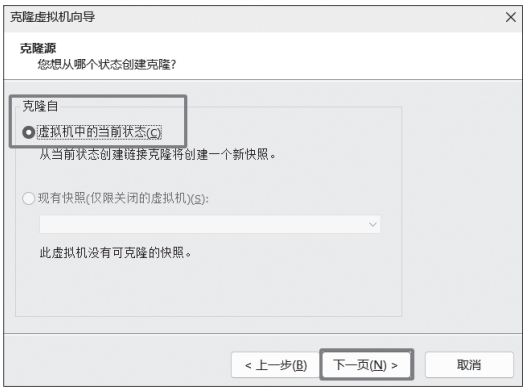


图 2-15 克隆状态选择

（4）在“克隆类型”界面选中“创建完整克隆”单选按钮，单击“下一页”按钮进入下一步操作，如图 2-16 所示。

（5）在“新虚拟机名称”界面中设置虚拟机名称为“hadoop002”，虚拟机位置为本地存储，如图 2-17 所示。

（6）等待虚拟机复制完成，如图 2-18 和图 2-19 所示。

（7）按照以上步骤再复制一个虚拟机，如图 2-20 所示。

图 2-16　选中“创建完整克隆”单选按钮

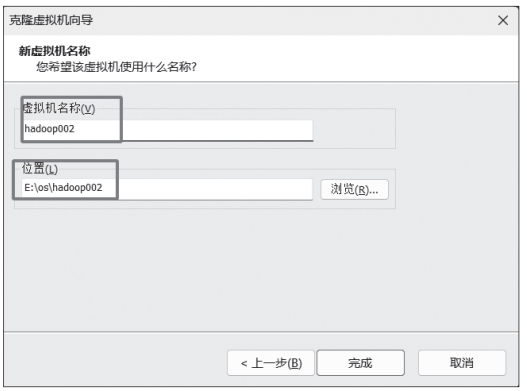


图 2-17　设置虚拟机名称及位置

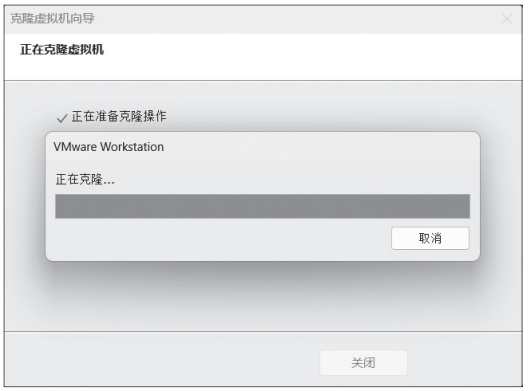


图 2-18　正在复制虚拟机

图 2-19　虚拟机复制完成

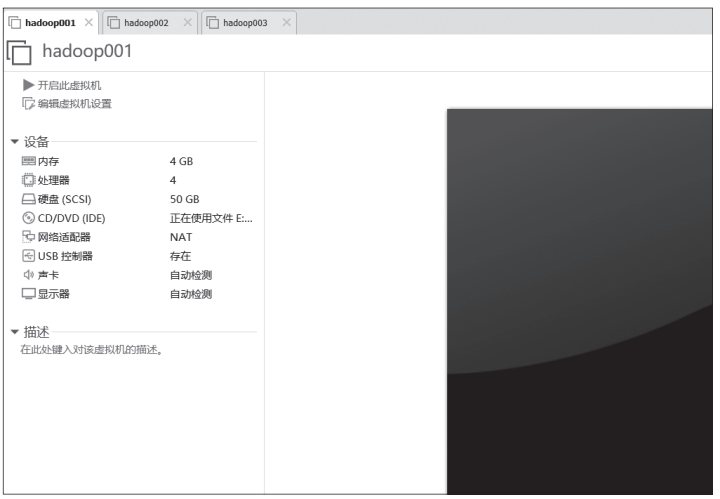


图 2-20　虚拟机创建完成

2.3.2　安装远程服务器管理工具

远程服务器管理工具是允许用户管理远程服务器的实用工具。它通常提供图形用户界面（GUI），允许用户执行各种任务。例如，配置服务器、管理服务器用户、管理

服务器资源和监视服务器性能。

市面上有很多远程服务器管理工具可用，选择远程服务器管理工具时需要考虑的一些因素包括特性、易用性、价格和可用的支持。

远程服务器管理有以下协议。

（1）SSH。SSH（secure shell）是一种远程管理协议，允许用户控制和修改远程服务器。SSH 通常用于登录远程服务器、执行命令和传输文件。SSH 是一种安全协议，通过加密来保护数据不被拦截。

