# 大数据应用部署与调优

# 职业技能等级标准

标准代码：510065

（2021年2.0版）

**南京云创大数据科技股份有限公司　制定**

2021 年 12 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：南京云创大数据科技股份有限公司、星环信息科技（上海）有限公司、天津南大通用数据技术股份有限公司、天津神舟通用数据技术有限公司、柏睿数据科技有限公司、中国外汇交易中心、上海亿发信息科技有限公司、金陵科技学院、国防科技大学、长春职业技术学院、重庆电子工程职业学院、天津电子信息职业技术学院、黄河水利职业技术学院。

本标准主要起草人：刘鹏、罗圣美、张真、张燕、杨洪山、赵伟、高阳、姜才康、陶建辉、夏志江、朱辉、何玮、惠军华、齐志、李腾、孙健、贾红军。

## 1 范围

本标准规定了大数据应用部署与调优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部署与调优职业技能培训、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。

GB/T 35295-2017 信息技术 大数据 术语

GB/T 37721-2019 信息技术 大数据分析系统功能要求

GB/T 37722-2019 信息技术 大数据存储与处理系统功能要求

GB/T 37973-2019 信息安全 技术大数据安全管理指南

GB/T 5271.1-2000 信息技术 词汇 第 1 部分：基本术语

ISO/IEC 20547-4 信息技术 大数据参考架构 第 4 部分：安全与隐私保护

ITU-T Y. 3600 大数据 基于云计算的要求和能力

## 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标准。

3.1

**Hadoop**

是一个开源的可运行于大规模集群上的分布式文件系统和运行处理基础框架，用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用集群的能力进行高速运算和存储。

3.2

**Spark**

一种专为大规模数据处理而设计的快速通用的计算引擎。该计算引擎基于Scala语言实现，可以像操作本地集合对象一样操作分布式数据集。

3.3

**MySQL**

业界流行的开源关系型数据库管理系统，是中小型网站开发的主要数据库。

3.4

**Python**

提供了高效的高级数据结构和面向对象编程，是多数平台编写脚本和快速开发应用的编程语言。

3.5

**Shell 脚本**

Shell脚本是使用Linux/Unix下的命令，编写的程序执行文件，程序文件方便多次使用，并且可以通过其中的语法实现多种处理动作。

## 4 适用院校专业

### 4.1 参照原版专业目录

中等职业学校：计算机应用、计算机网络技术、软件与信息服务、电子与信息技术、物联网技术应用、网络信息安全、网站建设与管理、网站安防系统安装与维护专业。

高等职业学校：计算机应用技术、计算机网络技术、计算机信息管理、计算机系统与维护、软件技术、软件与信息服务、云计算技术与应用、大数据技术与应用、电

子商务技术、人工智能技术服务专业。

高等职业教育本科学校：电子信息工程、物联网工程、计算机应用工程、网络工程、软件工程、大数据技术与应用、信息安全与管理、虚拟现实技术与应用等专业使用。

应用型本科学校：数据科学与大数据技术、计算机科学与技术、软件工程、网络工程、智能科学与技术、大数据管理与应用、信息管理与信息系统专业。

### 4.2 参照新版职业教育专业目录

中等职业学校：计算机应用、计算机网络技术、软件与信息服务、电子信息技术、物联网技术应用、网络信息安全、网站建设与管理、网站安防系统安装与维护专业。

高等职业学校：计算机应用技术、计算机网络技术、大数据技术、软件技术、云计算技术应用、人工智能技术应用专业。

高等职业教育本科学校：电子信息工程技术、物联网工程技术、计算机应用工程、网络工程技术、软件工程技术、大数据工程技术、信息安全与管理、虚拟现实技术、人工智能工程技术、工业互联网技术专业。

应用型本科学校：数据科学与大数据技术、计算机科学与技术、软件工程、网络工程、智能科学与技术、大数据管理与应用、信息管理与信息系统专业。

## 5 面向职业岗位（群）

【大数据应用部署与调优】（初级）：主要面向大数据运维人员、数据清洗工程师、需求工程师等职业岗位，主要完成数据采集、清洗、展示、相关全生命周期任务管理、大数据日常运维管理以及大数据组件模块的需求创建与记录等工作。

