项目 1

大数据及 Hadoop 概述

导读

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，用户可以利用 Hadoop 生态体系开发和处理海量数据。Hadoop 具有可靠及高效的处理性能，因此逐渐成为处理大数据的优选方案。本项目将深入介绍大数据（Big Data）以及 Hadoop 的相关概念，为后面知识的学习建立概念体系。

学习目标

（1）了解大数据的基本概念及特征；
（2）了解大数据处理技术；
（3）熟悉大数据的处理流程；
（4）了解 Hadoop 的架构及生态系统。

职业素养目标

（1）增强文化自信和民族自豪感，引入国家和个人层面的思考，树立正确的价值观；
（2）增强创新意识，鼓励学生在学习 Hadoop 的过程中，不断探索新的应用场景；
（3）培养团队合作精神，培养与他人合作、共同解决问题的能力；
（4）提升职业道德意识，在使用 Hadoop 开发分布式系统时，注重数据的安全性和隐私保护，遵守相关法律法规和行业标准；
（5）培养社会责任感，通过 Hadoop 的学习和实践，引导学生关注社会热点问题，利用所学知识为社会发展做出贡献。

任务 1.1 大数据概述

■ 任务描述

通过学习本任务介绍的大数据基本概念、大数据处理技术以及大数据处理流程，读者能够对这些知识有一定了解。

知识学习

大数据概述

1. 大数据的概念和特征

随着近年来计算机技术和互联网的发展，“大数据”一词被越来越频繁地提及，大数据的快速发展也在时刻影响着我们的生活。例如，在医疗方面，大数据能够帮助医生预测疾病；在电商方面，大数据能够向顾客个性化地推荐商品；在交通方面，大数据会帮助人们选择最佳出行方案。

在高速发展的信息时代，新一轮科技革命正在加速推进，技术创新日益成为重塑经济发展模式和促进经济增长的重要驱动力，而“大数据”无疑是核心驱动力。

大数据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，是需要新的处理模式才能获得更强的决策力、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。大数据关注海量数据的存储和分析计算问题。按由小到大的排列，数据的存储单位为：bit、Byte、KB、MB、GB、TB、PB、EB、ZB 等。大数据具有以下特征。

1）Volume（大量）

截至目前，全人类说过的话的数据量大约是 5EB。当前，典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB 量级。

2）Velocity（高速）

高速是大数据区别于传统数据处理技术的最显著特征之一。根据互联网数字中心（Internet Data Center，IDC）的“数字宇宙”的报告，预计到 2025 年，全球数据使用量将达到 175ZB。在如此海量的数据面前，处理数据的效率将决定企业的命运。

3）Variety（多样性）

大数据类型的多样性体现在它分为结构化数据、半结构化数据和非结构化数据上。相比于以往便于存储的以数据库 / 文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等数据，这些数据对数据的处理能力提出了更高要求。

4）Value（低价值密度）

价值密度的高低与数据总量的大小成反比。例如，对于一天的车道数据，我们只关心车流高峰时段的数据，因此如何快速对有价值的数据进行“提纯”，成为目前大数据背景

下待解决的难题。

因此，大数据是一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样化的数据类型和价值密度低四大特征。大数据包括结构化、半结构化和非结构化数据，其中非结构化数据所占的比重越来越大。

具体来说，电商网站的用户浏览行为记录、购买行为记录，社交网站的用户行为数据记录、用户关系数据，通信行业的用户通信行为记录、上网行为记录，App 应用的用户行为数据，交通部门的海量探测数据、路况监控数据，政府部门的民生数据、舆情数据等，由于用户基数大，形成的数据量动辄日增数百 TB 甚至 PB 级别，这些都是真实、具体的大数据。

2. 大数据处理技术

处理数据需要技术，而在处理规模不同的数据集时，就算处理需求一致，但由于存储难度和计算难度不同，使用的技术也必然不同。在进行大规模数据处理时，基本上需要解决以下两个核心问题。