SSH 协议主要有以下几个版本：SSH-1、SSH-2 和 OpenSSH。SSH-1 由 Tatu Ylönen 开发。SSH-2 是 SSH-1 的改进版本，更安全且功能更强大。OpenSSH 是 SSH 协议的免费开源实现，也是 Linux 操作系统默认使用的 SSH 实现。

（2）RDP。RDP（remote desktop protocol）即远程桌面协议，是一种远程管理工具，允许用户控制和修改远程服务器。RDP 通常用于登录远程服务器、执行命令和传输文件。RDP 是一种私有协议，不像 SSH 那样安全。

（3）VNC。虚拟网络计算（virtual network computing，VNC）是一种远程管理协议，允许用户控制和修改远程服务器。VNC 通常用于登录远程服务器、执行命令和传输文件。VNC 不像 SSH 和 RDP 那样安全。

MobaXterm 是一款 SSH 客户端，能帮助人们在 Windows 操作系统下去连接并操作 Linux 服务器，也是一款增强型终端、X 服务器和 UNIX 命令集工具箱。MobaXterm 可以开启多个终端窗口，轻松地使用 UNIX/Linux 上的 GNU UNIX 命令。MobaXterm 还有很强的扩展能力，可以集成插件来运行 GCC、Perl、Curl、Tcl/Tk/Expect 等程序。MobaXterm 分为免费开源版和收费专业版，免费开源版又分为便捷版（解压即用）和安装版（需要一步步安装）。

MobaXterm 支持 SSH、X11、RDP、VNC、FTP、MOSH 等连接，也支持 UNIX 命令，如 bash、ls、cat、sed、grep、awk、rsync 等。连接 SSH 终端后支持 SFTP 传输文件。

MobaXterm 的下载链接为：https://mobaxterm.mobatek.net/download-home-edition.html，下载界面如图 2-21 所示。