【大数据应用部署与调优】（中级）：主要面向大数据研发工程师、数据分析师、大数据产品工程师等职业岗位，主要从事大数据实时指标分析、性能调优、数据分析

仓库建模、分布式集群故障排查等工作。

【大数据应用部署与调优】（高级）：主要面向系统性能优化工程师、分布式系统架构师等职业岗位，主要完成大规模集群架构、调优、底层核心故障排查、集群迁移、组件架构升级以及容灾备份等架构设计、升级等工作。

## 6 职业技能要求

### 6.1 职业技能等级划分

大数据应用部署与调优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。

【大数据应用部署与调优】（初级）：根据操作文档的要求，独立完成云计算平台搭建、操作系统安装、Hadoop 平台配置与部署，能够对常见故障进行识别、判断和处理，满足日常运维、故障管理和日常巡检要求，以及承担数据采集、预处理、存储和处理等基础工作任务。

【大数据应用部署与调优】（中级）：根据操作手册的要求，独立完成 Spark、应用开发环境配置与部署，采用日志管理和配置管理等工具方法，完成复杂故障的识别、判断和处理，参与运维管理流程制定，掌握 SQL 语法和 Python 编程语言，承担大数据应用部署和调试等工作任务。

【大数据应用部署与调优】（高级）：根据业务规划，独立完成 Hadoop 和 Spark 系统组件升级，掌握安全管理、性能优化和容灾备份等设计方法，保证业务连续性要求，掌握数据获取、分析和可视化方法，承担深度运维和应用调优等工作任务。

### 6.2 职业技能等级要求描述

表 1 大数据应用部署与调优职业技能等级要求（初级）

| 工作领域 | 工作任务 | 职业技能要求 |
| --- | --- | --- |
| 1.平台搭建 | 1.1 云计算平台搭建 | 1.1.1 能根据操作文档搭建 OpenStack 私有云平台，安装 Nova、Swift、Glance、Dashboard 等组件。<br>1.1.2 能在专业人员指导下，协助完成私有云平台搭建过程中的排错工作。<br>1.1.3 能根据操作文档，正确使用公有云资源，如云服务器、对象存储服务、云网络服务、云安全服务、云数据库服务等。 |
| | 1.2 操作系统安装 | 1.2.1 能独立配置操作系统所需虚拟环境。<br>1.2.2 能独立完成操作系统的安装与设置。<br>1.2.3 能独立进行操作系统补丁安装。<br>1.2.4 能同时使用图形化和命令行界面进行操作。 |
| | 1.3 Hadoop 系统部署 | 1.3.1 能根据操作手册进行 Hadoop 单节点部署。<br>1.3.2 能根据操作手册进行 Hadoop 伪分布式部署。<br>1.3.3 能根据操作手册进行 Hadoop 集群部署。<br>1.3.4 能在搭建好的集群上运行自带测试用例，进行错误排查。 |
| 2.系统调优 | 2.1 运维管理 | 2.1.1 能运用供应商提供的系统运维管理工具或者可视化界面，独立完成大数据架构下各个模块（如：HDFS、Spark、配置信息）的日常运维管理操作，如：系统状态监测、系统日志收集、日常巡检等。<br>2.1.2 能运用 Shell 命令完成对各个节点的日常运维操作，如：运维脚本查询、存储使用情况、机器运行状态、网络防火墙配置。<br>2.1.3 能根据业务需求，运用数据库管理工具（Plsql，navicat），完成数据库日常监控和运维。<br>2.1.4 能协助高级技术支持人员梳理优化运维管理流程，补充系统运维操作手册等相关文档。 |
| | 2.2 故障管理 | 2.2.1 能运用日常运维实例和网络检索实例，定位日常简单故障，评估可能造成的影响范围，并及时上报。<br>2.2.2 能运用系统故障处理的常用方法和工具，独立分析常见故障的原因，提出改进建议和方法措施。<br>2.2.3 能根据软件提供的系统故障处理手册，运用故障诊断工具或系统自带的故障诊断命令，进行应急处理，保障系统的稳定运行。<br>2.2.4 能在故障闭环后，根据故障现象、解决过程、核心原因、处理方案等多方面因素，编写相关故障解决文档。 |