1）数据存储

由于大数据动辄数百 TB，甚至达到 PB 级别，无法用一个单机文件系统或者一个单机数据库进行存储。因此，在大数据技术体系中，一般采用分布式存储：将数据（文件）分散地存储到一个集群上的 N 台机器中。

2）数据运算

首先来了解什么叫运算。例如，某大型电商网站有大量的用户浏览行为记录，需要从这些记录日志中分析得到以下信息。

（1）最热门的 N 个商品。

（2）用户浏览网站的平均深度。

（3）用户浏览商品时的路径。

这些数据分析需求，最终都需要转化成运算程序来实现。而在海量数据的场景下，即使单机资源（无论是 CPU，还是内存）的配置达到极限，也无法在合理的限定时间内运算出结果，所以，在大数据技术体系下，数据运算主要通过运算资源（计算节点）的水平扩展来实现，即使用分布式集群运算系统。

3. 大数据处理流程

大数据处理流程一般分为五个步骤：数据采集、数据清洗和预处理、数据存储、数据分析和挖掘、数据可视化，大数据处理流程及常用工具如图 1-1 和图 1-2 所示。

图 1-1　大数据处理流程图

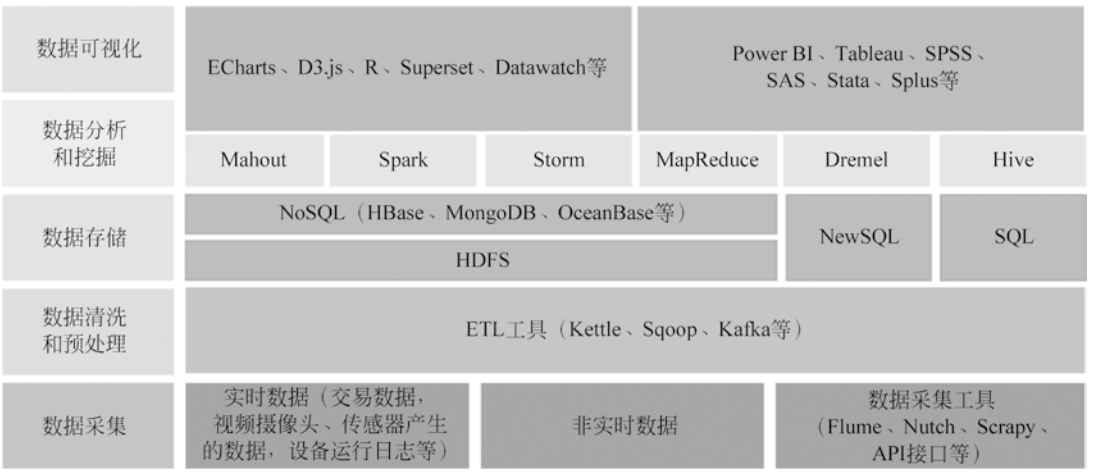

图 1-2　大数据处理常用工具

1）数据采集

数据的来源多种多样，包括移动互联网和社交网络等。这些结构化和非结构化数据是零散的，也就是存在所谓的数据孤岛。在这种情况下，这些数据的作用十分有限。此时，便可利用数据采集技术将这些数据写入数据仓库，将零散的数据整合在一起，进行分析。数据采集包括对交易数据、文件日志、各类传感器数据及设备运行日志的采集，涉及关系型数据库的接入和应用程序的接入等。常用的数据采集工具包括 Flume、Nutch、Scrapy 等。

2）数据清洗和预处理

采集的数据中包含大量重复或无用的数据。此时，需要对数据进行简单的清洗和预处理，从而将不同来源的数据整合成一致的、适合数据分析算法和工具读取的数据，如数据去重、异常处理和数据归一化等，然后将这些数据存储到大型分布式数据库或者分布式集群中。

