CDA LEVEL Ⅱ 大数据分析师考试大纲及解析

CERTIFIED DATA ANALYST LEVEL ⅡEXAMINATION OUTLINE

CDA 考试大纲是 CDA 命题组基于 CDA 数据分析师等级认证标准而设定的一套科学、详细、系统的考试纲要。考纲规定并明确了 CDA 数据分析师资格考试的具体范围、内容和知识点，考生可按照 CDA 考试大纲进行相关知识的复习。

CDA LEVEL Ⅱ 大数据分析师考试大纲

基础理论(占比 15%)

a. 数据分析基础（5%）
b. 描述性统计分析（5%）
c. Linux& Ubuntu 基础（2%）
d. hbase 理论及实战（3%）

Hadoop 理论(占比 15%)

a. hadoop 安装配置及运行机制解析（3%）
b. Hadoop 分布式文件系统（2%）
c. MapReduce 理论及实战（8%）
d. hadoop 生态环境简介（2%）

大数据分析之数据库理论及工具　占比(15%)

a. 数据库理论（2%）
b. mysql 理论及实战（3%）
c. sqoop 安装及应用（3%）
d. hive 安装部署及应用（5%）
e. Tableau 的具体功能及应用（2%）

大数据分析之数据挖掘理论　占比(10%)

a. 数据挖掘的基本思想（2%）
b. 数据挖掘之聚类算法（3%）
c. 数据挖掘之分类算法（2%）
d. 数据挖掘之主题推荐（3%）

大数据分析之 Mahout 工具及实战　占比(10%)

a. mahout 之聚类及实战（2%）
b. mahout 之分类及实战（3%）
c. mahout 之主题推荐及实战（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 与内核(8%)
c. 实时数据流处理框架—Spark Streaming(3%)
d. 交互式数据查询框架—Spark SQL (5%)
e. 数据挖掘框架—Spark MLib (占比 12%)
f. 图计算框架—Spark GrapX (占比 5%)

CDA　LEVEL Ⅱ大数据分析师考试大纲解析

根据 CDA 数据分析师认证考试大纲，经管之家 CDA 数据分析研究院给出了详细解析，以“领会”，“熟知”，“应用”三个不同的级别将每一个知识点进行分解，建议考生应该按照不同的知识掌握程度有目的性的进行复习。

1. 领会：要求应考者能够记忆规定的有关知识点的主要内容，并能够了解规定的有关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，并能根据考核的不同要求，做出正确的解释、说明和阐述。
2. 熟知：要求应考者必须熟悉的理论知识，并能够正确理解和记忆相关的理论方法，根据考核的不同要求，做出逻辑严密的解释、说明和阐述。
3. 应用：要求应考者必须掌握知识点的主要内容，并能够结合工具进行商业应用，根据考核的具体要求，做出问题的具体实施流程和策略。

PART 1

基础理论部分

- 数据分析基础

1．领会：数据分析和数据挖掘的概念，数据描述性统计分析，抽样估计和假设检验的基础知识，方差分析和回归分析的基础知识。

2．熟知：明确数据分析目标的意义，数据分析方法与数据挖掘方法的区别和联系；明确数据分析中不同人员的角色与职责；衡量数据集中趋势、离中趋势和数据分布的常用指标及计算方法，P 值检验的原理，方差分析和回归分析的应用前提

3．应用：根据不同数据类型选用不同的统计指标来进行数据的集中趋势、离中趋势和数据分布的衡量，方差分析和回归分析的实现。

- 描述性统计分析

1．领会：根据统计学的定义，了解统计学中的几个基本概念（变量、数据、数据类型），了解描述性统计分析的常用指标，了解数据来源的主要渠道，了解常用搜集数据的方法、特点及应用条件。