| | | |
|---|---|---|
| | 2.3 日常巡检 | 2.3.1 能使用基础运维工具完成硬件资源信息的收集汇总，协助硬件管理人员完成周期性更新勘误。<br>2.3.2 能使用基础运维工具完成软件部署信息的收集汇总，并独立完成软件服务部署信息的周期性更新勘误。<br>2.3.3 能使用基础运维工具完成网络资源分配信息的收集汇总，并协助网络资源管理人员完成对网络拓扑结构和资源 IP 的周期性更新勘误。<br>2.3.4 能根据既有常用组织架构体系，并通过供应商所提供的配置管理工具，完成对具体岗位职务的日常运维工作。 |
| 3.应用部署 | 3.1 数据采集 | 3.1.1 能运用大数据技术（SparkStreaming、Kafka 等），完成数据采集系统的搭建和基础配置。<br>3.1.2 能运用 Python 编写基础的爬虫脚本，完成对指定网页的数据爬取。<br>3.1.3 能运用 Flume 技术，完成大数据架构体系的系统日志收集。<br>3.1.4 能运用 ETL 工具（Kettle），完成 excel/json/csv 等通用格式的数据采集任务。<br>3.1.5 能运用 Kettle 构建定时调度作业，完成对已有采集任务的自动优化。 |
| | 3.2 数据预处理 | 3.2.1 能运用 ETL 工具（Kettle），完成数据转换，合并与拆分动作。<br>3.2.2 能运用 ETL 工具（Kettle），对关键字段完成数据脱敏。<br>3.2.3 能运用 Python 语言,实现对已采集数据的清洗。<br>3.2.4 能通过 ETL 工具(Sqoop),完成从关系数据库到 NoSQL 数据库的数据迁移。 |
| | 3.3 大数据存储 | 3.3.1 能运用大数据技术(如分布式文件系统 HDFS)维护、管理数据存储系统，优化存储资源利用率和计算效率。<br>3.3.2 能独立完成 HBase 分布式集群环境部署。<br>3.3.3 能运用 HBase Shell 完成日常节点维护和数据查询导出等维护动作。 |
| | 3.4 大数据处理 | 3.4.1 能运用 Spark 技术，完成多中间件日志格式的解析和系统各种基础指标的日常统计。<br>3.4.2 能运用 Neo4j 技术，完成对图形化数据的可视化输出。<br>3.4.3 能运用日常数据库技术，优化 ETL 工具流程，监控并维护例行数据 ETL 任务。<br>3.4.4 能运用标注工具 Labelme，完成对日常巡检图像数据的标注和预警。 |