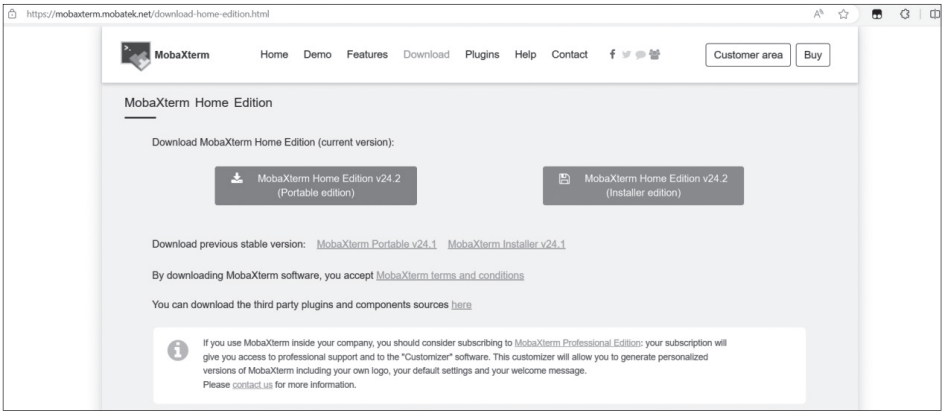


图 2-21　MobaXterm 下载界面

在 Windows 中安装 MobaXterm 工具的步骤如下。

（1）下载 Portable edition 免安装版本，如图 2-22 所示。

图 2-22　下载 Portable edition 免安装版本

（2）解压下载的压缩包，选择安装位置，完成解压，如图 2-23 所示。

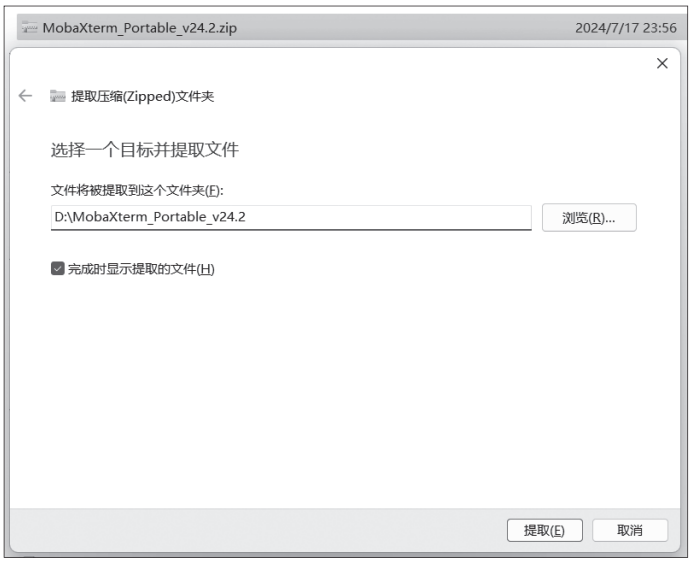


图 2-23　解压软件

（3）解压后，文件目录内有 3 个文件，双击“MobaXterm_Personal_24.2.exe”文件即可打开软件，如图 2-24 所示。

图 2-24　免安装版 MobaXterm

（4）在软件界面左侧窗格中右击“User sessions”选项，在弹出的快捷菜单中选择“New session”选项，弹出“Session settings”对话框，单击“SSH”按钮，在打开的界

