CDA LEVEL Ⅱ 大数据分析师考试大纲

CERTIFIED DATA ANALYST LEVEL Ⅱ EXAMINATION OUTLINE

一、总则

「CDA 数据分析师人才行业标准」是面向全行业数据分析及大数据相关岗位的一套科学化、专业化、正规化、系统化的人才技能准则。经管之家 CDA 数据分析师认证考试是评判「标准化人才」的唯一考核路径。CDA 考试大纲规定并明确了数据分析师认证考试的具体范围、内容和知识点，考生可按照大纲要求进行相关知识的学习，获取技能，成为专业人才。

二、考试形式与试卷结构

考试方式：线下统考，上机答题

考试题型：客观题（单选+多选）+ 上机建模题

考试时间：210 分钟

考试内容：第一阶段，90 分钟，客观题（单选+多选），上机答题；第二阶段 120 分钟，案例操作，自行携带电脑操作，案例数据将统一提供 CSV 文件。

（关于上机案例操作的相关说明：

1. 个人电脑需安装好 Hadoop 集群，必要组件及版本如下：
   Hadoop 2.7 及以上、Spark 2.1 及以上，Hive 2.2.0 或 1.2.2 及以上，MySQL 5.5 及以上、Sqoop 1.4.7 或 1.99.7；
2. 视个人电脑配置情况，搭建伪分布式或分布式集群均可，由于网络限制，不可使用云服务器进行操作，可在虚拟机中搭建 Hadoop 集群，虚拟机推荐使用 Ubuntu 或 CentOS，系统版本没有要求；
3. 建模数据量约 200-300M，为保证顺利建模，需预留至少 1G 内存空间，伪分布式也可运行；）

考试成绩：分为 A、B、C、D 四个层次，A、B、C 为通过考试，D 为不通过。

三、知识要求

针对不同知识，掌握程度的要求分为【领会】、【熟知】、【应用】三个级别，考生应按照

不同知识要求进行学习。

1．领会：考生能够领会了解规定的知识点，并能够了解规定知识点的内涵与外延，了解其内容要点和它们之间的区别与联系，并能做出正确的阐述、解释和说明。

2．熟知：考生须掌握知识的要点，并能够正确理解和记忆相关理论方法，能够根据不同要求，做出逻辑严密的解释、说明和阐述。此部分为考试的重点部分。

3．应用：考生须学会将知识点落地实践，并能够结合相关工具进行商业应用，能够根据具体要求，给出问题的具体实施流程和策略。

四、考试范围

- **PART 1 大数据基础理论　占比（8%）**
  - a. 大数据分析基础（1%）
  - b. Python 基础（5%）
  - c. Linux & Ubuntu 操作系统基础（2%）
- **PART 2 Hadoop 理论　占比（12%）**
  - a. Hadoop 安装配置及运行机制解析（2%）
  - b. HDFS 分布式文件系统（2%）
  - c. MapReduce 理论及实战（2%）
  - d. Hadoop 生态其他常用组件（6%）
- **PART 3 大数据分析之数据库理论及工具　占比（16%）**
  - a. 数据库导论（2%）
  - b. MySQL 理论及实战（3%）
  - c. HBase 安装及使用（3%）
  - d. Hive 安装及使用（5%）
  - e. Sqoop 安装及使用（3%）
- **PART 4 大数据分析之数据挖掘理论基础　占比（10%）**
  - a. 数据挖掘的基本思想（2%）
  - b. 数据挖掘基本方法介绍（2%）
  - c. 有监督学习算法（4%）
  - d. 无监督学习算法（2%）
- **PART 5 大数据分析之 Spark 工具及实战　占比（35%）**

