



娄底职业技术学院

Loudi Vocational & Technical College

大数据技术专业
人才培养方案评价资料

技能考核标准

娄底职业技术学院

二〇二三年八月

第一部分：大数据技术技能考核标准

一、专业代码及专业名称

专业名称：大数据技术

专业代码：510205

二、考核目标

本专业技能考核包括程序设计、数据库设计、网络爬虫、Hadoop 平台与组件、数据分析，数据可视化 6 个技能考核模块。用于检测学生的程序设计与开发能力、大数据采集能力、数据存储与处理分析能力、大数据平台部署与运维能力、大数据可视化开发能力以及从事大数据开发工作的程序编写规范、技术文档编写、交流与沟通、法律法规等职业素养。

三、考核内容

表 1：大数据应用与技术技能考核内容

| | |
|-----------------|----------------------|
| 模块一： 基本专业技能 | 项目 1：常见算法设计与实现 |
| | 项目 2：MySQL 数据库操作与查询 |
| 模块二： 岗位核心技能 | 项目 3：网络爬虫与分析 |
| | 项目 4：Hadoop 集群部署与使用 |
| | 项目 5：数据仓库 Hive 部署与使用 |
| | 项目 6：Flink 的部署与使用 |
| | 项目 7：Spark 的部署与使用 |
| 模块三： 跨岗位综合技能 | 项目 8：Python 数据可视化 |

模块 1 专业基本技能

本模块以企、事业单位应用项目为背景，完成项目开发平台的配置与使用、项目模型的设计与建立、程序代码的编写与运行等工作内容，基本涵盖大数据开发工程师、采集工程师、大数据存储与分析工程师等岗位从事项目设计与开发工作所需的基本技能。

项目 1 常见算法设计与实现

1. 技能要求

- (1)能熟练使用主流的软件开发平台，并进行相关参数的配置；
- (2)能使用平台进行项目的创建、开发、编译、运行及调试；
- (3)能使用面向对象思想对信息化项目进行建模与设计；
- (4)能将编程任务以流程图的形式描述出来；

(5)能使用数据类型、变量、常量、运算符、表达式、函数，并结合顺序、分支、循环三种控制结构实现项目的业务逻辑单元；

(6)能使用封装、继承、多态、类、接口、对象等语言机制，进行面向对象程序的编写，实现代码的可重用性；

(7)能将数组等基本数据结构及查找、排序等基础算法应用到程序代码的编写中，实现项目性能的提升。

2. 素养要求

(1)在悉相关规范和标准，能识读相关软件技术文档；

(2)具有较强的分析问题的能力、发散思维和创新意识；

(3)具有程序员必备的良好编程习惯，程序架构合理，模块结构合理，命名规范、注释清晰，缩进良好，具备较好的可读性和可维护性；

(4)具备程序员的严谨认真、规范的工作态度和正确的价值观。

本模块以企业事业单位信息管理系统项目开发为背景，完成大数据应用信息系统中数据库开发环境的配置与使用、数据库及数据表的设计、创建与管理、数据表的约束与关系、数据库访问和数据库备份管理等工作内容。基本涵盖了大数据数据采集工程师、大数据平台运维、实施工程师岗位从事应用系统数据库的设计与开发工作所需的基本技能。