2．熟知：根据数据的类型，选择不同的统计图表对数据进行描述，使用不同的统计量反映数据的集中趋势、离散程度及数据分布的形态。

3．应用：根据不同数据类型，选用不同的统计指标来进行数据的集中趋势、离中趋势和数据分布形态进行深入分析，展示商业目标。

- Linux 与 ubuntu 基础

1．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。

2．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，进行多个节点间的无密码登陆。

3．应用：安装配置 Linux 操作系统，进行多个节点间的无密码登陆。

- hbase 理论及实战

1．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。

2．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。

3．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 2

hadoop 理论

- hadoop 安装配置及运行机制解析

1．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

2．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。

3．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。

- Hadoop 分布式文件系统

1．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

2．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。

3．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。

- MapReduce 理论及实战

1．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。

2．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。

3．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。

- hadoop 生态环境

1．领会：ZooKeeper、Pig 的基本功能结构。

2．熟知：ZooKeeper、Pig 的安装配置参数及常用命令。

3．应用：ZooKeeper、Pig 的安装、运行。

PART 3

大数据分析之数据库理论及工具

- 数据库基础

1．领会：数据（Data）、数据库（Database，DB）、数据库管理系统（DBMS）、 数据库系统（DBS）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；

2．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； SQL 的数据更新功能；

- mysql 理论及实战

1．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

2．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；

3．应用：My SQL 平台下的 SQL 交互操作

- Hive 数据仓库基础

1 领会：Hive 数据仓库在 Hadoop 生态系统中的地位。

2 熟知： Hive 与 Pig、Hbase 的区别。

3 应用： 使用 Hive 进行频率统计。

- Hive 的基本命令

1 领会：Hive 中的数据库概念、修改数据库。

2 熟知：创建表、管理表、外部表、分区表、删除表。

3 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。

- Hive 中检索数据

1 领会：hive 中的命令语句是类 SQL 语句。

2 熟知：SELECT…FROM 语句。

3 应用： 使用列值进行计算、算数运算符、使用函数、列别名、嵌套 SELECT 据、WHERE 语句、group by 语句、集合运算、多表链接、内链接、外链接、笛卡尔积链接、order by 语句、抽样查询、视图。

- Sqoop 基础

1 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。

2 熟知：Sqoop 链接数据库需要 JDBC 的支持。

3 应用： 安装 Sqoop、从 Hadoop HDFS 向 mysql 导入数据、从 mysql 向 Hadoop HDFS 导入数据。

- Tableau 的具体功能

1. 熟知：Tableau 的具体功能，主要包括：数据源连接，连接数据库，数据提取、数据过滤等功能；重命名，隐藏、编辑属性、计算字段、组等数据字段的处理功能；部件、图表类型介绍、多个度量、筛选、排序、集、缺失值、参考线等图表制作功能；

2. 应用：使用 Tableau 进行数据的发布和共享，使用仪表盘建立动作，使用预测、趋势线、参数、表计算等功能进行高级分析

- Tableau 大数据分析及展示

1. 熟知：tableau 连接数据库的方式

2. 应用：使用 tableau 连接 mysql 数据，使用 tableau 连接 hive 数据库的方式，将数据库中的数据展示在 tableau 中，并用于生成大数据分析报告。