一般采用 ETL（Extract Transform Load，即抽取—转换—加载）工具将分布式、异构数据源中的数据（如关系数据、平面数据以及其他非结构化数据等）抽取到临时文件或数据库中。

3）数据存储

处理过后的数据保存在分布式存储系统中，如分布式文件系统 HDFS、分布式数据库 HBase 中。

4）数据分析和挖掘

数据分析需要用到 SPSS 等工具。数据挖掘主要指针对现有数据使用各种算法进行计算，达到预测以及实现一些高级别数据分析的需求。

5）数据可视化

可视化分析能够直观地呈现大数据分析和挖掘的结果，常用的工具包括 ECharts、D3.js 等。

综上所述，大数据处理流程是一个从数据采集到应用的过程，每个环节都需要精细设计和执行，以确保数据处理和分析的效率和准确性。

任务 1.2 Hadoop 概述

■ 任务描述

通过学习本任务讲解的 Hadoop 基本概念及技术生态，读者能够对 Hadoop 的作用有一定了解。

知识学习

Hadoop 概述

1. Hadoop 简介

Hadoop 最早由雅虎公司的技术团队根据谷歌公司公开论文中的思想，用 Java 语言开发，现在则隶属于 Apache 基金会。

Hadoop 以分布式文件系统 HDFS（Hadoop Distributed File System）和分布式计算框架 MapReduce 为核心，为用户提供了底层细节透明的分布式基础设施。HDFS 具备高容错性、高伸缩性等优点，允许用户将 Hadoop 部署在廉价的硬件上，构建分布式文件存储系统。

MapReduce 分布式计算框架则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、分布式应用程序，充分利用大规模的计算资源，解决传统高性能单机无法解决的大数据处理问题。

总之，Hadoop 是一种海量数据的处理工具，并已经被各行各业广泛应用于以下场景。

（1）大数据海量存储：分布式文件系统 HDFS 以及分布式数据库 HBase。

（2）日志处理：Hadoop 可处理大规模离线日志。

（3）海量计算：分布式并行计算 MapReduce。

（4）ETL：数据抽取到 Oracle、MySQL、DB2、MongDB 及主流数据库中。

（5）数据分析：使用 HBase 的扩展性应对大量读写操作。

（6）机器学习：如 Apache Mahout 项目（其常见应用领域：协作筛选、集群、归类）。

（7）搜索引擎：基于 Hadoop + Lucene 技术开发搜索引擎应用。

（8）数据挖掘：适用于用户行为特征建模、个性化广告推荐。

2. Hadoop 技术生态系统

自从 Hadoop 成为 Apache 基金会的顶级项目后，经过长时间的发展，围绕 Hadoop 出现了大量开源扩展技术框架，从而形成了一个庞大的 Hadoop 技术生态体系。Hadoop 技术生态系统如图 1-3 所示。