项目 2 MySQL 数据库操作与查询

1. 技能要求

(1) 能使用主流的数据库管理系统，在常用操作系统上进行相关参数的配置，完成数据库管理系统的安装；

(2) 能使用数据库相关工具，完成数据库备份与日常维护；

(3) 能根据 E-R 图，系统数据库模型完成数据库和表的创建；

(4) 能使用 SQL 语句完成数据库，表的创建修改和删除操作；

(5) 能使用 SQL 语句实现表的主、外键等约束条件添加、修改与删除操作；

(6) 能使用 SQL 语句实现数据表记录的插入、修改和删除操作；

(7) 能使用 SQL 语句实现联合查询、子查询并能实现对查询结果的筛选、排序、统计等操作；

(8) 能使用 SQL 语句实现视图、存储过程的创建、修改或删除操作。

2. 素养要求

(1) 在悉相关规范和标准，具备独立分析解决问题能力，能识读相关软件技术文档；

(2) 在应用系统数据库设计过程中，具有数据库管理员、程序员必备的数据库操作和管理习惯；

(3)数据表结构设计合理，SQL 语句执行效率高，表名、字段名命名规范具较好的可读性和可维护性；

(4)具备程序员的严谨认真、规范的工作态度和正确的价值观。

模块 2 岗位核心技能

项目 3 网络爬虫与分析

1. 技能要求

(1)能完成爬虫的集成环境搭建；

(2)能够掌握 requests、urllib3 的使用；

(3)能够完成 cookie，session 等查询与设置；

(4)能够使用 BeautifulSoup4、Xpath、lxml 完成网页解析；

(5)能够使用 python 文件操作保存爬取数据到 csv 中；

(6)能够使用 pymysql 文件操作保存爬取数据到 MySQL 数据库中。

2. 素养要求

(1)能够完成真实业务逻辑向代码转换；

(2)具有良好的编程习惯、较强的逻辑思维能力及综合运用知识的能力；

(3)具备程序员的严谨认真、规范的工作态度和正确的价值观；

(4)在进行网络爬虫的过程中能注意内容的合法性（不能非法爬取隐私信息 等）、规范性。

项目 4 Hadoop 集群部署与使用

1. 技能要求

(1)能够完成 Hadoop 伪分布式集群、分布式集群的安装与部署；

(2)熟悉 Hadoop 三大组件；

(3)能够使用 HDFS 完成文件读写操作；

-
- (4) 能够使用 MapReduce 完成对文件数据的常规操作；
 - (5) 能够完成 Yarn 调度器的配置与使用。

2. 素养要求

- (1) 能够独立解决技术问题，规范操作对应命令；
- (2) 熟悉相关规范和标准，能识读相关软件技术文档；
- (3) 具备程序员的严谨认真、规范的工作态度和正确的价值观。

项目 5 数据仓库 Hive 部署与使用

1. 技能要求

- (1) 能够完成 Hive 数据仓库的安装与部署；
- (2) 能够使用命令行和 JDBC 方式使用 Hive；
- (3) 熟悉 Hive 中数据类型、表类型；
- (4) 能够完成对 Hive 表数据的查询、添加、删除操作；
- (5) 能够灵活使用 Hive 中的函数进行分组、去重、排序、行列转换等操作。

2. 素养要求

- (1) 具有较强的分析与解决问题的能力，能够独立解决技术问题；
- (2) 具有良好的编程习惯、较强的逻辑思维能力及综合运用知识的能力；
- (3) 具备程序员的严谨认真、规范的工作态度和正确的价值观。

项目 6 Flink 的部署与使用

1. 技能要求

- (1) 熟悉 Flink Streaming 和 Flink Batch 程序；
- (2) 熟悉 Flink 本地、集群模式的安装与部署使用；
- (3) 熟悉 Flink 常用 API:DataStream, DataSet, Table；

(4) 熟悉 Flink 中 Broadcast, Accumulator 的使用;

(5) 熟悉 Flink State 的管理与恢复;

(6) 熟悉 Flink 中 Window, Time 的使用。

2. 素养要求

(1) 具有较强的分析与解决问题的能力;

(2) 掌握数理统计分析等分析方法, 并能初步进行市场预测;

(3) 具有良好的编程习惯、较强的逻辑思维能力及综合运用知识的能力;

(4) 具备程序员的严谨认真、规范的工作态度和正确的价值观。

项目 7 Spark 的部署与使用

1. 技能要求

(1) 能够完成 Spark 集群安装部署;

(2) 能够完成 Spark 开发环境的安装与配置部署;

(3) 熟悉 Spark 任务的三种提交方式;

(4) 熟悉 Spark 中 Transformation 和 Action;

(5) 熟悉 RDD 持久化开发;

(6) 熟悉共享变量 Broadcast Variable 和 Accumulator。

2. 素养要求

(1) 具有较强的分析与解决问题的能力;

(2) 具备良好的自学能力, 关注互联网新知识、新技术, 有一定创新意识;

