大数据应用开发（Python）

职业技能等级标准

（2020年1.0版）

广东泰迪智能科技股份有限公司制定

2020 年 7 月发布

目 次

前　言

本标准按照GB/T 1.1-2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

本标准起草单位：广东泰迪科技股份有限公司牵头发起，得到华为技术有限公司、蓝盾信息安全技术股份有限公司、广州智能装备研究院有限公司、中国联合网络通信股份有限公司、人民邮电出版社有限公司、网宿科技股份有限公司、广州思迈特软件有限公司、广州粤嵌通信科技股份有限公司、广州佰聆数据股份有限公司、深圳市怡亚通供应链股份有限公司、电子工业出版社有限公司、广东省人才研究会、中山大学、深圳职业技术学院等数十家大数据技术企业、行业协会、院校单位及专家学者的广泛参与支持。

本标准主要起草人：郝志峰、张良均、李兵、施兴、赵云龙、刘应吉、曾斌、钟锦辉、罗攀峰、欧陕兴、吴华夫、周永章、方海涛、蔡志杰、冯国灿、戎海武、肖刚、邓明华、余明辉、聂哲、张雅珍、苏晓、官金兰、阳永生、杨虎、刘保东。

1 范围

本标准规定了大数据应用开发（Python）职业技能等级对应的工作领域、工作任务及职业技能要求。

本标准适用于大数据应用开发（Python）职业技能培训、考核与评价，相关用人单位的人员聘用、培训与考核可参照使用。

2 规范性引用文件

下列文件对于本标准的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本标准。凡是不注日期的引用文件，其最新版本适用于本标准。

国家、行业有关标准如下：

GB/T 35295-2017 信息技术大数据术语

GB/T 5271.17-2010 信息技术词汇第17部分：数据库

GB/T 5271.1-2000 信息技术词汇第1部分：基本术语

GB/T 33745-2017 物联网 术语

3 术语和定义

国家、行业标准界定的以及下列术语和定义适用于本文件。

3.1 数据 data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

[GB/T 5271.1-2000，术语和定义 01.01.02]

3.2 大数据 big data

具有体量巨大、来源多样、生成几块、且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[GB/T 35295-2017，术语和定义 2.1.1]

3.3 关系数据库 relational database

数据按关系模型来组织的数据库。

[GB/T 5271.17-2010，术语和定义 17.04.05]

3.4 数据采集 data acquisition

又称数据获取，是利用一种装置，从系统外部采集数据并输入到系统内部的一个接口。

[GB/T 33745-2017，术语和定义 2.5.5]

3.5 数据处理 data processing

数据操作的系统执行。

[GB/T 5271.1-2000，术语和定义 01.01.06]

3.6 数据管理 data management

在数据处理系统中，提供对数据的访问、执行或监视数据的存储，以及控制输入输出操作等功能。

[GB/T 5271.1-2000，术语和定义 01.08.02]

3.7 数据分析

为提取有用信息和形成结论面对数据加以详细研究和概括总结的过程。

[GB/T 33745-2017，术语和定义 2.5.4]

3.8 数据挖掘 data mining

从大量的数据中通过算法搜索隐藏于其中信息的过程。

[GB/T 33745-2017，术语和定义 2.5.3]

4 适用院校专业

4.1 中等职业学校

大数据应用与技术、软件开发技术、计算机网络技术、数据商务、数据管理、商务数据分析与应用、软件与信息服务、计算机应用、通信技术、电子信息技术、云计算技术。

4.2 高等职业学校

大数据技术与应用、商务数据分析与应用、人工智能技术服务、智能产品开发、信息统计与分析、统计与会计核算、软件与信息服务、移动应用开发、云计算技术与应用、移动互联应用技术。

4.3 应用型本科学校

数据科学与大数据技术、大数据管理与应用、人工智能、计算机科学与技术、应用统计学、数学与应用数学、信息与计算科学、软件工程。

5 面向职业岗位（群）

主要面向包含大数据应用开发相关业务的互联网企业、传统企事业单位等的大数据应用开发、项目管理以及解决方案部门，从事数据采集、数据分析、数据挖掘、文本挖掘、算法调优等工作任务。面向的主要岗位包括数据分析师、爬虫工程师、数据可视化工程师、大数据开发工程师、算法工程师、项目经理、文本挖掘工程师等。

