

# 大数据开发工程师（基础级）考试大纲

## Big Data Development Engineer Level I Examination Outline

BDDE 考试大纲是 CBDA 国际认证培训中心基于大数据开发工程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 BDDE 大数据开发工程师认证考试的具体范围、内容和知识点，考生可按照 BDDE 考试大纲进行相关知识的复习。

### 大数据基础（占比 10%）

1. 大数据概念（占比 2%）
2. 大数据特点。（占比 3%）
3. 大数据处理流程（占比 5%）

### 大数据 Hadoop 技术（占比 70%）

1. Hadoop 基础（占比 5%）
2. Hadoop 安装配置（占比 10%）
3. 分布式文件系统 HDFS（占比 15%）
4. 分布式计算框架 MapReduce 编程（占比 40%）

### 大数据数据仓库 Hive（占比 20%）

1. Hive 概述（占比 2%）
2. Hive 客户端访问（占比 2%）
3. Hive 数据类型（占比 1%）
4. Hive DDL、DML、Select 语法（占比 15%）

# 大数据开发工程师（基础级）考试大纲解析

## Big Data Development Engineer Level I Examination Note

根据 BDDE 大数据开发工程师认证考试大纲，泰迪智能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: 大数据基础

### ➤ 大数据概念

1. 领会：大数据基本概念、起源、目标及发展历程。
2. 熟知：明确大数据的基本概念、发展历程，针对大数据的起源及要解决的问题（即目标）有一定认知。

### ➤ 大数据特点

1. 领会：了解大数据特点、对比传统小数据不同之处、应用场景及行业。
2. 熟知：明确大数据 3V、4V 特点、针对不同的特点有明确的认知，对大数据的应用场景及行业有比较清晰的认识。

### ➤ 大数据处理流程

1. 领会：大数据存储、大数据传输，大数据预处理，建模预测及评估调优。
2. 熟知：明确大数据存储、传输的通用技术、大数据预处理常用过程：数据抽取、属性选择数据质量分析、数据分布分析、数据清洗、属性变换、数据规约，根据实际业务场景对不同数据进行不同的建模及调优。

## Part 2: 大数据 Hadoop 技术

### ➤ Hadoop 基础

1. 领会：了解 Hadoop 基本概念、发展历程。
2. 熟知：明确 Hadoop 的特点、架构及组成。

### ➤ Hadoop 安装配置

1. 领会：能进行基本的 Hadoop 完全分布式集群安装配置、会进行基本的集群启动关闭机监控。
2. 熟知：明确 Hadoop 各个配置文件中参数的意义。

### ➤ 分布式文件系统 HDFS

1. 领会：了解 HDFS 架构及组件构成。
2. 熟知：熟练使用 `hdfs dfs shell`、`hdfs dfsadmin shell`。
3. 应用：针对不同的需求，使用 HDFS Shell 完成对应功能。

### ➤ 分布式计算框架 MapReduce 编程

1. 领会：了解 MapReduce 流程：输入、分区、Mapper、Shuffle 和 Sort、Reducer、输出。
2. 熟知：能进行基本的 Mapper、Reducer、Driver 编程。
3. 应用：能根据不同的需求，编写对应的 Mapper 及 Reducer 逻辑来实现对应的功能。

## Part 3: 大数据数据仓库 Hive 技术

### ➤ Hive 概述

1. 领会：了解大数据仓库 Hive 的基本架构、原理和适用场景。

### ➤ Hive 客户端访问

1. 领会：了解访问大数据仓库 Hive 的三种方式。
2. 熟知：能根据不同的场景选择不同的大数据仓库访问访问。

### ➤ Hive 数据类型

1. 领会：了解 Hive 的基本数据类型：INT、DOUBLE、STRING 等。

➤ Hive DDL、DML、Select 语法

1. 领会：了解 Hive 的数据定义预研基本语法、数据操作预研基本语法、查询基本语法。
2. 熟知：能进行 Hive 建表、修改表、数据导入导出、基本查询操作。
3. 应用：能根据不同需求，使用不同的命令来建立符合应用场景的 Hive 表或导入导出数据。

## 参考资料

- 大数据时代, [英] 维克托·迈尔-舍恩伯格, [英] 肯尼思·库克耶 著; 盛杨燕, 周涛 译, 浙江人民出版社
- Hadoop 权威指南, [美] 汤姆·怀特 (Tom White) 著; 王海, 华东, 刘喻, 吕粤海 译, 清华大学出版社
- Hadoop 大数据分析与挖掘实战, 张良均 樊哲 赵云龙 李成华 著, 机械工业出版社
- Hadoop 与大数据挖掘, 张良均 樊哲 位文超 刘名军等 著, 机械工业出版社
- Hadoop 大数据开发基础, 余明辉 张良均 著, 人民邮电出版社
- Hive 编程指南, [美] Edward Capriolo, Dean Wampler, Jason Rutherglen 著; 曹坤 译, 人民邮电出版社

# 大数据开发工程师（专业级）考试大纲

## Big Data Development Engineer Level II Examination Outline

BDDE 考试大纲是 CBDA 国际认证培训中心基于大数据开发工程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 BDDE 大数据开发工程师认证考试的具体范围、内容和知识点，考生可按照 BDDE 考试大纲进行相关知识的复习。