(3) 具有良好的编程习惯、较强的逻辑思维能力及综合运用知识的能力;

(4) 具备程序员的严谨认真、规范的工作态度和正确的价值观。

模块3 跨岗位综合技能

项目8 Python 数据可视化

1. 技能要求

- (1) 能够使用 Pandas 完成数据的清洗与统计；
- (2) 能够使用 matplotlib 完成数据的展示；
- (3) 能够使用 pyecharts 完成数据的展示；
- (4) 能够对程序进行优化处理。

2. 素养要求

- (1) 能够根据行业领域需求进行可视化设计；
- (2) 能识读相关软件技术文档并能编写逻辑清晰的分析报告；
- (3) 具有良好的编程习惯、较强的逻辑思维能力及综合运用知识的能力；
- (4) 具备程序员的严谨认真、规范的工作态度和正确的价值观。

四、评价标准

1. 评价方式:

本专业技能考核采取过程考核与结果考核相结合,技能考核与职业素养考核相结合。根据考生操作的规范性、熟练程度和用时量等因素评价过程成绩;根据设计作品、运行测试结果和提交文档质量等因素评价结果成绩。

2. 分值分配:

本专业技能考核满分为 100 分,其中专业技能占 90 分,职业素养占 10 分。

3. 技能评价要点:

根据模块中考核项目的不同,重点考核学生对该项目所必须掌握的技能和要求。虽然不同考试题目的技能侧重点有所不同,但完成任务的工作量和难易程度基本相同。各模块和项目的技能评价要点内容如表 1 所示。

表 2: 大数据技术专业技能考核评价要点

| 序号 | 类型 | 模块 | 项目 | 评价要点 |
|----|--------|------|-----------|--|
| 1 | 专业基础技能 | 程序设计 | 常见算法设计与实现 | 1. 正确配置开发环境。5 分 2. 项目的设计步骤清晰、方法科学合理。5 分 3. 正确运用各种图例画出流程图。10 分 4. 正确定义变量、常量、名称符合编码规范。10 分 5. 正确使用运算式、表达式。 |

| | | | | |
|---|--------|-------|----------------|--|
| | | | | <p>10 分</p> <p>6. 正确使用顺序、分支、循环三种控制结构。20 分</p> <p>7. 正确使用数组等基本数据结构。10 分</p> <p>8. 正确使用封装、继承、多态、类等面向对象语言机制、实现代码复用。20 分</p> <p>9. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。10 分</p> |
| | | 数据库设计 | MySql 数据库操作与查询 | <p>1. 正确使用数据库管理工具，完成用户管理与配置。5 分</p> <p>2. 正确使用 E-R 图完成数据库表结构设计。5 分</p> <p>3. 正确使用 SQL 语句完成数据库和表创建。10 分</p> <p>4. 正确使用 SQL 语句完成表约束管理。10 分</p> <p>5. 正确使用 SQL 语句完成数据增、删、改、查等操作。40 分</p> <p>6. 正确使用 SQL 完成视图、存储过程等复杂操作。20 分</p> <p>7. 遵守对应编码规范。10 分</p> |
| 2 | 岗位核心技能 | 数据爬虫 | 爬取常见网页数据 | <p>1. 正确完成爬虫环境搭建与配置。10 分</p> <p>2. 正确使用 requests、urllib3、Scrapy 完成数据爬取。30 分</p> <p>3. 正确使用 BeautifulSoup4、Xpath、lxml</p> |

| | | | |
|--|--|--------------|--|
| | | | <p>完成网页解析。30 分</p> <p>4. 正确使用文件操作保存数据到 csv 中。10 分</p> <p>5. 正确使用 pysql 文件操作保存爬取数据到 MySQL 数据库中。10 分</p> <p>6. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。。10 分</p> |
| | | Hadoop 平台与组件 | <p>Hadoop 集群部署与使用</p> <p>1. 正确完成 Hadoop 伪分布式集群、分布式集群的安装与部署。30 分</p> <p>2. 正确完成 Hadoop 对应组件。Hive, Hbase, Flume, Kafka 等的安装与配置。20 分</p> <p>3. 正确使用命令完成 HDFS 文件读写操作。10 分</p> <p>4. 正确使用 java 完成 MapReduce 开发对应数据常规操作。10 分</p> <p>5. 正确完成项目上传并测试通过。10 分</p> <p>6. 正确完成 Yarn 调度器的配置与使用。10 分</p> <p>7. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。10 分</p> |
| | | | <p>数据仓库 Hive 部署与使用</p> <p>1. 正确完成 Hive 数据仓库的安装与部署。10 分</p> <p>2. 正确使用命令行和 JDBC 方</p> |

