



湖南石油化工职业技术学院  
Hunan Petrochemical Vocational Technology College

# 课程标准

课程名称： 行业数据综合实训

课程代码： 31091640

适用专业： 大数据技术与应用

制订时间： 2021年3月

湖南石油化工职业技术学院

## 目 录

1 课程概述.....	1
1.1 课程的性质.....	1
1.2、课程定位.....	1
1.3 课程设计思路.....	1
2. 课程基本目标.....	1
1、知识目标.....	1
2、职业技能目标.....	2
3、职业素质养成目标.....	2
3、课程教学内容及学时安排.....	2
3.1 课程主要内容说明.....	2
3.2 课程组织安排说明.....	3
3.3 课程教学内容.....	3
4 实施建议.....	5
4.1 教学组织建议.....	5
4.2 教学评价.....	6
4.3 教材选用.....	6
4.4 课程主讲教师和教学团队要求说明.....	6
4.5 课程教学环境和条件要求.....	6
4.6 教学资源的开发与利用.....	7

# 《行业数据综合实训》课程标准

课程名称：行业数据综合实训

课程代码：31091640

总学时数：72 学时（理论课学时数：36； 实践课学时数：36）

适用专业：大数据技术与应用专业

## 1 课程概述

### 1.1 课程的性质

《行业数据综合实训》是大数据技术与应用专业的一门专业核心课（技术技能课），旨在对学生的程序设计思想和技能进行强化，培养学生利用主流 scrapy 框架进行爬虫项目的设计和开发的能力。

### 1.2、课程定位

本课程以 Hive 数据仓管岗位的基本要求为指导，依据该岗位真实业务内容与流程选取课程内容、构建学习单元，将目前 Hive 必备功能组件如网页数据下载、数据分析、数据存储、网页地柜爬取等技术作为项目中的系列任务。课程内容编排符合循序渐进的认知规律，培养学生的 Hive 实际应用能力。

先导课程：Python、Hadoop 大数据、Linux 操作系统。

### 1.3 课程设计思路

课程设计围绕大数据开发的基础和方法，考虑到实际企业生产环境中的应用场景，课程中采用实际案例，从案例出发、启动、实施、验证各环节剖析讲解，与真实环境同步，本课程通过大量的案例，引导学生思考问题、分析问题，学生在理解知识点的过程中遇到困难时，可以继续往后学习。可以通过对各学习情景中学习目标、主要内容、授课方式、师生要求等各项内容的描述，来规范课程所要求的内容。真正做到在项目实施过程中各个环节的考察和评价，来完成对课程的评鉴与考核。

## 2. 课程基本目标

### 1、知识目标

- （1）掌握 Hive 设计理念
- （2）掌握 Hive 数据添加

- (3) 掌握 Hive 数据删除
- (4) 掌握 Hive 数据修改
- (5) 熟练掌握 HiveQL 查询
- (6) 熟练 Hive 装载数据
- (7) 熟练掌握 HiveQL: 视图
- (8) 熟练掌握 HiveQL: 索引
- (9) 熟练掌握 Hive 调优
- (10) 熟练掌握 Hive 函数

## 2、职业技能目标

- (1) 能够熟练 Hive 数据添加、删除、修改
- (2) 能够熟练 HiveQL 查询
- (3) 能够进行 Hive 数据装载
- (4) 能够进行 Hive 调优
- (5) 能够完好规范编写技术文档
- (6) 沟通能力强, 能够与小组其他人通力合作
- (7) 能够独立分析解决技术问题
- (8) 自学能力强, 能够快速准确地查找参考资料

## 3、职业素质养成目标

- (1) 培养学生良好的职业道德
- (2) 培养按时、守时的软件交付观念
- (3) 培养学生良好的自我表现、与人沟通的能力
- (4) 培养学生的团队精神和服务意识
- (5) 培养学生的自主学习能力
- (6) 培养学生勇于创新、敬业乐观的工作作风

## 3、课程教学内容及学时安排

### 3.1 课程主要内容说明

本课程是大数据专业课程。Hive 是 Hadoop 生态系统中必不可少的一个工具, 它提供了一种 SQL (结构化查询语言) 方言, 可以查询存储在 Hadoop 分布式文件系统(HDFS)中的数据或其他和 Hadoop 集成的文件系统, 如 MapR-FS、Amazon