a. Spark 基础理论（2%）

b. Spark RDD 基本概念及常用操作（3%）

c. Spark 流式计算框架 Spark Streaming、Structured Streaming（5%）

d. Spark 交互式数据查询框架 Spark SQL（5%）

e. Spark 机器学习算法库 Spark MLlib 基本使用方法（15%）

f. Spark 图计算框架 GraphX（5%）

◆ **PART 6 大数据分析之数据可视化方法 占比（4%）**

a. 数据可视化入门基础（1%）

b. Python 数据可视化入门（2%）

c. Python 高级数据可视化方法（1%）

◆ **PART 7 大数据分析实战 占比（15%）**

a. 利用 HDFS Shell 操作 HDFS 文件系统（1%）

b. 利用 Hive SQL 进行数据清洗（2%）

c. 利用 Sqoop 进行数据传输（1%）

d. 利用 Spark SQL 进行数据读取（2%）

e. 利用 Spark MLlib 进行机器学习建模（8%）

f. 利用 Python 进行建模结果数据可视化（1%）

五、考试内容

PART 1 大数据基础理论

◆ **1、大数据分析基础**

【领会】

大数据技术诞生技术背景

大数据技术实际应用

分布式处理技术概念

数据分析和数据挖掘的概念

【熟知】

明确数据分析的目标和意义

明确分布式技术在进行海量数据处理时起到的关键作用

数据分析方法与数据挖掘方法的区别和联系

明确数据分析流程中不同软件工具的作用

常用描述性统计方法

常用数据挖掘方法

◆ **2、Python 基础**

【领会】

Python 语言的特点、语法、应用场景

【熟知】

Python 基础语法，包括基本数据类型、运算符、条件控制语句、循环语句等；

Python 函数式编程，常用高阶函数，包括 map 函数、reduce 函数、filter 函数及模块相关功能

Python 面向对象编程特性，包括类和实例、继承、多态

利用 Python 链接数据库

Python 可视化常用包及其基本使用方法

◆ **3、Linux 与 Ubuntu 基础**

【领会】

Linux 入门

Linux 与 Ubuntu 的关系

Ubuntu 的安装及配置

Ubuntu 文件组织形式

Ubuntu 操作系统的常用命令

SSH 理论基础

了解其他常用 Linux 系统，如 CentOS，RedHat，SUSE 等

【熟知】

Ubuntu 操作系统命令及使用命令编辑文件

IP 地址的基础理论

SSH 命令使用方法

利用 SSH 基于密匙的安全验证进行多个节点间的无密码登陆

【应用】

安装配置 Linux 操作系统

利用 SSH 基于密匙的安全验证进行多个节点间的无密码登陆

掌握部分 shell 命令进行 Linux 操作，如 awk、grep、sed 典型的文本处理工具

PART 2 Hadoop 理论

◆ 1、Hadoop 安装配置及运行机制解析

【领会】

分布式系统设计的基本思想

Hadoop 概念、版本、历史

Hadoop 单机、伪分布及集群模式的安装配置步骤

如何通过命令行和浏览器观察 Hadoop 的运行状态

【熟知】

Hadoop 单机、伪分布及集群模式的安装配置过程和内容

Hadoop 参数格式

Hadoop 参数的修改与优化

Hadoop 的安全模式

【应用】

进行 Hadoop 集群的配置

查看和管理 Hadoop 集群

Hadoop 运行的日志信息查看与分析

◆ 2、HDFS 分布式文件系统

【领会】

HDFS 的概念及设计

HDFS 体系结构及运行机制，

NameNode、DataNode、SecondaryNameNode 的作用及运行机制

HDFS 的备份机制和文件管理机制

【熟知】

HDFS 的运行机制

NameNode、DataNode、SecondaryNameNode 的配置文件

HDFS 文件系统的常用命令

【应用】

使用命令及 Java 语句操作 HDFS 中的文件

使用 JPS 查看 NameNode、DataNode、SecondaryNameNode 的运行状态

◆ **3、MapReduce 理论及实战**

【领会】

MapReduce 的概念及设计

MapReduce 运行过程中类的调用过程

Mapper 类和 Reducer 类的继承机制

job 的生命周期

MapReduce 中 block 的调度及作业分配机制

【熟知】

MapReduce 程序编写的主要内容

MapReduce 程序提交的执行过程

MapReduce 程序在浏览器的查看

【应用】

Mapper 类和 Reducer 类的主要编写内容和模式

job 的实现和编写

编写基于 MapReduce 模型的 wordcount 程序

相应 jar 包的打包和集群运行

◆ **4、Hadoop 生态其他常用组件**

【领会】

HBase 基本功能、Hive 基本功能、Sqoop 基本功能、ZooKeeper 的基本功能、Flink 基本功能

【熟知】

HBase 的安装配置及常用命令、Hive 的安装配置及常用命令、Sqoop 的安装配置及常用命令、ZooKeeper 的安装配置及常用命令、Flink 安装配置及常用命令