| | | | | |
|--|--|------|--------------|--|
| | | | | <p>式使用 Hive。10 分</p> <p>3. 正确使用 Hive 中数据类型、表类型。10 分</p> <p>4. 正确使用 Hive 表数据的查询、添加、删除操作。30 分</p> <p>5. 正确使用 Hive 中的函数进行分组、去重、排序、行列转换等操作。30 分</p> <p>6. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。10 分</p> |
| | | 数据分析 | Spark 的部署与使用 | <p>1. 正确完成 Spark 集群环境安装和部署。10 分</p> <p>2. 正确使用 Spark 任务的三种提交方式。20 分</p> <p>3. 正确使用 Spark 中 Transformation 和 Action。30 分</p> <p>4. 正确使用共享变量 Broadcast Variable 和 Accumulator。30 分</p> <p>5. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。10 分</p> |
| | | | Flink 的部署与使用 | <p>1. 正确完成 Flink 本地、集群模式的安装与部署。10 分</p> <p>2. 根据要求正确使用 Flink Streaming 和 Flink Batch。10 分</p> <p>3. 正确使用 Api:DataStream、DataSet、Table 完成操作。20 分</p> |

| | | | | |
|---|----------------|-----------|---------------------|---|
| | | | | <p>4. 正确使用 Broadcast、Accumulator。20 分</p> <p>5. 正确使用 Flink State 的管理与恢复。10 分</p> <p>6. 正确使用 Flink 中的 window 和 Time。20 分</p> <p>7. 程序书写结构良好、注释清晰、可维护性好、语法正确、功能正确完备，开发过程遵循软件开发规范。10 分</p> |
| 3 | 拓展 岗位 技能 | 数据可 视化 | Python 数据可 视化 | <p>1. 正确使用 Pandas 完成数据导入与数据清洗，统计，并输出结果。30 分</p> <p>2. 正确使用 matplotlib 或者 pyecharts 完成数据图形展示，包括不限于曲线图，柱状图，饼状图，词云图。30 分</p> <p>3. 图形表示正确，并且有对应标注显示，配色合理，界面美观。10 分</p> <p>4. 图形组合合理，绘制子图方法正确 绘制三维图形方法恰当 绘制交互图表方法合理 趋势分析选择图形合理 。20 分</p> <p>5. 程序书写结构良好，注释清晰，可维护性好，语法正确，功能正确且完备，开发过程遵循软件开发规范。10 分</p> |

五、考核方式

本专业技能考核方式为现场操作考核，成绩评定采用过程考核与结果考核相结合。具体方式如下：

1、模块选择：专业基本技能 2 个模块为必选模块，岗位核心模块 3 个模块为必选模块，跨岗位综合技能为可选模块。

2、学生抽取：从大数据技术专业全日制在籍大三学生中随机抽取 10%学生参加，如抽查专业学生不足 100 人，则抽取学生 10 人。

3、模块抽取：参考学生按照规定比例随机抽取考试模块，其中，专业基本技能模块不低于 30%，岗位核心技能模块不低于 60%。考试方案可以根据实际情况对参考学生比例进行适当调整，各模块考生人数按四舍五入计算。

4、试题抽取：学生在相应模块题库中随机抽取 1 道试题考核。

5、考位抽取：学生根据对应模块与试题号到达对应考试区域进行考位随机抽选，确定考位号，并登记备案。

六、实施条件

场地要求：大数据技术实训机房

软硬件要求：根据模块的不同，在试题中均附有相应的软硬件要求

表 3：程序设计模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|--------------------------------------|-----------------------------|------------------|
| 1 | 计算机 | 双核 CPU，内存 4G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | JDK1.8 或以上，Python3.0 或以上 | Java 和 Python 开发环境 | |
| 4 | Intelij IDEA2019 或以上，Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | MSDN 或 JDK 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

