大数据技术应用--数据处理师

职业能力等级评价标准

（试行稿）

1 项目概况

1.1 项目名称

数据处理师

1.2 项目定义

指根据特定的需求从事数据规范设计、数据模型设计开发、数据清洗加工、数据质量管控等数据生命周期全流程管理的从业人员。

1.3 能力等级

本项目共设三个等级，分别为:初级、中级、高级。

1.4 能力特征

具有一定的学习、表达、计算能力，手指、手臂灵活，动作协调。

1.5 职业能力等级评价要求

1.5.1 申报条件

具备以下条件之一者，可申报初级：

（1）累计从事相关职业工作1年（含）以上。

（2）相关专业在校学生。

具备以下条件之一者，可申报中级：

（1）取得本项目或相关职业初级评价证书（含职业资格证书、职业技能等级证书等）后，累计从事相关职业工作2年（含）以上。

（2）累计从事相关职业工作4年（含）以上。

（3）取得相关专业毕业证书。

具备以下条件之一者，可申报高级：

（1）取得本项目或相关职业中级评价证书（含职业资格证书、职业技能等级证书等）后，累计从事相关职业工作3年（含）以上。

（2）累计从事相关职业工作6年（含）以上。

（3）具有高等职业学校、高级技工学校、技师学院相关专业毕业证书，并取得本项目或相关职业中级评价证书（含职业资格证书、职业技能等级证书等）。

（4）具有大专及以上相关专业毕业证书，并取得本项目或相关职业中级评价证书（含职业资格证书、职业技能等级证书等）后，累计从事相关职业工作1年（含）以上。

1.5.2 申报条件注释

（1）满足本项目高级别申报条件可申报本项目低级别。

（2）相关职业:大数据技术应用相关职业。

（3）相关专业（根据《普通高等学校高等职业教育专业目录（2015年）》归类）：

电子信息类6101；

计算机类6102；

通信类6103；

经济贸易类6305（限经济信息管理专业630505）。

1.5.3 评价方式

职业能力等级评价考试包括理论知识、技能操作两个科目，较高等级必要时可增加综合评审。

理论知识考试以笔试为主，条件成熟时试点开展网络考试，主要考核从业人员从事本职业应掌握的基本要求和相关知识要求。技能操作考核主要采用现场操作、模拟操作、面试答辩等方式进行，主要考核从业人员从事本职业应具备的技能水平。综合评审通常采取审阅申报材料、技术答辩等方式进行全面评议和审查。理论知识考试和技能操作考核均采用百分制，成绩达到60分以上者为合格。