表 2 大数据应用部署与调优职业技能等级要求（中级）

| 工作领域 | 工作任务 | 职业技能要求 |
| --- | --- | --- |
| 1.平台搭建 | 1.1 Spark 系统部署 | 1.1.1 能根据操作手册，独立进行多种形式的 Spark 部署。<br>1.1.2 能部署与 Hadoop 组件相结合的 Spark on Yarn 集群。<br>1.1.3 能独立搭建高可用 Spark 集群。<br>1.1.4 能进入 Spark 命令行界面，运行测试验证搭建效果。 |
| | 1.2 数据库软件安装 | 1.2.1 能运用操作系统（如 Windows，Linux）安装工具，独立完成 MySQL 数据库的软件下载。<br>1.2.2 能运用 Xshell，完成 MySQL 数据库软件的传输以及解压。<br>1.2.3 能正确完成 MySQL 数据库在 Windows 环境下的安装、配置文件的编写、环境变量的配置。<br>1.2.4 能正确完成 MySQL 数据库在 Linux 环境下的安装、服务开启以及变量配置。 |
| | 1.3 Python 系统安装 | 1.3.1 能独立完成 Python 系统的下载以及安装。<br>1.3.2 能独立完成 Python 系统的环境变量配置。<br>1.3.3 能独立完成 Python 库安装。<br>1.3.4 能独立完成 IDE 集成软件开发环境（如 Pycharm）的软件安装和高级功能参数配置。 |
| 2.系统调优 | 2.1 性能管理 | 2.1.1 能独立完成集群下 HDFS 配置优化，包括数据块大小设置、I/O 读写能力优化。<br>2.1.2 能独立完成集群下 MapReduce 对 I/O 参数、并行传输数据以及 Tasktracker 的 RPC 请求数优化。<br>2.1.3 能独立完成集群下 Spark 配置优化，包括读取、写入和架构资源配置。<br>2.1.4 能独立完成集群下架构环境中的各个组件性能优化（Zookeeper、消息队列、容器配置）。 |
| | 2.2 日志管理 | 2.2.1 能通过 Shell 命令以及厂商所提供的运维日志管理相关文档，完成对指定服务的实时日志查询，并能根据日志结果判断服务运行的状况。<br>2.2.2 能独立构建日志管理相关规范，按照日志类型、日志级别、输出内容重要度、有效日期等制定对日志的统一分类归档标准。<br>2.2.3 能够根据 Flume 采集的日志，完成对日常日志稽核的统计与分析。<br>2.2.4 能通过构建定时调度作业，完成对服务器日志文件的定时清理与归档。 |

| | | |
|---|---|---|
| | 2.3 配置管理 | 2.3.1 能根据已有架构，制定周期性巡检计划，补充完善巡检内容。<br>2.3.2 能根据实际机房标准和硬件标准，制定硬件及环境巡检标准，设置检查项。<br>2.3.3 能根据业务连续性要求和软件健康性标准，完成软件网络巡检标准，设置检查项。<br>2.3.4 能根据巡检中遇到的可疑或者异常情况，及时督促协调厂商，完成对异常情况的排查处理。 |
| 3.应用部署 | 3.1 数据查询 | 3.1.1 能独立编写操作数据库的语法（如创建、查看、选择、删除数据库等）。<br>3.1.2 能独立编写数据表数据的新增、批量新增以及删除的操作。<br>3.1.3 能独立编写数据表数据的查询，包括：条件查询、联表查询、分组查询等操作。<br>3.1.4 能独立完成对数据表进行结构上的修改、调整及权限添加、回收。 |
| | 3.2 编程基础 | 3.2.1 能深入理解程序设计语言、常见编程方法以及 Python 编码规范。<br>3.2.2 能独立完成 Python 简单程序开发（交互式编程和脚本式编程两种方式）。<br>3.2.3 能深入理解 Python 多个语句构成代码组以及命令行参数，完成样例开发。 |
| | 3.3 应用开发 | 3.3.1 能独立完成 Python 赋值语句的编写，使用 Python 顺序结构和选择结构进行编程。<br>3.3.2 能独立完成 Python 图形程序编写。<br>3.3.3 能采用函数定义及调用方法，独立完成 Python 函数的使用。<br>3.3.4 能掌握 Python 模块和导入方法，独立完成第三方模块安装和使用。 |

**表 3 大数据应用部署与调优职业技能等级要求（高级）**

| 工作领域 | 工作任务 | 职业技能要求 |
|---|---|---|
| 1.平台搭建 | 1.1 Hadoop 系统升级 | 1.1.1 能独立进行 Hadoop 升级配置。<br>1.1.2 能根据业务需求，选择合适的 HDFS 升级配置，并进行升级。<br>1.1.3 能在 HDFS 出现升级问题时快速进行回滚操作，防止数据丢失。 |