面中输入 IP 地址（Remote host）、用户名（Specify username），单击“OK”按钮，如图 2-25 所示。

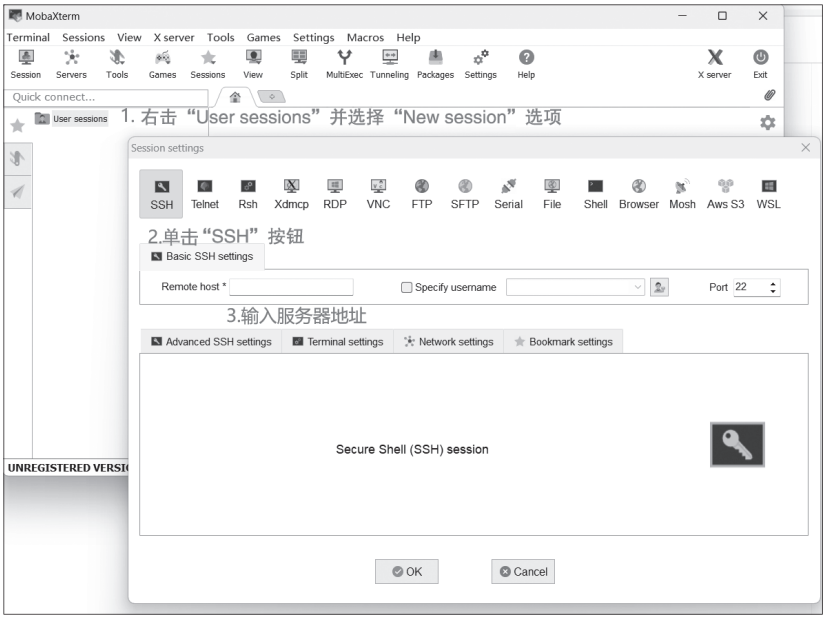


图 2-25　账号登录

（5）双击创建好的 session，输入密码即可登录服务器，并进行业务操作，如图 2-26 所示。

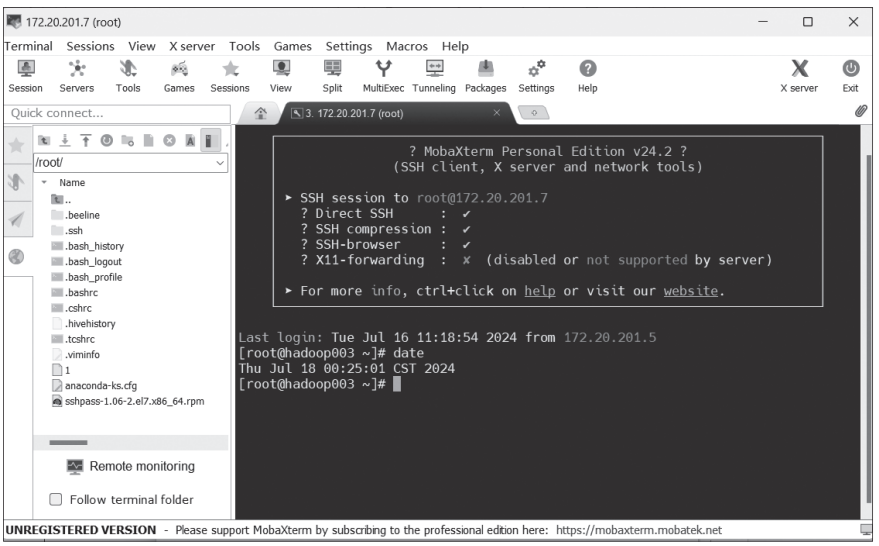


图 2-26　服务器登录

2.3.3　安装 JDK

登录 Oracle 官网，下载安装包。下载链接为：https://www.oracle.com/java/technologies/javase/javase8u211-later-arcHIVE-downloads.html。

（1）下载 JDK 1.8 安装包，将 JDK 安装包上传到 hadoop001 服务器的 /opt/software/package 目录下。

（2）解压 JDK 安装包到 /opt/software/bigdata/jdk 目录下。

```
tar -xvf jdk-8u381-linux-x64.tar.gz -C /opt/software/bigdata/jdk/
```

（3）配置环境变量，执行“vim /etc/profile”命令，将以下配置放到文件末尾。使用同样的方法在节点 hadoop002、hadoop003 上配置环境变量。

```
# JAVA_HOME
JAVA_HOME=/opt/software/bigdata/jdk/jdk1.8.0_381
JRE_HOME=${JAVA_HOME}/jre
CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
PATH=${JAVA_HOME}/bin:$PATH
```

（4）按“Esc”键，执行“:wq”命令，保存并退出。

（5）重新加载环境变量。

```
source /etc/profile
```

（6）验证 Java 是否安装成功。

```
java -version
```

如图 2-27 所示，出现 Java 版本号，则表示安装成功。

图 2-27　确认 Java 安装成功

（7）查看 jps 进程脚本。

```
#!/bin/bash
source/etc/profile
# 获取控制台指令
cmd=$*
# 判断指令是否为空
if [ ! -n "$cmd" ]
then
        echo "command can not be null !"
```

```
            exit
    fi
    # 获取当前登录用户
    user='whoami'
    # 在从机执行指令，这里需要根据具体的集群情况配置，host 与具体主机名一致，同上
    for host in hadoop001 hadoop002 hadoop003
    do
            echo "================current host is $host==
==============="
            echo "--> excute command \"$cmd\""
            ssh $user@$host source "/etc/profile;${cmd}"
    done
```

演示脚本功能，执行结果如图 2-28 所示。因为笔者连接的服务器已经安装了部分软件，所以能够通过脚本查看每个服务器上已经运行的 Java 程序。

```
[root@hadoop001 ~]# sh xcall.sh jps
================current host is hadoop001================
--> excute command "jps"
603560 NameNode
1011341 StandaloneSessionClusterEntrypoint
1011691 TaskManagerRunner
102532 RunJar
102660 RunJar
915361 Master
603748 DataNode
618052 ResourceManager
618724 JobHistoryServer
996898 HRegionServer
996129 HMaster
588935 QuorumPeerMain
1168 Jps
915548 Worker
618235 NodeManager
604286 DFSZKFailoverController
604054 JournalNode
================current host is hadoop002================
--> excute command "jps"
310213 DataNode
```

图 2-28　jps 脚本运行结果

2.4　知识拓展——从 Hadoop 核心到新兴组件的机遇与困境

除了官方认可的 Hadoop 生态圈组件之外，还有很多优秀的组件，这些组件的应用也非常广泛。例如，基于 Hive 查询优化的 Presto、Impala、Kylin 等。