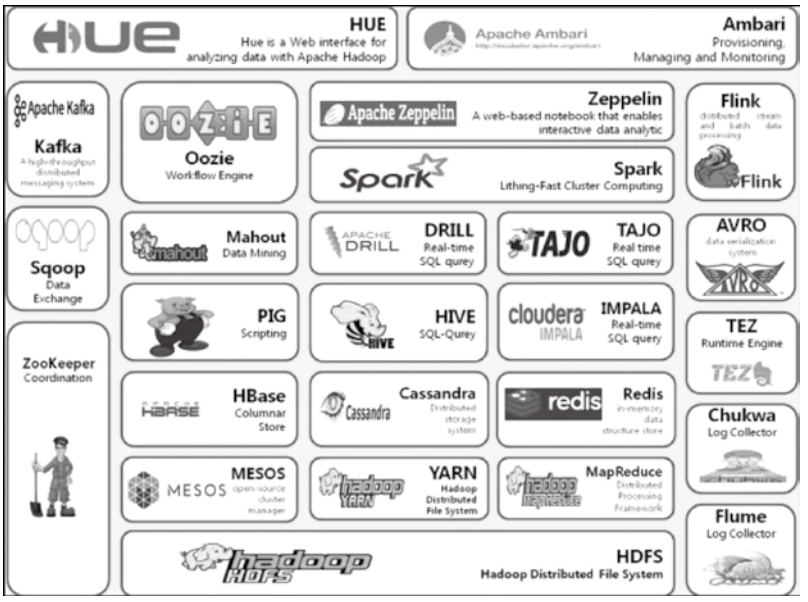

图 1-3　Hadoop 技术生态系统

1）Hadoop 技术生态系统的核心框架组件

（1）HDFS/MapReduce：这两个组件是 Hadoop 的两大核心组件，HDFS 提供分布式文件系统，MapReduce 则提供分布式运算程序编程框架。

（2）Hive：直接基于 MapReduce 开发数据处理和分析的分布式运算程序的技术门槛高、开发效率低，而 Hive 则提供了一个 SQL 脚本作为 MapReduce 运算程序之间的转换桥梁，用户可以基于 Hive 编写类 SQL 脚本，从而快速实现各类数据统计分析功能的开发需求。

（3）HBase：HDFS 是只能追加数据的文件系统，不支持数据的修改，而 HBase 的出现解决了该问题。HBase 运行在 HDFS 之上，是一个分布式、随机访问、面向列的数据库系统，它允许应用程序开发人员直接读写 HDFS 数据。只是，HBase 并不支持 SQL 语句，属于 NoSQL 数据库的一种。然而，HBase 提供了基于命令行的界面以及丰富的 API 函数来操作数据。

2）Hadoop 技术生态系统的外围框架组件

Hadoop 技术生态系统除了核心组件，还包含了非常多外围框架组件，有如下常见的框架组件。

（1）ZooKeeper：Hadoop 技术生态系统中一个非常基础的服务框架，是各分布式框架公用的一个分布式协调服务系统。它通过为各类分布式框架提供状态数据的记录和监听功能，使各类分布式系统的开发变得更加便捷。

（2）Mahout：一个开源的机器学习库，它能使 Hadoop 用户高效地进行诸如数据分析、数据挖掘以及集群等一系列操作。它提供的算法经过性能优化能够在 HDFS 文件系统上高效地运行 MapReduce 框架，对大数据集特别高效。

（3）Ambari：提供一套基于网页的界面来管理和监控 Hadoop 集群，让 Hadoop 集群的部署和运维变得更加简单。它提供了一系列功能，如安装向导、系统警告、集群管理、

任务性能优化等。

（4）Kafka：一个分布式、高吞吐量、支持多分区和多副本的基于 ZooKeeper 的分布式消息发布订阅系统。Kafka 的设计初衷是构建一个用来处理海量日志、用户行为和网站运营统计等的数据处理框架，目前与 Spark 等分布式实时处理组件结合使用，用于实时流式数据分析。

（5）Sqoop：用来在各类传统的关系型数据库（如 MySQL、Oracle 等）和 Hadoop 生态体系中的各类分布式存储系统（如 HDFS、Hive、HBase 等）之间进行数据迁移，从而让开发人员快速地将业务系统数据库中的数据加载到 Hadoop 中，通过综合其他日志数据进行分析。此外，Sqoop 还能方便地将分析结果导出到关系型数据库中，以便进行查询分析和数据可视化。

（6）Flume：用来进行日志的采集、汇聚，它能从各类数据源中读取数据，并将这些数据汇聚到 HDFS、HBase、Hive 等各种类型的大型存储系统中。并且，在使用 Flume 时，用户几乎不用进行任何编程，只需要在 Flume 的配置文件中对数据源和汇聚存储系统的属性进行配置，即可快速搭建一个大型分布式数据采集系统。