的 s3 和像 HBase (Hadoop 数据库) 和 cassandra 这样的数据库中的数据。

大多数数据仓库应用程序都是使用关系数据库进行实现的,并使用 SQL 作为查询语言。Hive 降低了将这些应用程序转移到 Hadoop 系统上的难度。凡是会使用 SQL 语言的开发人员都可以很轻松地学习并使用 Hive。如果没有 Hive,那么这些用户就必须学习新的语言和工具,然后才能应用到生产环境中。另外,相比其他工具, Hive 更便于开发人员将基于 SQL 的应用程序转移到 Hadoop 中。如果没有 Hive,那么开发者将面临一个艰巨的挑战,如何将他们的 SQL 应用程序移植到 Hadoop 上。不过, Hive 和其他基于 SQL 的环境还是有一些差异的。如今,可供 Hive 用户和 Hadoop 开发者使用的文档并不多,所以我们决定撰写这本书来填补这个缺口。我们将对 Hive 进行全面详实的介绍,主要适用于 SQL 专家,如数据库设计人员和业务分析师。我们也谈到了深入的技术细节,可以帮助 Hadoop 开发人员对 Hive 进行调优和定制。

### 3.2 课程组织安排说明

- (1) 采用机房教学,教师讲解和学生操作同步进行
- (2) 递进式教学,按知识和技能分阶段进行知识讲解、项目实践和技能考核
- (3) 理论联系实际,理论讲授和知识运用交叉进行,采用企业化的教学案例

### 3.3 课程教学内容

序号	学习单元	知识、技能、素质、要求	学习成果	建议学时
1	Hive 概述	<ul style="list-style-type: none"><li>● 能够初步了解 Hive 的概念,了解 Hive 的历史、发展、功能等</li><li>● 具备信息收集和信息处理能力</li><li>● 具备自学能力,能适应行业的不断变革发展</li><li>● 具备一定的设计素养</li></ul>	Hive 基础知识	4
2	Hive 安装	<ul style="list-style-type: none"><li>● 安装 Hadoop</li></ul>	能够安装并启动 Hive	8

		<ul style="list-style-type: none"> <li>● 安装Hive</li> <li>● 启动Hive</li> </ul>		
3	Hive 基础操作	<ul style="list-style-type: none"> <li>● 使用 JDBC 连接元数据</li> <li>● Hive 命令</li> <li>● 命令行界面</li> <li>● CLI 选项</li> <li>● 从文件中执行 Hive 查询</li> </ul>	能够使用 Hive 对数据进行基础操作	6
4	数据类型和文件格式	<ul style="list-style-type: none"> <li>●基本数据类型</li> <li>●集合数据类型</li> <li>●文本文件数据编码</li> <li>●读时模式</li> </ul>	能够掌握 Hive 数据类型	6
5	HiveQL: 数据定义	<ul style="list-style-type: none"> <li>●修改数据库</li> <li>●创建表</li> <li>●管理表</li> <li>●分区表</li> <li>●外部分区表</li> </ul>	能够使用 Hive 数据定义	6
6	数据操作	<ul style="list-style-type: none"> <li>●向管理表中装载数据</li> <li>●通过查询语句向表中插入数据</li> <li>●单个查询语句中创建表并加载数据</li> <li>●导出数据</li> </ul>	能够进行数据操作	6
7	HiveQL: 查询	<ul style="list-style-type: none"> <li>●SELECT...FROM 语句</li> <li>●使用正则表达式来指定列</li> <li>●使用列值进行计算</li> <li>●算术运算符</li> </ul>	能够使用 HiveQL 查询	6
8	Hive 查询语法	<ul style="list-style-type: none"> <li>●CASE...WHEN...THEN 句式</li> <li>●什么情况下 Hive 可以避免进行 MapReduce</li> <li>●WHERE 语句</li> <li>●GROUP BY 语句</li> <li>●JOIN 语句</li> </ul>	能够使用使用 Hive 查询	8

