

# CDA LEVEL II 考试大纲

## CERTIFIED DATA ANALYST LEVEL II EXAMINATION 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 数据分析师认证考试的具体范围、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 大数据分析师考试大纲

#### 基础理论(占比 20%)

- a. 数据分析基础 (10%)
- b. JAVA 基础 (5%)
- c. Linux 基础 (2%)
- d. Ubuntu 基础 (3%)

#### Hadoop 理论(占比 20%)

- a. hadoop 安装配置及运行机制解析 (2%)
- b. Hadoop 分布式文件系统 (3%)
- c. MapReduce 理论及实战 (5%)
- d. hbase 理论及实战 (8%)
- e. hadoop 生态环境简介 (2%)

#### 大数据分析之数据挖掘理论(占比 15%)

- a. 数据挖掘的基本思想 (2%)
- b. 数据挖掘之聚类算法 (5%)
- c. 数据挖掘之分类算法 (5%)
- 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 15%)

- a. mahout 之聚类及实战 (5%)
- b. mahout 之分类及实战 (5%)
- 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 30%)

- a. Spark 基础理论(占比 2%)
- b. Spark RDD 与内核(占比 3%)
- c. 实时数据流处理框架—Spark Streaming(占比 2%)
- d. 交互式数据查询框架—Spark SQL(占比 3%)
- e. 数据挖掘框架—Spark MLib (占比 15%)
- f. 图计算框架—Spark GrapX (占比 5%)

# CDA 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。

➤ Java 基础

1. 领会：Eclipse 的编程入门，面向对象的思想基本介绍，类、对象、接口、封装、继承，Java 的集合类——数组、Set、List、Map，数据库基础知识及 SQL 语法。

2. 熟知：JDK 的安装配置，Java 基本知识、数据类型以及基本语法，描述 Java 的文件操作、包的概念及如何打包。

3. 应用：可以使用 Eclipse 工具进行简单的 JAVA 应用程序开发。

➤ 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。

## 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。

➤ 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。

➤ hadoop 生态环境

1. 领会：ZooKeeper、Pig、Hive、Sqoop 的基本功能结构。

2. 熟知：ZooKeeper、Pig、Hive、Sqoop 的安装配置参数，Hive、Sqoop 的原理及常用命令。

3. 应用：ZooKeeper、Pig、Hive、Sqoop 的安装、运行。

## PART 3

### 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 4

### 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 5

### 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 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。
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 MLlib

1. 领会：Spark MLlib 的基本框架与原理，Spark MLlib 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 MLlib 中的矩阵向量运算库 jblas，掌握 Spark MLlib 中的梯度下降算法。
3. 应用：LinearRegressionWithSGD 源码分析与调试，Spark MLlib 中协同过滤算法的源码分析与调试，Spark MLlib 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 GraphX

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。

## 参考资料

张孝祥编著，《java 就业培训教程》，清华大学出版社，2003 年 9 月

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，2015 年 8 月 1 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

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

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

CDA 考试报名链接: [exam.cda.cn](http://exam.cda.cn)

经管之家 CDA 数据分析研究院

2016.2