【应用】

HBase、Hive、Sqoop、Flink 及 ZooKeeper 的安装与运行

PART 3 大数据分析之数据库理论及工具

◆ **1、数据库导论**

【领会】

数据、数据库、数据库管理系统、数据库系统、数据仓库的概念

数据管理发展的三个阶段，不同阶段数据管理的特点，特别是数据库系统的特点

数据依赖及数据规范化理论、数据模型理论及方法

【熟知】

SQL 的基本概念和特点

SQL 的数据定义功能

SQL 的数据查询功能

CRUD 操作

SQL 的数据更新功能

不同 NoSQL 数据库的特点及使用场合

◆ **2、MySQL 理论及实战**

【领会】

数据库、表、索引和视图的相关概念

数据库完整性约束的概念、定义及使用方法

数据库、表、索引和视图的维护方法

【熟知】

MySQL 中 SELECT 命令的基本格式

掌握单表查询的方法和技巧

掌握多表连接查询的方法和技巧

掌握嵌套查询、集合查询的方法和技巧

【应用】

MySQL 平台下的 SQL 交互操作

◆ **3、Hive 数据仓库基础**

【领会】

Hive 数据仓库在 Hadoop 生态系统中的地位

【熟知】

Hive 与 HBase 的区别

【应用】

使用 Hive 进行频率统计

◆ **4、Hive 的基本命令**

【领会】

Hive 中的数据库概念、修改数据库

【熟知】

创建表、管理表、外部表、分区表、删除表

【应用】

向表中增加数据

通过查询语句向表中插入数据

单个查询语句中创建表并加载数据

导出数据

◆ **5、Hive 中检索数据**

【领会】

Hive 中的命令语句是类 SQL 语句

【熟知】

SELECT…FROM 语句

【应用】

使用列值进行计算、算术运算符、使用函数、列别名、嵌套 SELECT 语句、WHERE 语句、group by 语句、集合运算、多表连接、内连接、外连接、笛卡尔积连接、order by 语句、抽样查询、视图。

◆ **6、Sqoop 基础**

【领会】

Sqoop 是一个数据转储工具，它能够将 Hadoop HDFS 中的数据转储到关系型数据库中，也能将关系型数据库中的数据转储到 HDFS 中。

【熟知】

Sqoop 链接数据库需要 JDBC 的支持

【应用】

Sqoop 的安装方法

从 Hadoop HDFS 向 MySQL 导入数据

从 MySQL 向 Hadoop HDFS 导入数据

◆ **7、HBase 理论及实战**

【领会】

HBase 的基础概念、数据模型、存储模型

HBase 集群配置参数分析

HBase 集群查看方式

【熟知】

HBase shell 常用的操作命令

HBase 的参数配置

HBase 的每个数据单元的操作方式

区域服务器(Region Server)和主服务器(Master Server)的管理模式

HBase 的存储模式

【应用】

HBase 的伪分布和集群的安装及配置

HBase 的 API 操作项目实战

PART 4 大数据分析之数据挖掘理论基础

◆ 1、数据挖掘的基本思想

【领会】

常用数据挖掘的概念和方法介绍

【熟知】

数据挖掘的常用算法

数据挖掘的过程

数据挖掘的常用工具及数据挖掘的应用场景

◆ 2、数据挖掘基础知识

【熟知】

数据、算法基本概念

算法基本分类方法

有监督学习算法中的训练样本、测试样本、特征变量、目标变量（标签）等常用术语的相关定义

◆ 3、有监督学习算法

【领会】

有监督学习算法基本定义

分类算法与回归算法区别与联系

【熟知】

掌握以下常用有监督学习算法基本原理：

最近邻分类器 KNN、朴素贝叶斯、线性回归、逻辑回归、决策树、集成算法（包括 Bagging 和 Boosting 两类算法）、支持向量机（SVM）、神经网络、协同过滤等

