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 大数据分析师考试大纲

基础理论(占比 30%)

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

Hadoop 理论(占比 30%)

a. hadoop 安装配置及运行机制解析（5%）
b. Hadoop 分布式文件系统（10%）
c. MapReduce 理论及实战（5%）
d. hbase 理论及实战（5%）
e. hadoop 生态环境简介（5%）

大数据分析理论、工具及实战　占比(40%)

a. 大数据分析之数据挖掘理论（10%）
b. 大数据分析工具之 Mahout （10%）
c. 大数据分析分析工具之刀 Spark （20%）

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 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

大数据分析理论、工具及实战

- 数据挖掘理论基础

1．领会：数据挖掘的基本思想，聚类、分类、关联规则及预测概述，kmeans、canopy 算法思想，朴素贝叶斯算法、logstic 算法、随机森林算法思想，基于物品、用户的推荐算法、ALS-WR 算法思想。

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 算法原理，点和点之间的距离、类和类之间的距离的计算，分类算法中的混淆矩阵、ROC 曲线、AUC 值的含义及用法。

- 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。

- 大数据分析工具之 Spark

1．领会：Spark 的基本功能结构、基本原理，Spark 内存式读写机制，Spark SQL 原理及数据整合应用。

2．熟知：Spark 数据建模流程（logistics 回归，决策树，朴素贝叶斯方法），Spark 推荐应用（ALS 方法，FP-growth 方法），Spark GraphX 图计算方法应用。

3．应用：Spark 单机和集群的安装、运行，使用 Spark 进行聚类、分类和主题推荐的大数据分析。

参考教材：

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：http://cda.pinggu.org/online_registration.html

经管之家 CDA 数据分析研究院
2015.9