6 职业技能要求

6.1 职业技能等级划分

大数据应用开发（Python）职业技能等级分为三个等级：初级、中级、高级。三个级别依次递进，高级别涵盖低级别职业技能要求。

【大数据应用开发（Python）】（初级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的基础设施管理、应用软件开发部门，从事数据分析师、爬虫工程师、数据可视化工程师等工作岗位，能根据业务要求完成数据采集、数据处理、

数据分析、数据可视化等工作任务。

【大数据应用开发（Python）】（中级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的大数据应用软件开发部门，从事数据分析师、数据可视化工程师、大数据开发工程师、项目经理等工作岗位，能根据业务要求完成数据挖掘、数据可视化、基础项目管理等工作任务。

【大数据应用开发（Python）】（高级）：主要面向互联网企业以及向互联网转型的政府、企事业单位的大数据应用软件开发部门，从事算法工程师、大数据开发工程师、文本挖掘工程师、项目经理等工作岗位，能根据业务要求完成数据挖掘、文本挖掘、项目管理等工作任务。

6.2 职业技能等级要求描述

表 1 大数据应用开发（Python）职业技能等级要求（初级）

工作领域	工作任务	职业技能要求
1.平台管理	1.1 软件安装	1.1.1 能够根据操作规范，独立完成 Linux 系统的安装 1.1.2 能够根据操作规范，完成 Python 环境安装 1.1.3 能够根据操作规范，完成常见关系型数据库的安装 1.1.4 能够根据操作规范，完成数据库管理工具的安装 1.1.5 能够根据操作规范，完成数据可视化工具的安装
	1.2 软件管理	1.2.1 能够根据操作规范，进行关系型数据库用户管理 1.2.2 能够根据操作规范，进行关系型数据库权限管理 1.2.3 能够根据操作规范，在线扩展 Python 第三方库 1.2.4 能够根据操作规范，离线扩展 Python 第三方库

	1.3 系统管理	1.3.1 能够根据操作规范，进行 Linux 系统用户管理 1.3.2 能够根据操作规范，进行 Linux 系统权限管理 1.3.3 能够根据操作规范，进行 Linux 系统的内存管理 1.3.4 能够根据操作规范，进行 Linux 系统的状态监控
2.数据采集与存储	2.1 数据质量评估	2.1.1 能够根据业务需求及数据质量标准，进行数据规范性评估 2.1.2 能够根据业务需求及数据质量标准，进行数据完整性评估 2.1.3 能够根据业务需求及数据质量标准，进行数据准确性评估 2.1.4 能够根据业务需求及数据质量标准，进行数据一致性评估 2.1.5 能够根据业务需求及数据质量标准，进行数据时效性评估 2.1.6 能够根据数据质量评估结果，独立完成数据质量评估报告
	2.2 数据采集	2.2.1 能够根据业务需求，制定网页数据采集方案 2.2.2 能够根据业务需求，进行网址分析、网页分析 2.2.3 能够根据业务需求，使用 Python 采集网页数据 2.2.4 能够根据业务需求，存储采集的结构化数据
	2.3 数据存储	2.3.1 能够根据业务需求，选择数据库管理工具 2.3.2 能够根据业务需求，将结构化文件数据导入关系型数据库 2.3.3 能够根据业务需求，导出关系型数据库数据为结构化文件 2.3.4 能够根据业务需求，运用 SELECT 语句实现数据查询
3．数据分析与可视化	3.1 数据处理	3.1.1 能够根据业务需求与数据现状，进行数据中缺失值的识别与处理 3.1.2 能够根据业务需求与数据现状，进行数据中异常值的识别与处理