### 大数据 Hadoop 技术（占比 30%）

1. Hadoop 集群调优（占比 5%）
2. 分布式文件系统 HDFS Java API 操作（占比 5%）
3. 分布式计算框架 MapReduce 编程（占比 20%）

### 大数据数据仓库 Hive（占比 15%）

1. Hive 复杂 DDL、DML、自定义函数语法（占比 10%）
2. Hive 调优（占比 5%）

### 大数据内存计算 Spark（占比 40%）

1. Spark 基础（占比 3%）
2. Spark 集群（占比 5%）
3. Spark RDD 及编程（占比 20%）
4. Spark 生态圈（占比 15%）

### 大数据数据库 HBase（占比 10%）

1. HBase 基础（占比 3%）
2. HBase 数据模型（占比 2%）
3. MapReduce 操作 HBase 数据（占比 5%）

## 大数据 workflow Oozie (占比 5%)

1. Oozie 基础 (占比 1%)
2. Oozie 工作流 (占比 4%)

# 大数据开发工程师（专业级）考试大纲解析

## Big Data Development Engineer Level II Examination Note

根据 BDDE 大数据开发工程师认证考试大纲，泰迪智能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: 大数据 Hadoop 技术

### ➤ Hadoop 集群调优

1. 领会：Hadoop 配置文件参数，。
2. 熟知：能根据不同任务情况，来调整集群，使集群达到较优性能。

### ➤ 分布式文件系统 HDFS Java API 操作

1. 领会：明确 HDFS Java API 的各种调用情况。
2. 熟知：能根据具体需求编写调用接口，对分布式文件系统 HDFS 进行各种操作。

### ➤ 分布式计算框架 MapReduce 编程

1. 领会：熟知 MapReduce 流程：输入、分区、Mapper、Shuffle 和 Sort、Reducer、输出。
2. 熟知：能进行高级的 Mapper、Reducer、Driver、Partitioner、Combiner、自定义键值对、自定义 FileInputFormat 和 FileOutputFormat 编程。
3. 应用：能根据不同的需求，编写较高性能的 MapReduce 程序来实现对应的功能，同时掌握二次排序及全局排序的内容。

## Part 2: 大数据数据仓库 Hive 技术

- Hive 复杂 DDL、DML、自定义函数语法
  1. 领会：对大数据仓库 Hive 复杂数据定义语言、数据操作预研有一定了解、能够进行常用内置函数操作。
  2. 熟知：对表分区、桶表操作、语句执行计划有一定了解，能够进行自定义函数操作。
- Hive 调优
  1. 领会：能进行深入的 Hive 表优化、SQL 优化。
  2. 熟知：优化 Hive Job，在 Map/Shuffle/Reduce 阶段。

## Part 3: 大数据内存计算 Spark

- Spark 基础
  1. 领会：对 Spark 基本概念、基本原理、特点有一定了解。
  2. 熟知：知道 Spark 和 Hadoop 的特点对比，了解 Spark 生态圈及其应用场景
- Spark 集群
  1. 领会：会配置 Spark 集群，了解参数的意义，能进行集群的启动与关闭，会查看各个组件对应的监控信息。
  2. 熟知：能根据实际应用来进行集群参数配置。
- Spark RDD 及编程
  1. 领会：弹性分布式数据集 RDD 基本概念与原理，其数据存储策略及其使用场景。
  2. 熟知：能进行 Transformation 和 Action 编程。
  3. 应用：能够根据具体需求来编写对应的 Spark 程序。
- Spark 生态圈
  1. 领会：Spark SQL、Spark Streaming、Spark MLlib、Spark GraphX 架构原理及适用场景。
  2. 熟知：能使用 Spark SQL 、 Spark MLlib 进行系统架构。



## Part 4: 大数据数据库 HBase

- HBase 基础
  1. 领会：了解大数据数据库 HBase 基本架构及原理、特点。
- HBase 数据模型
  1. 领会：了解 HBase 的三维数据模型，能对不同维度进行分析。
- MapReduce 操作 HBase 数据
  1. 领会：能使用 MapReduce 对 HBase 数据进行导入导出。

## Part 3: 大数据工作流 Oozie

- Oozie 基础
  1. 领会：了解大数据工作流 Oozie 基本概念、架构。
- Oozie 工作流
  1. 领会：会进行 Oozie 的 MapReduce、Spark、Hive 工作流配置及运行。

## 参考资料

- Hadoop 权威指南, [美] 汤姆·怀特 (Tom White) 著; 王海, 华东, 刘喻, 吕粤海译, 清华大学出版社
- Hadoop 大数据分析 with 挖掘实战, 张良均, 樊哲, 赵云龙, 李成华 著, 机械工业出版社
- Hadoop 与大数据挖掘, 张良均, 樊哲, 位文超, 刘名军等 著, 机械工业出版社
- Hive 编程指南, [美] Edward Capriolo, Dean Wampler, Jason Rutherglen 著; 曹坤 译, 人民邮电出版社
- Spark 大数据技术与应用, 肖芳 张良均等 著, 人民邮电出版社
- Spark 快速大数据分析, [美] 卡劳 (Holden Karau), [美] 肯维尼斯科 (Andy Konwinski), [美] 温德尔 (Patrick Wendell), [加] 扎哈里亚 (Matei Zaharia) 著; 王道远 译, 人民邮电出版社
- HBase 权威指南, [美] Lars George 著; 代志远, 刘佳, 蒋杰 译, 人民邮电出版社