下面简单介绍其中比较重要的成员。

2.4.1　基于 Hive 查询优化的组件

（1）Presto：开源分布式 SQL 查询引擎，适用于交互式分析查询，数据量支持 GB

到 PB 级。Presto 可以处理多数据源，是一款基于内存计算的 MPP 架构查询引擎。

（2）Kudu：与 HBase 类似的列存储分布式数据库，能够提供快速更新和删除数据的功能，是一款既支持随机读写、又支持 OLAP 分析的大数据存储引擎。

（3）Impala：基于 MPP 架构的高效的快速查询引擎，基于 Hive 并使用内存进行计算，兼具 ETL 功能，具有实时、批处理、多并发等优点。

（4）Kylin：开源分布式分析型数据仓库，提供 Hadoop/Spark 之上的 SQL 查询接口及多维分析（OLAP）能力，支持超大规模数据的压秒级查询。

（5）Flink：一款高吞吐量、低延迟的针对流数据和批数据的分布式实时处理引擎，是实时处理领域的新星。

（6）Hudi：由 Uber 开发并开源的数据库解决方案，Hudi（Hadoop updates and incrementals）支持 HDFS 数据的修改和增量更新操作。

如今，Hadoop 已经演化成了一个生态系统，系统内的组件千差万别，其中，经久不衰的当属 HDFS 和 Hive 两大组件，昙花一现的包括 HBase、MapReduce、Presto 等，风华正茂的当属 Spark 和 Flink。

大数据成功的核心原因是开源，但它存在的最大的问题也是开源。很多组件虽然依靠开源可以快速成熟，但是一旦成熟，就会出现生态紊乱和版本割裂的情况，其中最典型的就是 Hive。

Hive 1.x 之前的版本功能不完善，1.x 版和 2.x 版算是逐步优化到基本可用了，到了 3.x 版又出现了各种问题，并且大部分云平台的 Hive 版本都停留在 2.x 版，新版本推广乏力。另外，Hive 的计算引擎也饱受争议，Hive 支持的计算引擎主要有 MapReduce、Tez、Spark、Presto。十多年来 MapReduce 的计算速度并没有提升；Tez 虽然计算速度快，但是安装需要定制化编译和部署；Spark 的计算速度最快，但是对 JDBC 支持不友好；Presto 计算速度快且支持 JDBC，但是语法又和 Hive 不一致。

2.4.2 基于 Hadoop 的大数据平台

总的来说，基于 Hadoop 的大数据平台具有以下特点。

（1）扩容能力强：能够存储和处理 PB 级的数据。Hadoop 生态基本采用 HDFS 作为存储组件，吞吐量高、稳定可靠。

（2）成本低：可以利用廉价、通用的机器组成集群分发、处理数据。集群数量总计可达数千个节点甚至上万个节点。

（3）高效率：通过分发数据，Hadoop 可以在数据所在节点上并行处理，处理速度非常快。

（4）可靠性强：Hadoop 能自动维护数据的多份备份，并在任务失败后能自动重新部署计算任务。

2.4.3 Hadoop 生态

Hadoop 生态存在以下缺点。

（1）Hadoop 采用文件存储系统，所以读写时效性较差。

（2）Hadoop 生态系统日趋复杂，组件之间的兼容性差，安装和维护比较困难。

（3）Hadoop 各个组件功能相对单一，优点很明显，缺点也很明显。

（4）云生态对 Hadoop 的冲击十分明显，云厂商定制化组件导致版本分歧进一步扩大，无法形成合力。

2.5 任务自测

1. 选择题

（1）（单选题）Hadoop 是一个能够对大量数据进行分布式处理的软件框架，能够处理 PB 级数据。（　　）

A. 正确　　B. 错误

（2）（多选题）Hadoop 生态系统的优势包含（　　）。

A. 高扩展性　　B. 低成本

C. 开源工具成熟　　D. 属于大型关系数据库系统

（3）（多选题）（　　）属于 Hadoop 生态系统的开源工具。

A. Hive　　B. HBase　　C. MySQL　　D. ZooKeeper

（4）（多选题）Hadoop 的核心组成部分包含（　　）。

A. HDFS 存储系统　　B. Hive 数据仓库

C. MapReduce 运算框架　　D. HBase 分布式数据库

2. 简答题

Hadoop 是什么？其核心由两部分组成，分别是什么？