1.5.4 监考人员、考评人员与考生配比

理论知识考试和技能操作考核中的监考人员与考生配比不低于1：15，且每个考场不少于2名监考人员。技能操作考核中考评人员为3人以上单数。

1.5.5 评价时间

理论知识考试时间不少于90分钟；技能操作考核时间:初级不少于120分钟，中级/高级不少于180分钟。

1.5.6 评价场所设备

理论知识考试：在标准教室或标准联网多媒体计算机教室进行。

技能操作考核：在标准联网多媒体计算机教室进行，考生计算机需要按照考核要求安装考试系统客户端及相关应用软件，考试结束后能完成环境的还原。

2 基本要求

2.1 职业道德

（1）遵纪守法，爱岗敬业。

（2）认真严谨，忠于职守。

（3）勤奋好学，不耻下问。

（4）钻研业务，勇于创新。

（5）精益求精，工匠精神。

2.2 基础知识

2.2.1 计算机及网络知识

（1）计算机组成知识。

（2）计算机基础操作知识。

（3）计算机常用应用软件的安装及使用方法。

（4）计算机网络基础知识。

2.2.2 数据库知识

（1）关系型数据库的基本概念。

（2）SQL 基础。

（3）Oracle、MySQL、NoSQL等数据库基础。

2.2.3 云计算、大数据知识

（1）云/大数据基本概念。

（2）云产品的操作使用。

（3）典型大数据平台操作使用。

2.2.4 编程知识

（1）Java 基础知识。

（2）Python 基础知识。

（3）其他编程类基础知识。

2.2.5 主流操作系统基础知识

（1）Linux基本操作知识。

（2）Windows基本操作知识。

（3）Unix基本操作知识。

2.2.6 信息安全及合规相关知识

（1）信息安全基础知识。

2.2.7 数据管理相关国家标准

（1）数据管理知识体系。

（2）数据管理能力成熟度评估模型（DCMM）。

（3）数据管理宏观体系及关键职能操作方法。

3 工作要求

本标准初级、中级、高级的技能要求和相关知识要求依次递进，高级别涵盖低级别的要求。

3.1 初级

职业功能	工作内容	技能要求	相关知识要求
1.数据规范制定	1.1 数据结构收集	1.1.1 能使用数据库连接工具登录数据库 1.1.2 能查询获取数据库的表结构并记录	1.1.1 数据库连接工具的使用方法 1.1.2 数据库表结构查询方法
	1.2 数据字典收集	1.2.1 能登录数据库识别数据字典表 1.2.2 能识别表中数据编码字段，并进行标识 1.2.3 能登录数据库对数据字典表结构及记录进行收集	1.2.1 数据字典表的识别方法 1.2.2 数据编码字段的识别方法与标识方法 1.2.3 数据字典表的记录方法
2.数据清洗与加工	2.1 数据源配置	2.1.1 能使用数据清洗平台配置关系型数据源 2.1.2 能使用数据清洗平台配置大数据平台数据源 2.1.3 能使用数据清洗平台配置非结构化数据源 2.1.4 能使用数据清洗平台配置半结构化数据源	2.1.1 关系型数据源配置方法 2.1.2 大数据平台数据源配置方法 2.1.3 非结构化数据源配置方法 2.1.4 半结构化数据源配置方法
	2.2 清洗工作流数据输入组件配置	2.2.1 能结合数据清洗要求选择正确的输入组件 2.2.2 能对数据输入组件进行正确参数配置	2.2.1 输入组件的作用 2.2.2 输入组件的参数配置方法
	2.3 清洗组件配置	2.3.1 能选择正确组件完成数据不一致的清洗 2.3.2 能选择正确组件完成数据缺失的清洗 2.3.3 能选择正确组件完成数据重复的清洗 2.3.4 能选择正确组件完成数据合并的清洗 2.3.5 能选择正确组件完成数据排序的动作	2.3.1 数据不一致处理方法 2.3.2 数据缺失的处理方法 2.3.3 数据重复的处理方法 2.3.4 数据合并的处理方法 2.3.5 数据排序的处理方法

			2.3.6 数据错误的处理方法
	2.4 数据加工组件的配置	2.4.1 能选择正确组件完成数据分组统计 2.4.2 能选择正确组件完成不同字段数据的计算	2.4.1 数据分组统计的处理方法 2.4.2 不同字段数据的计算处理方法
	2.5 清洗工作流数据输出组件配置	2.5.1 能结合数据清洗要求选择正确的输出组件 2.5.2 能对数据输出组件进行正确参数配置	2.5.1 输出组件的作用 2.5.2 输出组件的参数配置方法
	2.6 任务调度配置	2.6.1 能使用ETL工具对任务的调度周期、定时时间进行配置 2.6.2 能使用ETL工具对任务进行调试与运行	2.6.1 ETL工具对任务的调度周期配置方法 2.6.2 ETL工具任务的调试与运行方法
3.数据质量管控	3.1 数据质量检测	3.1.1 能根据数据质量指标要求检测空值、乱码、值域检查、唯一性检查、关联性检查、及时性检查 3.1.2 能根据检测记录方法对结果进行记录与反馈	3.1.1 数据质量指标要求 3.1.2 数据质量检测记录方法
	3.2 数据质量问题报告编写与处理	3.2.1 能根据数据质量报告编写规则编写数据质量问题报告 3.2.2 能根据数据质量报告结果对数据质量问题作出初步分析与处理	3.2.1 数据质量的编写规则 3.2.2 数据质量报告中简单问题修复方法

3.2 中级

职业功能	工作内容	技能要求	相关知识要求
1.数据规范制定	1.1 数据探查	1.1.1 能使用数据探查方法查询数据库表 1.1.2 能对数据记录数量、大小、完整性、字段含义等数据情况进行收集并记录	1.1.1 数据库连接工具的使用方法 1.1.2 运用SQL语句进行数据探索的方法
	1.2 映射关系整理	1.2.1 能整理源业务系统表数据字典与标准数据字典的映射关系 1.2.2 能创建数据映射表并插入记录	1.2.1 映射关系的整理方法 1.2.2 映射关系表的创建方法