掌握用于分类算法评价的指标体系，包括混淆矩阵内的一级评估指标，以及基于混淆矩阵的准确率（accuracy）、召回率（recall）、精确度（precaution）、F 指标、AUC 曲线等各项指标的评估方法。

掌握包括 SSE、SSR、MSE、R 方、调整 R 方在内的回归类算法评估指标。

【应用】

能借助常用数据挖掘工具实现上述常用有监督学习算法，并完成相应数据挖掘工作

◆ **4、无监督学习算法**

【领会】

无监督学习算法基本定义

聚类算法和关联规则算法基本定义

【熟知】

掌握常用关联规则挖掘算法，包括 Apriori、FP-Growth 等

掌握常用聚类算法，包括层次聚类、K-Means 快速聚类、DBSCAN 聚类算法等

【应用】

能借助常用数据挖掘工具实现上述常用无监督学习算法，并完成相应数据挖掘工作

PART 5 大数据分析之 Spark 工具及实战

◆ **1、Spark 基础理论**

【领会】

Spark 大数据生态系统的功能与结构

Spark、Hadoop 之间的区别与联系

Spark 大数据生态系统的特点

Scala 基本语法

【熟知】

Spark 生态系统中的四大核心组件

Spark 与 MapReduce 的对比与分析

Spark 与 MapReduce 适用的应用场景

Spark 的多种运行模式

【应用】

熟练掌握 Standalone 模式下 Spark 集群的搭建步骤

配置文件中参数的具体含义

◆ **2、Spark RDD 基本概念与常用操作**

【领会】

Spark RDD 基本概念

Spark API

Spark 任务调度策略

【熟知】

Spark RDD 的特性

RDD 上的转换操作、执行操作、持久化操作

RDD 之间的宽依赖关系与窄依赖关系

【应用】

基于 Spark API 编写词频统计程序，并在词频统计程序基础上进行功能扩展，SparkContext、TaskScheduler、DAGScheduler 等核心代码的分析与调试。

◆ **3、Spark 流式计算框架 Spark Streaming、Structured Streaming**

【领会】

Kafka 分布式消息分发机制

Spark Streaming 应用场景

Spark Streaming 基本概念

Spark DStream 的存储级别

Structured Streaming 计算框架

【熟知】

批处理间隔、离散数据流 Spark DStream、窗口、滑动间隔、窗口间隔等重要概念

熟练使用 Spark DStream 的相关操作

Spark Streaming 的三种应用模式，以及实现三种模式的相关操作

【应用】

搭建 Kafka 环境，能够将 Kafka 作为高级数据源时使用 Spark Streaming，基于 HDFS 上文本数据创建 Spark DStream，并利用相关操作进行数据分析，基于网络中实时数据创建 Spark DStream，并结合窗口等概念和相关操作进行数据分析，基于无状态模式处理 HDFS 上的文本数据，基于 stateful 与 window 模式处理网络实时数据。

◆ **4、Spark 交互式数据查询框架 Spark SQL**

【领会】

Spark SQL 的发展历程

Spark SQL 的性能

Spark SQL、Hive、Shark 之间的联系

Spark SQL 的应用场景

hive/console 的安装过程与基本原理

【熟知】

基于 Hadoop 搭建 Spark SQL 的测试环境

掌握 LogicalPlan、SqlParser、Analyzer、Optimizer 等组件

SchemaRDD 的基本概念与相关操作

不同数据源的运行计划

不同查询的运行计划

查询优化策略

【应用】

HiveContext 与 SQLContext 的基础应用，利用 Spark SQL 对 JSON 文件、Parquet 文件以及 Hive 上的数据进行交互式查询。

◆ **5、Spark 机器学习算法库 Spark MLlib 基本使用方法**

【领会】

Spark MLlib 的基本框架与原理

Spark MLlib 中 ML 库与 MLlib 库区别

【熟知】

Spark MLlib 中矩阵向量运算方法

Spark MLlib 中常用统计计算方法

【应用】

能够利用 ML Pipelines 构建机器学习流

能够利用 TF-IDF、Word2Vec、CountVectorizer 等进行特征抽取、转化和选择

能够利用 ML 进行机器学习模型建模，至少掌握以下常用模型建模方法，包括决策树、逻辑回归、KMeans 聚类、GMM 高斯混合模型聚类、协同过滤、随机森林、SVM 等模型