（7）Spark：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。

（8）Flink：当前最流行的开源大数据内存计算框架之一，常用于实时计算场景。

（9）Oozie：一个管理 Hadoop 作业（job）的工作流程调度管理系统。

Hadoop 技术生态系统中各组件的架构，如图 1-4 所示。

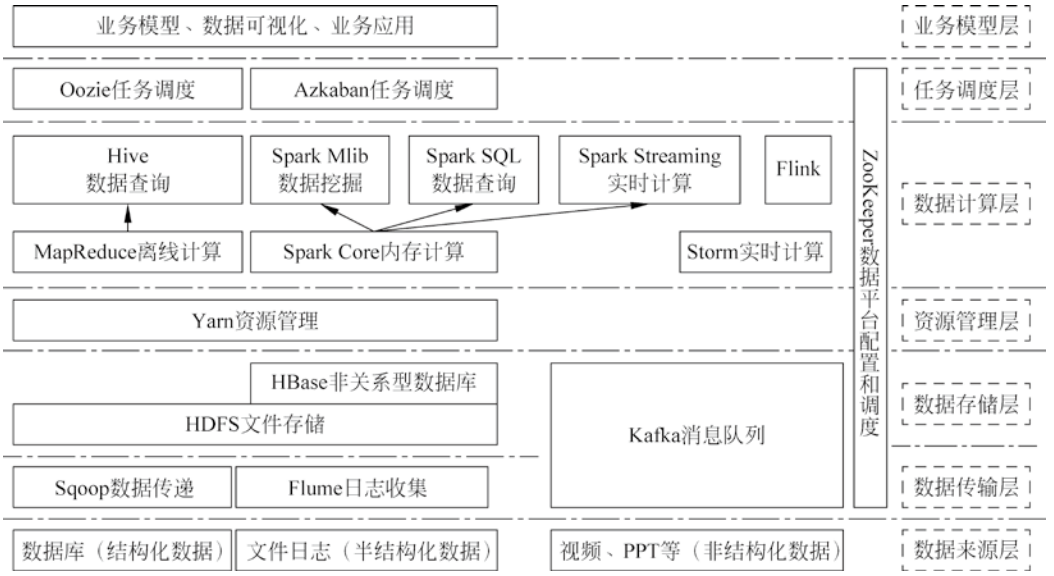

图 1-4　Hadoop 技术生态系统中各组件的架构

3）典型的大数据处理系统架构

Hadoop 技术生态系统中组件众多，上文介绍的组件只是其中的一小部分，不过大部分组件的使用场景十分有限，大部分情况下用不到。典型的大数据处理系统架构如图 1-5 所示。