		3.1.3 能够根据业务需求与数据现状，进行数据的其他清洗操作 3.1.4 能够根据业务需求与数据现状，进行数据合并
	3.2 数据分析	3.2.1 能够根据业务需求与数据现状，进行描述性统计分析 3.2.2 能够根据业务需求与数据现状，进行相关性分析 3.2.3 能够根据业务需求与数据现状，进行对比分析 3.2.4 能够根据业务需求与数据现状，进行交叉分析
	3.3 数据可视化	3.3.1 能够根据业务需求，选择数据可视化工具 3.3.2 能够根据业务需求，使用数据可视化工具对数据进行基本的操作与配置 3.3.3 能够根据业务需求，绘制基础的可视化图形 3.3.4 能够根据业务需求，辅助业务人员完成数据可视化大屏

表 2 大数据应用开发（Python）职业技能等级要求（中级）

工作领域	工作任务	职业技能要求
1.平台管理	1.1 软件安装	1.1.1 能够根据操作规范，完成 Linux 系统集群的安装与配置 1.1.2 能够根据操作规范，完成 Hadoop、Storm、Spark 大数据系统的安装与配置 1.1.3 能够根据操作规范，完成 IDE 集成开发环境的安装与基础配置 1.1.4 能够根据操作规范，完成分布式数据库、分布式文件系统的安装与基础配置 1.1.5 能够根据操作规范，完成 ETL 工具的配置与安装
	1.2 软件管理	1.2.1 能够根据操作规范，对大数据平台进行状态监控、异常分析 1.2.2 能够根据监控与分析结果，对常见问题进行处理 1.2.3 能够根据操作规范，完成大数据平台的

		升级操作 1.2.4 能够根据操作规范，完成大数据组件性能优化
	1.3 系统管理	1.3.1 能够根据操作规范，完成 Linux 系统文件管理与编辑 1.3.2 能够根据操作规范，完成 Linux 系统压缩与解压 1.3.3 能够根据操作规范，完成 Linux 系统的磁盘管理与维护 1.3.4 能够根据操作规范，完成 Linux 系统的网络设置与维护
2.数据采集与存储	2.1 数据采集	2.1.1 能够根据业务需求，进行终端协议分析 2.1.2 能够根据业务需求，进行客户端数据采集 2.1.3 能够根据业务需求，进行手机 APP 数据采集 2.1.4 能够根据业务需求，完成采集的非结构化数据的存储
	2.2 数据存储	2.2.1 能够根据业务需求，运用 SQL 语句实现常规数据查询 2.2.2 能够根据业务需求，进行关系型数据库性能优化 2.2.3 能够根据业务需求，进行关系型数据库的备份与恢复 2.2.4 能够使用 Python 访问关系型数据库，进行数据操作
	2.3 数据整合	2.2.1 能够根据业务需求，选择不同的 ETL 工具 2.2.2 能够根据业务需求，实现关系型数据库数据的抽取 2.2.3 能够根据业务需求，实现本地文件数据的抽取 2.2.4 能够根据业务需求，将数据装载至关系型数据库
3.数据分析与可视化	3.1 数据处理	3.1.1 能够根据业务需求与数据现状，进行数据标准化处理 3.1.2 能够根据业务需求与数据现状，进行离散化处理 3.1.3 能够根据业务需求与数据现状，进行独热编码处理

		3.1.4 能够根据业务需求与数据现状，进行业务指标构建
	3.2 数据挖掘	3.2.1 能够根据业务需求，构建分类模型 3.2.2 能够根据业务需求，构建聚类模型 3.2.3 能够根据业务需求，构建回归模型 3.2.4 能够根据业务需求，构建智能推荐模型 3.2.5 能够根据业务需求，构建关联规则模型
	3.2 数据可视化	3.3.1 能够根据业务需求使用数据可视化工具将数据以图表的形式进行展示 3.3.2 能够根据业务需求，在业务主管的指导下根据数据分析可视化结果，形成有效的数据分析报告 3.3.3 能够通过数据分析可视化结果，得出有效的分析结论 3.3.4 能够根据业务需求，实现数据可视化大屏设计
4.项目管理	4.1 需求管理	4.1.1 能够根据项目现状，制定需求管理计划 4.1.2 能够收集业务需求并进行整理归档 4.1.3 能够根据业务，使用常用的需求收集工具与技术 4.1.4 能够根据业务，筛选需求
	4.2 进度管理	4.2.1 能够根据业务需求，对项目活动进行排序 4.2.2 能够根据业务需求，对项目活动所需资源进行规划 4.2.3 能够根据业务需求，制定项目进度计划 4.2.4 能够根据计划执行情况，调整项目计划
	4.3 变更管理	4.3.1 能够根据项目变更原则，制定研发计划 4.3.2 能够根据项目变更工作流程，推动项目研发 4.3.3 能够根据项目需求，控制变更频次 4.3.4 能够根据项目需求，制定版本发布和回退计划