PART 4

大数据分析之数据挖掘理论

- 数据挖掘概述

1．领会：数据挖掘的基本思想，数据挖掘的概念

2．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。

- 数据挖掘之聚类算法

1．领会：聚类算法概述，常用的聚类算法

2．熟知：类与类之间的距离，点与点之间的距离，聚类的有效性函数，层次聚类、快速聚类、kmeans 聚类、canopy 聚类等算法的原理和思想

3．应用：能使用数据挖掘工具 R 软件或 SPSS 软件使用常用聚类算法进行数据分析

- 数据挖掘之分类算法

1. 领会：分类算法概述，常用的分类算法，分类中的训练样本、测试样本、特征变量、目标变量等常用术语

2. 熟知：AUC、TPR、TNR 分类等算法模型性能评估指标，ROC 曲线，贝叶斯分类、决策树分类、随机森林等等算法的原理和思想

3. 应用：能使用数据挖掘工具 R 软件或 SPSS 软件使用常用分类算法进行数据分析

- 数据挖掘之主题推荐算法

1. 领会：主题推荐算法概述，常用的主题推荐算法

2. 熟知：欧几里德距离、皮尔逊相关系数、余弦相似性等计算物品和内容相似性的方法，TF-IDF 统计方法，基于物品、用户的推荐算法、ALS-WR 算法原理和思想

3. 应用：能使用数据挖掘工具 R 软件或 SPSS 软件使用常用主题算法进行数据分析

PART 5

大数据分析之 mahout 工具及实战

- mahout 之聚类及实战

1. 熟知：mahout 常用的聚类算法命令及各命令的参数，各个参数的使用场景

2. 应用：使用 mahout 大数据分析工具进行 kmeans、canopy 算法聚类，聚类算法结果分析

大数据分析工具之 Mahout

1．领会：kmeans、canopy 算法、朴素贝叶斯算法、logstic 算法、随机森林算法、基于物品、用户的推荐算法、ALS-WR 算法 mapreduce 实现原理及过程。

2．熟知：kmeans、canopy 算法、朴素贝叶斯算法、logstic 算法、随机森林算法、基于物品、用户的推荐算法、ALS-WR 算法的实现过程、结果查看命令，各种算法在 mahout 中执行的命令及参数调整

3．应用：使用 mahout 大数据分析工具进行聚类、分类和主题推荐。

- mahout 之分类及实战

1. 熟知：mahout 常用的分类算法命令及各命令的参数，各个参数的使用场景

2. 应用：使用 mahout 大数据分析工具进行朴素贝叶斯算法、logstic 算法、随机森林算法分类，分类算法结果分析

- mahout 之主题推荐及实战

1. 熟知：mahout 常用的主题推荐命令及各命令的参数，各个参数的使用场景

2. 应用：使用 mahout 大数据分析工具进行基于物品、用户的推荐算法、ALS-WR 算法进行主题推荐，推荐结果的实际应用分析

PART 6

大数据分析之 Spark 工具及实战

- Spark 基础理论

1. 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。

2. 熟知：Spark 生态系统中的四大核心组件，Spark 与 MapReduce 的对比与分析，二者所适用的应用场景，Spark 的多种运行模式

3. 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。

- Spark RDD 与内核

1. 领会：Spark RDD 基本概念，Spark API，Spark 任务调度策略

2. 熟知：Spark RDD 中的转换操作、执行操作、持久化操作，RDD 之间的宽依赖关系与窄依赖关系，Spark 基于 DAG 图实现的容错机制。

3. 应用：基于 Spark API 编写 WordCount 程序，并在 WordCount 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。

- 实时数据流处理框架—Spark Streaming

1. 领会：Spark Streaming 应用场景，Spark Streaming 基本概念，Spark DStream 的存储级别；

2. 熟知：批处理间隔、离散数据流 Spark DSteam、窗口、滑动间隔、窗口间隔等重要概念，熟练使用 Spark DSteam 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。

3. 应用：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。

- 交互式数据查询框架—Spark SQL

1. 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。

2. 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。

3.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。

- 数据分析框架—Spark MLib

1. 领会： Spark MLib 的基本框架与原理，Spark MLib 目前支持的三种常见数据挖掘问题（分类、聚类和协同过滤）。

2. 熟知：掌握 Spark MLib 中的矩阵向量运算库 jblas，掌握 Spark MLib 中的梯度下降算法。

3. 应用：LinearRegressionWithSGD 源码分析与调试，Spark MLib 中协同过滤算法的源码分析与调试，Spark MLib 中 K-Means 算法源码中的相关参数，K-Means 源码分析与调试，从源码角度分析并掌握 K-Means 的重要步骤。

- 图计算框架—Spark GrapX

1. 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。

2. 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。

3. 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。

参考教材

Jonathan R.Owens JonLentz Brian Femiano 著，傅杰 赵磊 卢学裕译，《hadoop 实战手册》，人民邮电出版社，2014 年 3 月 1 版
王雨竹，高飞著，《MySQL 入门经典》，机械工业出版社，2013 年 4 月 1 版
曹正凤编著，《从零进阶!-数据分析的统计基础》，电子工业出版社，2016 版
sean Owen Robin Anil Ted Dunning Ellen Friedman 著，王斌 韩冀中 万吉译，《Mahout 实战》，人民邮电出版社，2014 年 3 月 1 版
夏俊鸾等著，《Spark 大数据处理技术》，电子工业出版社，2015 年 1 月 1 版