能够利用 CrossValidator（交叉验证）和 TrainValidationSplit（训练验证分割）进行模型评估与参数调优

◆ **6、Spark 图计算框架 GrapX**

【领会】

Spark GraphX 简介

Spark GraphX、GraphLab、Pregel 的联系与区别

Spark GraphX 中表视图与图视图的两种数据的转换

图论基本概念

【熟知】

Spark GraphX 中数据的主要表示形式

图的存储模型

Spark GraphX 提供的切分策略

图的构建操作

图的属性操作

图的结构操作

【应用】

Spark GraphX 源码分析与调试

基于 Pregel 的 API 实现图的 PageRank 和最短路径算法

PART 6 大数据分析之数据可视化方法

◆ **1、数据可视化入门基础**

【领会】

数据可视化应用场景

常用数据可视化工具

常用可视化图形介绍及其基本作用，常用可视化图形包括条形图、柱状图、饼图、折线

图、雷达图等

◆ **2、Python 数据可视化**

【领会】

Python 可视化发展近况及其优势

【熟知】

Python 数据可视化常用包的安装与更新，包括 Matplotlib、Seaborn 等

利用 Matplotlib、Seaborne 绘制常用可视化图形

◆ **3、Python 高级可视化方法**

【领会】

Echarts 基本情况与主要应用背景

【熟知】

Pyecharts 的安装与更新

利用 Pyecharts 绘制常见可视化图形

PART 6 大数据分析实战

◆ **1、利用 HDFS Shell 操作 HDFS 文件系统**

【熟知】

HDFS 常用命令，包括创建文件目录命令、文件传输命令、文件修改及删除命令等

◆ **2、利用 Hive SQL 进行数据清洗**

【熟知】

熟悉 Hive SQL 基本语法，并能在数据预处理中灵活利用 Hive 工具，通过创建 Hive 表，利用 Hive SQL 进行数据查询与数据清洗

◆ **3、利用 Sqoop 进行数据传输**

【熟知】

能够灵活使用 Sqoop shell 命令进行文件在 Hadoop 中与 MySQL 数据库中的转化操作，以达到文件传输要求

◆ **4、利用 Spark SQL 进行数据读取**

【熟知】

能够灵活应用 Spark SQL 读取文件，并能够将其他数据类型按要求转化为 DataFrame，以方便后续机器学习建模工作

◆ **5、Spark MLlib 进行机器学习建模**

【熟知】

能够根据分析要求，灵活调用 MLlib 中的相关算法进行分析，并能进一步构建机器学习流，能够利用调参工具对模型进行调优，能够利用模型评估指标最终建模结果进行评估。

◆ **6、利用 Python 进行建模结果数据可视化**

最终建模完成后，结合实际业务场景和演示需求，将建模结果导入本地，并利用 Pyhton 工具，合理选择对应第三方库，对建模结果进行数据可视化演示。

六、推荐学习书目

说明：推荐学习书目中考生可根据自身需求选择性学习。参考书目不需全部学完，根据考纲知识点进行针对性学习即可。

[1] Jake VanderPlas. Python 数据科学手册[M]. 人民邮电出版社，2018.（必读）

[2] Tom White. Hadoop 权威指南（第三版）[M]. 清华大学出版社， 2015.（必读）

[3] 王雨竹，高飞. MySQL 入门经典[M]. 机械工业出版社，2013.（必读）

[4] Pang-Ning Tan 等. 数据挖掘导论[M]. 人民邮电出版社，2011.（必读）

[5] 林子雨等. Spark 编程基础[M]. 人民邮电出版社，2018.（必读）

[6] Hold Karau 等. Spark 快速大数据分析[M]. 人民邮电出版社，2015. （必读）

[7] Sandy Ryza 等. Spark 高级数据分析[M]. 人民邮电出版社，2015. （选读）

CDA Institute

经管之家 CDA 数据分析研究院