表 3 大数据应用开发（Python）职业技能等级要求（高级）

工作领域	工作任务	职业技能要求

1.数据采集与存储	1.1 数据采集	1.1.1 能够根据业务需求，进行大数据采集系统的配置 1.1.2 能够根据业务需求，进行大数据采集操作 1.1.3 能够使用 Python 调用大数据采集工具，获取采集数据 1.1.4 能够根据业务需求，设计大数据采集方案
	1.2 数据存储	1.2.1 能够运用非关系型数据库工具，进行非结构化数据查询 1.2.2 能够根据业务需求，进行非关系型数据库的备份与恢复 1.2.3 能够根据业务需求，进行非关系型数据库的性能优化 1.2.4 能够根据业务需求，使用 Python 访问非关系型数据库，实现非结构化数据的操作 1.2.5 能够根据业务需求，使用 Python 访问分布式文件系统，进行文件操作
	1.3 数据整合	1.3.1 能够根据业务需求，实现数据转换操作 1.3.2 能够根据业务需求，实现 ETL 全流程编排 1.3.3 能够根据业务需求，实现定时 ETL 1.3.4 能够根据业务需求，完成数据仓库方案设计
2.数据分析与可视化	2.1 数据挖掘	2.1.1 能够根据业务需求实现算法选型 2.1.2 能够运用算法优化工具，实现算法参数调优，提升算法性能 2.1.3 能够根据业务需求使用分布式技术实现算法的并行计算，提升计算效率 2.1.4 能够根据业务需求，使用自动机器学习框架，进行数据挖掘
	2.2 文本挖掘	2.2.1 能够根据业务需求，实现文本分词与去停用词 2.2.2 能够根据业务需求，实现文本向量化 2.2.3 能够根据业务需求，实现文本分类 2.2.4 能够根据业务需求，实现文本聚类 2.2.5 能够根据业务需求，实现关键词提取 2.2.6 能够根据业务需求，实现情感分析
	2.3 深度学习建模	2.3.1 能够根据业务需求，选择合适的深度学习框架

		2.3.2 能够根据业务需求，实现全连接神经网络 2.3.3 能够根据业务需求，实现卷积神经网络 2.3.4 能够根据业务需求，实现循环神经网络 2.3.5 能够算法结果，进行深度学习算法评价
3.项目管理	3.1 立项管理	3.1.1 能够根据业务状况，完成项目建议书 3.1.2 能够根据可行性研究步骤，完成项目可行性研究报告 3.1.3 能够根据可行性研究报告，进行项目效益的预测与评估 3.1.4 能够根据项目招投标流程，跟踪招投标进度
	3.2 质量管理	3.2.1 能够根据质量管理流程，规划质量管理 3.2.2 能够根据现有的质量管理标准体系，实施质量保证 3.2.3 能够正确使用项目质量管理规划阶段技术与工具 3.2.4 能够正确使用项目质量管理执行阶段技术与工具
	3.3 人力资源管理	3.3.1 能够根据人力资源管理的流程，绘制项目组织图 3.3.2 能够根据项目需求，组建项目团队 3.3.3 能够根据项目需求，制定人力资源管理计划 3.3.4 能够根据项目需求，制定团队绩效评价
	3.4 风险管理	3.3.1 能够结合现有情况，识别项目风险 3.3.2 能够运用定性分析，分析项目风险 3.3.3 能够运用定量分析，分析项目分享 3.3.4 能够针对可能出现的风险制定风险应对方案

参考文献

[1] GB/T 35295-2017 信息技术大数据术语

[2] GB/T 5271.17-2010 信息技术词汇第 17 部分：数据库

[3] GB/T 5271.1-2000 信息技术词汇第 1 部分：基本术语

[4] GB/T 33745-2017 物联网 术语