2.表模型设计	2.1 物理模型的设计	2.1.1 能确定表物理模型的数据库类型、库表名称、主键、外键、索引等表级设计要素 2.1.2 能确定表物理模型的字段名称、字段大小、字段类型、默认值等字段属性设计要素	2.1.1 物理模型的设计方法 2.1.2 建模工具设计物理模型的操作方法
	2.2 库表模型的创建	2.2.1 能通过建模工具将物理模型转换为数据库执行脚本 2.2.2 能执行物理模型的数据库创建脚本	2.2.1 物理模型转换为数据库执行脚本的方法 2.2.2 数据库执行脚本的方法
3.数据清洗与加工	3.1 较复杂的数据转换处理	3.1.1 能编写SQL对数据进行转换处理 3.1.2 能调用SQL函数进行转换处理 3.1.3 能编写自定义hive UDF对数据进行转换处理	3.1.1 hive SQL对数据转换的处理方法 3.1.2 SQL的函数使用 3.1.3 hive的UDF创建方法
	3.2 维度表处理	3.2.1 能利用ETL工具对雪花维度表、星型维度表进行加载 3.2.2 能利用ETL工具对缓慢变化维度表进行处理	3.2.1 雪花模型、星型模型的基本知识及相应ETL加载方法 3.2.2 缓慢变化维度表的知识及相应ETL加载方法
	3.3 事实表处理	3.3.1 能使用ETL工具加载周期快照事实表 3.3.2 能使用ETL工具加载累积快照事实表 3.3.3 能使用ETL工具加载事务事实表 3.3.4 能使用ETL工具加载聚集事实表	3.3.1 周期快照事实表知识及ETL工具对周期快照事实表的加载方法 3.3.2 累积快照事实表知识及ETL工具对累积快照事实表的加载方法 3.3.3 事务事实表知识及ETL工具对事实事实表的加载方法 3.3.4 聚集事实表知识及ETL工具对聚集事实表的加载方法
	3.4 任务调度	3.4.1 能制定ETL工具的任务调度策略，包括周期、运行时间、任务依赖等 3.4.2 能用ETL工具对历史数据进行补录入	3.4.1 ETL工具的不同任务调度策略规则 3.4.2 ETL工具对历史数据的处理方法 3.4.3 ETL工具对错误数

		3.4.3 能用ETL工具对错误数据任务进行排查及重跑	据、任务异常的处理方法
4.数据质量管控	4.1 数据质量策略的设定	4.1.1 能设定数据质量测量指标 4.1.2 能设定数据质量业务规则	4.1.1 数据质量测量指标知识 4.1.2 数据质量业务规则设定方法
	4.2 数据质量监控任务开发	4.2.1 能使用ETL工具开发数据质量监控任务 4.2.2 能使用ETL工具运行调试数据质量任务	4.2.1 ETL工具数据质量监控任务开发方法 4.2.2 ETL工具数据质量监控任务运行调试方法
	4.3 数据质量分析	4.3.1 能根据数据质量分析方法定位数据质量原因 4.3.2 能根据质量原因制定相应的数据质量解决方法	4.3.1 数据质量分析方法 4.3.2 数据质量解决方法制定流程
	4.4 数据质量问题处理	4.4.1 能配置ETL任务对数据质量进行修正 4.4.2 能根据方法验证数据质量问题处理完成结果	4.4.1 ETL工具数据质量修正任务的配置方法 4.4.2 数据质量处理完成结果的验证方法
5.数据处理培训	5.1 培训组织管理	5.1.1 能通过PPT 或图形工具制作培训宣传文稿 5.1.2 能完成现场培训环境的准备 5.1.3 能搭建现场培训演示环境 5.1.4 能对初级工现场培训授课和演示操作与指导	5.1.1 常用办公软件使用方法 5.1.2 演示环境的搭建方法 5.1.3 培训的规范和流程
	5.2 课件开发	5.2.1 能通过常用办公软件完成课件的开发 5.2.2 能结合课件知识点开发考试试题	5.2.1 PPT制作方法 5.2.2 试题开发规范常识
	5.3 培训质量评估	5.3.1 能制作培训评估模板 5.3.2 能根据评估反馈改进评估模板 5.3.3 能运用统计分析类方法对评估进行分析	5.3.1 评估系统知识或统计分析知识
	5.4 讲师培养	5.4.1 能对初级工进行指导 5.4.2 能组织新晋讲师的评审	5.4.1 TTT 培训技巧

