任务1-1 了解Spark及其生态系统

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | 深入理解Spark的基本架构、RDD（弹性分布式数据集）、DAG（有向无环图）等核心概念，掌握Spark的工作原理和数据处理流程。 |
| 具体要求 | 1.在不同环境下搭建和配置Spark；  2.了解Spark的核心概念和基本架构。 |
| 二、任务目标 | |
| 知识目标 | * 掌握 Spark 的核心概念及在大数据处理中的优势； * 熟悉 Spark 生态系统各组件的功能与应用场景。 |
| 能力目标 | * 能够运用 Spark 进行高效的数据处理和分析； * 具备整合 Spark 生态系统组件解决实际问题的能力。 |
| 三、知识储备 | |
| 1.Spark简介。  Apache Spark是一个开源的分布式大数据处理框架，自2009年在加州大学伯克利分校的AMPLab诞生以来，已经发展成为大数据生态系统中的核心组件之一。如果用一句话定义Apache Spark，那么Apache Spark是用于大规模数据（large-scala data）处理的统一（unified）分析引擎。Spark的设计初衷是为了解决Hadoop MapReduce在迭代算法和交互式数据分析方面的局限性，通过提供更高效的内存计算模型来加速大规模数据处理。    2. Spark与Hadoop的对比  Apache Spark和Apache Hadoop作为大数据处理领域的两大巨头，各自拥有独特的优势和应用场景。Hadoop作为较早出现的大数据框架，以其强大的分布式文件系统HDFS和MapReduce编程模型为核心，为大规模数据存储和批处理奠定了基础。它的设计理念是将复杂的任务分解为可以在商用硬件集群上并行执行的简单任务，这使得Hadoop特别适合处理超大规模的数据集，尤其是当数据量远超过集群总内存容量时。Hadoop的MapReduce模型虽然概念简单，但在处理需要多次迭代的复杂算法时效率较低，因为每次迭代结果都需要写入磁盘。 | |

|  |
| --- |
| 操作指南 |
| 一、制定计划 |
| |  |  |  |  | | --- | --- | --- | --- | | **序号** | **作业项目** | **序号** | **作业项目** | | 1 |  | 5 |  | | 2 |  | 6 |  | | 3 |  | 7 |  | | 4 |  | 8 |  | | 计划审核 | 审核意见： | | | |
| 二、实施方案 |
| 1. 进入Spark官网  打开浏览器，输入Spark官网的网址，进入Spark官网，官网如图所示：https://spark.apache.org/  IMG_256  2. 进入下载界面  在官网首页找到导航栏中的“Download”按钮，进入下载界面，下载页面如图所示：  IMG_256  3.下载Spark压缩包  可以选择下载最新的版本，也可以选择下载往期稳定的版本，本案例主要使用版本3.2进行演示，Spark压缩包的下载页面如图所示：  IMG_256  4.创建文件夹  进入虚拟机，创建文件夹：  mkdirs /export/server  mkdirs /export/data  mkdirs /export/software  5.解压Spark压缩包  将Spark压缩包上传至虚拟机，跳转到上传的文件夹，并且解压Spark压缩包：tar -zxvf spark-3.2.0-bin-hadoop3.2.tgz -C /export/server/  至此，Spark已经完成解压，可直接使用。 |
| 三、实施记录 |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **序号** | **关键步骤记录** |  | **序号** | **解决问题记录** | | 1 |  |  | 1 |  | | 2 |  |  | 2 |  | | 3 |  |  | 3 |  | | 4 |  |  | 4 |  | | 5 |  |  | 5 |  | | 6 |  |  | 6 |  | | 7 |  |  | 7 |  | | 8 |  |  | 8 |  | | 9 |  |  | 9 |  | |
| 四、任务拓展 |
| 官网的“Get Started”页面中提供了多种Spark下载搭建的方式，请尝试使用不同的方式完成Spark搭建。 |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元一 Spark入门 | | | | | | |
| 任务名称 | | 任务1-1 了解Spark及其生态系统 | | | | | | |
| 班级 |  | | 学号 | |  | 姓名 | |  |
| 组名 |  | | 学时 | |  | 日期 | |  |
| 组员  分工 |  | | |  | | |  | |
|  | | |  | | |  | |
| 二、任务检查评价单 | | | | | | | | |
| |  |  |  | | --- | --- | --- | | **评价内容** | **评价标准** | **得分** | | 知识运用  （20分） | 掌握相关理论知识，理解本次任务要求，制定详细计划，计划条理清晰，逻辑正确（20分） |  | | 理解相关理论知识，能根据本次任务要求、制定合理计划（15分） | | 了解相关理论知识，有制定计划（10分） | | 无制定计划（0分） | | 专业技能  （40分） | 结果验证全部满足。（40分） |  | | 结果验证只有一个功能不能实现，其它功能全部实现（30分） | | 结果验证只有一个功能实现，其它功能全部没有实现（20分） | | 结果验证功能均未实现（0分） | | 核心素养  （20分） | 具有良好的自主学习能力、分析解决问题的能力、整个任务过程中有指导他人（20分） |  | | 具有较好的学习能力和分析解决问题的能力，任务过程中无指导他人（15分） | | 能够主动学习并收集信息，有请教他人进行解决问题的能力（10分） | | 不主动学习（0分） | | 课堂纪律  （20分） | 设备无损坏、设备摆放整齐、工位区域内保持整洁、无干扰课堂秩序（20分） |  | | 设备无损坏、无干扰课堂秩序（15分） | | 无干扰课堂秩序（10分） | | 干扰课堂秩序（0分） | | 得分（满分100） | |  | | 综合评价 | ☆☆☆☆☆ | | | | | | | | | | |