		●ORDER BY 和 SORT BY		
9	HiveQL: 视图	<ul style="list-style-type: none"> <li>●使用视图来降低查询复杂度</li> <li>●使用视图来限制基于条件过滤的数据</li> <li>●动态分区中的视图和 map 类型</li> <li>●视图零零碎碎相关的事情</li> </ul>	能够使用 HiveQL: 视图	6
10	HiveQL: 索引	<ul style="list-style-type: none"> <li>●创建索引</li> <li>●重建索引</li> <li>●显示索引</li> <li>●删除索引</li> <li>●实现一个定制化的索引处理器</li> </ul>	能够进行 HiveQL: 索引	4
11	Hive 调优	<ul style="list-style-type: none"> <li>●使用 EXPLAIN</li> <li>●EXPLAIN EXTENDED</li> <li>●JOIN 优化</li> <li>●调整 mapper 和 reducer 个数</li> <li>●动态分区调整</li> <li>●单个 MapReduce 中多个 GROUP BY</li> </ul>	能够进行 Hive 调优	4
12	复习考试			8
合计学时				72

## 4 实施建议

### 4.1 教学组织建议

① 教学方法多样化，教学内容真实化。建议教师在家偶尔过程中通过案例激发学生思考，基于真实的第三方数据和抓取的外部数据来布置任务，驱动教学，从而提高学生的学习积极性，提高学生实操能力。

② 教学手段现代化。利用多媒体、网络平台、信息系统、视频录像等现代化手段，强化实际操作技能的训练，提高课堂教学效率。

③ 教学组织团队化。运营管理实践工作都是以团队形式完成的，教学过程中同样采取分组方式来组织实践教学。每项实践活动都是一个完整的工作过程，因此都可以成立类似于企业的一个工作小组；每个小组由 5~6 人组成，小组工作要按企业化运作，实行组长负责制；并在班级或年级内开展小组竞赛，培养学生的团队协作能力和职业意识，提高学生的管理能力。

## 4.2 教学评价

组成模块	考核项目	考核标准	评价分值
平时成绩	课堂学习	包括出勤、课堂表现及课堂积极回答问题等	15
	课后作业	是否按时、按质、按量完成教师布置的课后练习	15
项目评审	个人项目	学习过程中的两个综合案例	20
期末考试	笔试	由教师评定的笔试成绩	50

## 4.3 教材选用

依据本书中的课程标准选择与编写教材。教材应充分体现工学结合、任务驱动、项目教学的课程设计思想，要体现通用性、实用性、先进性，项目活动载体的选择要科学。优先选用高职教育国家规划教材、省级规划教材、数据科学与大数据技术专业系列规划教材。

## 4.4 课程主讲教师和教学团队要求说明

“行业数据综合实行”课程建议安排在校内实训室上课，实现教学场所与职业场所的一体化，以提高学习效率。建议本课程至少配套一个大数据实训室，实训室内应安装 Python、Hadoop、Linux 等课程所需组件，且具备本课程相对应得数据源资料。

本课程建议采用专任教师和兼职教师共同组成“双师型”课程教学团队。

专任教师应具有大数据，爬虫等相关专业背景。行业兼职教师配备建议具备两年以上爬虫工作经验；具有责任心和一定的教育教学能力。

## 4.5 课程教学环境和条件要求

计算机硬件要求：WindowsXP 或 Windows7 及以上，CPU 主频>2Ghz，内存容量>=2GB

软件要求：Centos6.5、Hadoop2.7.x

其他要求：该课程要求在理论实践一体化教室完成，以实现教、学、做合一，同时要求安装多媒体教学软件，方便下发教学任务和收集学生课堂实践结果；基于工作任务的项目课程最适合开展“教学做”一体化教学，实训室应包括多媒体



教学系统、产品展示、模拟软件，能同时开展讲授、训练和项目教学。

#### 4.6 教学资源开发与利用

本课程应建设实训指导手册、习题集、实训项目库、企业案例库、教学视频、教学动画等教学资源。若有条件，建设微课、MOOC\SPOC 等。可参考卓京教育等平台的数据资源，结合校企合作的实际项目，开发教学资源。