表 4：数据库设计模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|-------------------------|-----------------------------|------------------|
| 1 | 计算机 | 双核 CPU，内存 4G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | MySQL5.7 或以上 Navicat | 数据库管理系统 | |

表 5：网络爬虫模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|--------------------------------------|-----------------------------|------------------|
| 1 | 计算机 | 双核 CPU，内存 8G 或以上，win10 操作系统 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | Python3.0 或以上，爬虫相关库 | 爬虫环境 | |
| 4 | Intelij IDEA2019 或以上，Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | Python 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

表 6：Hadoop 平台与组件模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|--|--|------------------------------|
| 1 | 计算机 | 四核 CPU，内存 8G 或以上，win10 操作系统 或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | VMware Workstation Pro12 及以上 Xshell, tabby, electerm 等 SSH 连接工具 | 虚拟机 SSH 工具 | 虚拟机 Hadoop 环境完备。 参考人员自选工具 |
| 4 | JDK 帮助文档 Hadoop 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

表 7. 数据分析模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|--|--|---|
| 1 | 计算机 | 四核 CPU，内存 16G 以上，win10 操作系统 或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | VMware Workstation Pro12 及以上 (安装好对应 hadoop 及 hive 环境) Xshell, tabby, electerm 等 SSH 连接工具 | 虚拟机 SSH 工具 | 多台虚拟内 hadoop, Flink, Spark 相关环境完备。 参考人员自选连接工具 |
| 4 | JDK 帮助文档 Spark, Flink 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

表 8: 数据可视化模块实施条件

| 序号 | 设备、软件名称 | 规格/技术参数、用途 | 备注 |
|----|---|--|------------------|
| 1 | 计算机 | 四核 CPU，内存 8G 或以上，win10 操作系统 或者同等条件的云服务器 | 用于软件开发和软件部署，每人一台 |
| 2 | Office 或 wps | 编写文档 | |
| 3 | Python3.0 或以上 (安装 pyecharts、matplotlib 等图形库) | 开发环境 | |
| 4 | Intelij IDEA2019 或以上， Pycharm2019 或以上 | 软件开发 | 参考人员自选开发工具 |
| 5 | MSDN 或 JDK 帮助文档 | 帮助文档 | 参考人员可以使用帮助文档 |

七、相关附录

1. 相关法律法规

2002 年 《计算机软件著作权登记办法》

2001 年 《计算机软件保护条例》

2000 年 《计算机病毒防治管理办法》

1997 年 《计算机信息系统安全专用产品检测和销售许可证管理办法》

1994 年 《中华人民共和国计算机信息系统安全保护条例》

2. 相关规范与标准

本专业标准主要依据的计算机行业技术标准和规范如表 2 所示：

表 9：引用技术标准和规范

| 序号 | 标准号 | 中文标准名称 |
|----|-----------------|----------------------|
| 1 | GB/T18142-2000 | 信息技术书籍元素值格式记法 |
| 2 | GB/T 35295-2017 | 信息技术 大数据 术语 |
| 3 | GB/T 38673-2020 | 信息技术 大数据 大数据系统基本要求 |
| 4 | GB/T 38675-2020 | 信息技术 大数据计算系统通用要求 |
| 5 | GB/T 38633-2020 | 信息技术 大数据 系统运维和管理功能要求 |

| | | |
|----|-----------------|------------------------|
| 6 | GB/T 38643-2020 | 信息技术 大数据 分析系统功能测试要求 |
| 7 | GB/T 38676-2020 | 信息技术 大数据 存储与处理系统功能测试要求 |
| 8 | GB/T 37721-2019 | 信息技术 大数据分析系统功能要求 |
| 9 | GB/T 37722-2019 | 信息技术 大数据存储与处理系统功能要求 |
| 10 | GB/T 37973-2019 | 信息安全技术 大数据安全管理指南 |
| 11 | GB/T 35274-2017 | 信息安全技术 大数据服务安全能力要求 |