| | | |
|---|---|---|
| | | 1.1.4 能根据部署手册，独立升级 YARN。 |
| | 1.2 Spark 系统升级 | 1.2.1 能根据业务需求，选择合适的 Spark 升级配置，并独立完成升级操作。<br>1.2.2 能根据日志快速排查 Spark 升级时产生的兼容性错误。<br>1.2.3 能自行测试多个 Spark 组件在完成升级后是否能够正常使用，排查兼容性问题。<br>1.2.4 能运用项目文档编写工具和模板，独立整理和编写 Spark 升级报告文档和技术支持文档。 |
| 2.系统调优 | 2.1 安全管理 | 2.1.1 能深度理解资产安全和应用安全的架构体系，屏蔽部分系统安全风险。<br>2.1.2 能构建部门内安全管理规范及人员角色权限体系，并制定安全管理规范。<br>2.1.3 能独立完成所有组件漏洞版本的及时更新，并提供漏洞补救方案。<br>2.1.4 能运用禅道/Confluence 等产品，规划构建企业内部的漏洞风险管理知识库。 |
| | 2.2 系统优化 | 2.2.1 能独立完成 Hadoop 集群环境各个组件的配置优化，整合系统、网络及硬件资源。<br>2.2.2 能独立解决 Hadoop 集群环境各个组件负载优化和负载倾斜问题。<br>2.2.3 能独立完成集群环境下的复杂架构问题的排查与处理，给出问题根本解决方案。<br>2.2.4 能独立完成 Hadoop 集群下的性能优化。 |
| | 2.3 高可用 | 2.3.1 能通过借鉴已有经验和相关案例，构建大数据系统的容灾备份方案。<br>2.3.2 能针对企业核心业务，提供应急预案，并能按照预案内容进行日常演练。<br>2.3.3 能从架构层面完善优化系统的健壮性，并回溯构建开发规范，提高业务代码的健壮性和稳定性。 |
| 3.应用部署 | 3.1 数据获取 | 3.1.1 能熟练使用 Python 语言及相关工具包，发送网络请求，并获得响应。<br>3.1.2 能熟练使用 Python 语言，完成文档解析，并过滤出符合要求的数据。<br>3.1.3 能将获取到的有效数据数据存储在 MySQL 数据库中。<br>3.1.4 能利用编程自动探索和发现新的数据，并且自动持续地获取数据。 |

| | | |
|---|---|---|
| | 3.2 数据分析 | 3.2.1 能用编程方法完成数据频数的分析。<br>3.2.2 能用编程方法完成数据集中趋势的分析。<br>3.2.3 能用编程方法完成数据离散程度的分析。<br>3.2.4 能用编程方法完成数据分布的分析。 |
| | 3.3 数据可视化 | 3.3.1 能利用大数据平台提供的各种工具为用户提供海量数据环境下的直接查询功能。<br>3.3.2 能制作柱状图、饼图、折线图等基本图形，表示数据挖掘的结果。<br>3.3.3 能制作词云、热点图、三维图等图形，表示数据挖掘结果。<br>3.3.4 能撰写数据分析报告,并转换为被用户理解的知识。 |

# 参考文献

[1] GB/T 32400-2015 信息技术 云计算 概览与词汇

[2] GB/T 35295-2017 信息技术 大数据 术语

[3] GB/T 35295-2017 信息技术 大数据 术语

[4] GB/T 5271.1-2000 信息技术 词汇 第 1 部分：基本术语

[5] GB/T 37973-2019 信息安全 技术大数据安全管理指南

[6] GB/T 37722-2019 信息技术 大数据存储与处理系统功能要求

[7] GB/T 37721-2019 信息技术 大数据分析系统功能要求

[8] ISO/IEC 20547-4 信息技术 大数据参考架构 第 4 部分：安全与隐私保护

[9] ITU-T Y.3600 大数据 基于云计算的要求和能力

[10]《中华人民共和国职业分类大典》（2015 年版）

[11]国家职业技能标准编制技术规程（2018年版）

[12] 教育部关于印发《职业教育专业目录（2021年）》的通知（教职成〔2021〕2号）

[13] 《教育部关于公布2019年度普通高等学校本科专业备案和审批结果的通知》（教高函〔2020〕2号）

[14] 《教育部关于公布2020年度普通高等学校本科专业备案和审批结果的通知》（教高函〔2021〕1号）