3.3 高级

职业功能	工作内容	技能要求	相关知识要求
1.数据规范制定	1.1 规范数据元设计	1.1.1 能根据分类梳理规则，梳理数据元的分类 1.1.2 能根据属性描述方法，梳理数据元的属性 1.1.3 能根据数据元关联规则，识别字段对应的数据元	1.1.1 数据元分类规则 1.1.2 数据源的属性类别 1.1.3 数据元关联规则
	1.2 数据字典设计	1.2.1 能根据数据字典整理规则整理规范的数据字典 1.2.2 能通过数据质量的创建规则创建数据字典并插入记录	1.2.1 数据字典的整理方法 1.2.2 数据字典的创建方法
2.表模型设计	2.1 概念模型设计	2.1.1 能使用建模工具设计主题分类 2.1.2 能使用建模工具创建实体名称及描述 2.1.3 能使用建模工具创建实体属性 2.1.4 能通过建模工具设计实体的关联关系	2.1.1 主题的抽象方法及建模工具的主题创建方法 2.1.2 建模工具创建实体名称及描述的方法 2.1.3 建模工具创建实体属性的方法 2.1.4 实体关系知识及建模工具确定实体关系设定方法
	2.2 逻辑模型设计	2.2.1 能使用建模工具将概念模型转化为逻辑模型 2.2.2 能通过建模工具对逻辑模型进行设计	2.2.1 建模工具将概念模型转化为逻辑模型的方法 2.2.2 建模工具的逻辑模型的设计方法
3.数据清洗与加工	3.1 复杂的数据转换	3.1.1 能编写hadoop mapreduce程序对复杂转换进行处理 3.1.2 能使用ETL工具中的脚本组件对复杂转换进行处理	3.1.1 hadoop mapreduce的原理及mapreduce的开发流程 3.1.2 ETL工具的JS、JAVA等脚本组件的使用方法
	3.2 任务调度策略设计	3.2.1 能设计任务调度的策略 3.2.2 能设计任务调度的周期、时间点、运行有效时间方法等参数	3.2.1 ETL工具和操作系统的调度的选择方法 3.2.2 任务调度的依赖关系及配置方法 3.2.3 任务调度具体参数的设计方法

	3.3 清洗与加工规则设计	3.3.1 能选定不同维度表的缓慢变化处理方式 3.3.2 能识别需要数据清洗的字段并进行标识 3.3.3 能确定清洗表的表来源及多张表来源的关联方法 3.3.4 能确定清洗字段的来源字段及对应清洗规则	3.3.1 缓慢变化表的类型及使用场景 3.3.2 清洗与加工规则的梳理流程 3.3.3 表之间的关联方法 3.3.4 清洗规则制定方法
	3.4 性能调优	3.4.1 能识别数据处理任务性能不足的问题 3.4.2 能根据ETL组件性能优化方法进行组件性能优化 3.4.3 能根据SQL性能优化方法进行SQL性能优化	3.4.1 性能不足的识别方法 3.4.2 ETL组件性能不足处理方法 3.4.3 SQL性能优化方法
4.数据质量管控	4.1 数据质量维度设计	4.1.1 能根据数据质量维度方法设计质量指标 4.1.2 能设计数据质量的度量	4.1.1 数据质量指标的设计方法 4.1.2 数据质量的度量方法
	4.2 数据质量评估	4.2.1 能根据完整性度量评估数据丢失或数据不可用情况 4.2.2 能根据规范性度量评估数据未按统一格式存储的情况 4.2.3 能根据一致性度量评估数据的值在信息含义存在冲突的情况 4.2.4 能根据准确性度量评估数据不一致或超期的情况 4.2.5 能根据唯一性度量评估数据重复或数据属性重复的情况 4.2.6 能根据关联性度量评估关联的数据缺失或未建索引情况	4.2.1 完整性的度量评估方法 4.2.2 规范性的度量评估方法 4.2.3 一致性的度量评估方法 4.2.4 准确性的度量评估方法 4.2.5 唯一性的度量评估方法 4.2.6 关联性的度量评估方法
5.数据处理培训	5.1 讲师培养及管理	5.1.1 能对初、中级工进行指导 5.1.2 能结合能力地图指导讲师的提升路线	5.1.1 岗位能力模型
	5.3 计划与规范制定	5.2.1 能制定培训管理规范 5.2.2 能根据技术发展路线修改培训管理规范 5.2.3 能制定周期性不同级别的	5.2.1 培训体系管理知识

		培训计划 5.2.4 能根据技术发展及规则要求持续修订计划	

4 权重表

4.1 理论知识权重表

项目		技能等级		
		初级	中级	高级
		（%）	（%）	（%）
基本要求	职业道德	5	5	5
	基础知识	25	20	15
相关知识	数据规范制定	20	15	15
	表模型设计	—	25	20
	数据清洗与加工	30	20	15
	数据质量管控	20	10	20
	数据处理培训	—	5	10
合计		100	100	100

4.2 技能要求权重表

项目		技能等级		
		初级	中级	高级
		（%）	（%）	（%）
技能要求	数据规范制定	30	20	25
	表模型设计	—	30	25
	数据清洗与加工	50	30	20
	数据质量管控	20	15	20
	数据处理培训	—	5	10
合计		100	100	100