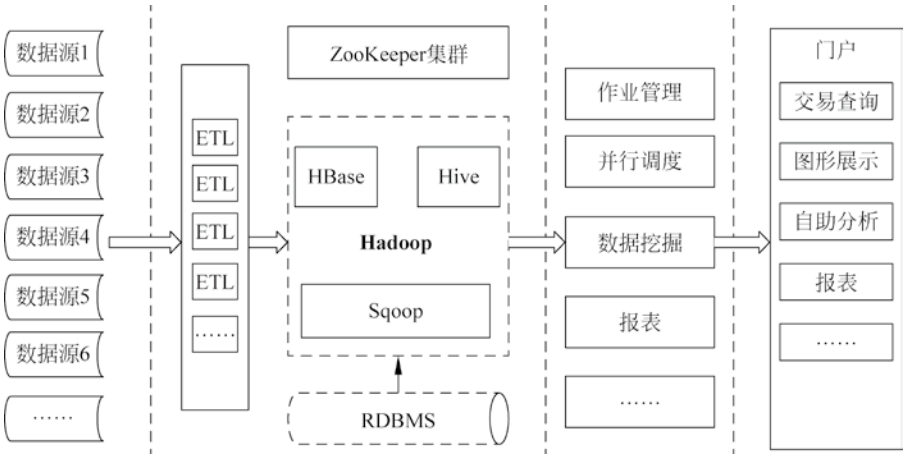

图 1-5　典型的大数据处理系统架构

课后练习

一、单选题

1. Hadoop 最初是由（　　）公司创建的。
   A. 雅虎　　B. 谷歌　　C. Apache　　D. 微软
2. Hadoop 是一个能够对大量数据进行（　　）处理的软件框架。
   A. 分布式　　B. 集中式　　C. 串行　　D. 并行
3. 在 Hadoop 中，负责数据存储的是（　　）组件。
   A. MapReduce　　B. HBase　　C. HDFS　　D. Yarn
4. Hadoop 可以运行在哪些操作系统上？（　　）
   A. 只能在 Windows 上运行　　B. 只能在 Linux 上运行
   C. 可以在多种操作系统上运行　　D. 只能在 macOS 上运行
5. 以下（　　）不是 Hadoop 的特点。
   A. 高可靠性　　B. 高可扩展性　　C. 高效性　　D. 实时性
6. Hadoop 中的 MapReduce 主要用于（　　）。
   A. 数据存储　　B. 数据处理　　C. 数据传输　　D. 数据展示
7. 大数据通常指的是（　　）特征的数据集。
   A. 数据量小，价值低　　B. 数据量大，类型单一
   C. 数据量小，类型多样　　D. 数据量大，类型多样
8. 以下（　　）组件不是 Hadoop 技术生态系统中的一部分。
   A. HBase　　B. Hive　　C. MySQL　　D. Pig
9. 在大数据处理过程中，（　　）是数据预处理的一个重要环节。
   A. 数据采集　　B. 数据存储　　C. 数据清洗　　D. 数据传输
10. Hadoop 的（　　）组件可以实现数据的实时查询和分析。
    A. HBase　　B. Hive　　C. HBase + Phoenix　　D. MapReduce

二、多选题

1. Hadoop 的主要组成部分包括（　　）。

A. HDFS　　B. MapReduce　　C. Yarn　　D. HBase

2. 关于 HDFS，以下说法是正确的是（　　）。

A. HDFS 是 Hadoop Distributed File System 的缩写

B. HDFS 是一个高容错性系统

C. HDFS 适合存储大量的小文件

D. HDFS 中的数据默认存储 3 份

3. Hadoop 技术生态系统可以应用在（　　）场景。

A. 日志分析　　B. 数据仓库　　C. 推荐系统　　D. 实时流处理

4. 大数据分析的主要方法包括（　　）。

A. 批处理　　B. 流处理　　C. 图计算　　D. 机器学习

5. 以下（　　）技术可以用于 Hadoop 数据的实时处理。

A. HBase　　B. Hive　　C. Spark Streaming　　D. Flink

项目 2

Hadoop 分布式集群安装及部署

导读

学习 Hadoop，需要一个可运行的 Hadoop 集群，而搭建一个 Hadoop 集群，则需要准备多台 Linux 服务器。如果购买真正的计算机来安装 Linux 系统作为服务器，在学习阶段显然成本太高，好在这并不是唯一的解决方案。我们可以通过虚拟机技术，快速获得多台虚拟 Linux 机器，并在这些虚拟机器上部署 Hadoop 集群。

学习目标

（1）了解 Linux 操作系统；
（2）掌握 Linux 操作系统的安装及配置方法；
（3）掌握 Hadoop 伪分布式系统的配置过程；
（4）掌握 Hadoop 完全分布式系统的配置过程。

技能目标

（1）能够独立搭建和管理 Hadoop 集群；
（2）能够处理 Hadoop 的常见问题和故障；
（3）能够对 Hadoop 集群进行性能调优。

职业素养目标

（1）增强文化自信和民族自豪感，引导树立正确的价值观；
（2）增强创新意识，鼓励学生不断探索新的 Hadoop 应用场景和技术创新点；
（3）培养团队合作精神，在 Hadoop 的学习和实践中，强调团队协作的重要性，培养与他人合作、共同解决问题的能力；
（4）提升职业道德意识，在使用 Hadoop 集群时遵守相关法律法规和行业标准。