ITSTAR 2017 大数据高薪实战《3+1+1 特训营》系统课程 V2.1

在 DT 时代发展的今天，数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。我们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。

古语云：三分技术，七分数据，得数据者得天下。今天大数据发展已经是国家级战略，所有企业都会直接和间接的应用到大数据的服务和产品。而今天我们面临最大的问题就是大数据的人才和行业需求严重失衡，据数据统计我国大数据人工智能人才缺口达到 1000 万 + ，企业对于这方面的人才已经形成如饥似渴，供不应求的局面。大数据学习已经到了迫在眉睫的时刻。

以下是目前国内完整大数据高端实战实用学习流程体系：

课程分类：大数据高级研发工程师，大数据架构师
课程讲师：赵老师
适合人群：初级零基础入门，中高级深造进修。
应用技术：Hadoop 集群，Storm 实时流式计算，Spark 内存计算架构，项目架构（数据抓取、存储、计算处理、可视化）完整系统
设计项目：京东商城、百度、阿里巴巴
课程形式：在线直播授课+课后录播视频+课后作业、技术解答 + 新知识点拓展

第一阶段：linux + Hadoop 分布式架构计算处理模块

一． linux 基础：

通过 Linux 的学习，系统地掌握企业级 LINUX 操作系统，熟练操作与管理 LINUX 系统及运行在 LINUX 系统上的各种应用与服务；如今大数据平台只能部署在 LINUX 环境下，学完本模块内容对大数据系统的部署、管理、维护与优化打下坚实的操作系统基础；想成为一名优秀的大数据工程师，熟练掌握脚本语言对大数据进行分析处理，不仅可以提高效率，也是必须掌必备技能。

1. 安装虚拟机以及 Linux 操作系统
2. Linux 桌面系统
3. Linux 文件和目录管理
4. VIM 的使用
5. Linux 终端常用命令，局域网工作机制和网络地址配置
6. WinSCP 及 SecureCRT 安装与介绍
7. 常用文件、权限、系统操作命令、ssh 免密登陆
8. 虚拟机相关问题、系统补充命令、安装 jdk 及相关软件

二．Hadoop 离线计算处理模块

Hadoop 是由 Apache 基金会所开发的分布式系统基础架构，一个高度可扩展的存储平台，因为他可以存储和分发横跨数百个并行操作的廉价的服务器数据集群。不同于传统的关系型数据库系统不能扩展处理大量的数据，Hadoop 是能给企业提供涉及成百上千 TB 的数据节点上运行的应用程序。

具有高可靠性、高可拓展性、高效性、高容错性，像我们常用阿里、腾讯、百度（手机百度、百度地图、百度云盘、百度贴吧、百度知道等，还支持了包括凤巢、网盟等百度推广产品，超过百万广告客户；）、今日头条等都用到 Hadoop 技术开发，同时也是我们大数据开发中的核心部分。

1. Hadoop 简介，起源于背景知识

Hadoop 的简介、体系结构、背景环境，Hadoop 的案例分析

什么是大数据，OLTP 与 OLAP

数据仓库 Data Warehouse

Google 的基本思想，实验环境

实验环境

环境	软硬件环境		版本/配置
开发环境	模块设计工具	Office	Office 2016 Professional Edition
	开发工具	Eclipse IDEA	Oxygen 2016.3.16
	文档编写	Office	Office 2016 Professional Edition
	程序运行服务器	Redhat	Redhat 7.4
软件环境	语言工具包	JAVA	1.8.0_144
		SCALA	2.11.11
		Python	2.7.5
	大数据组件	Hadoop	2.7.4
		Spark	2.2.0
		Storm	1.1.1
		Zookeeper	3.4.10
		Kafka	2.11-0.11.0.0
		Habse	1.3.1
		Redis	4.0.0
		Mysql	5.7.19
		Hive	2.3.0
		ELK	5.5.2
		Grafana	4.4.3
		ScrapyRedis	0.6.8
		CDH	5.12.0
测试环境	程序运行服务器	CentOS	Centos 7

2. **Apache Hadoop 的体系结构（重点）**

分布式存储(HDFS）

YARN:分布式计算(MapReduce)

Hbase 的体系结构

3. Hadoop2.X，基本安装，eclipse 配置

Hadoop 安装部署与预备条件

Hadoop 的目录结构

Hadoop 安装部署的三种模式

Hadoop 运行，优先级，定时作业，开机与启动

Hadoop 环境安装和测试

HadoopSSH 免密登陆

4. Hadoop 应用案例分析

互联网应用的架构

日志分析

Hadoop 在淘宝的应用

5. HDFS 文件系统

HDFS 概述，HDFS-Shell 操作，HDFS 运行原理机制

HDFS 的访问，HDFS 创建文件

HDFS 文件保存到云端、删除文件、判断文件是否存在

HDFS 遍历所有节点、HDFS 搜索文件主机，HDFS 重命名和上传，hadoop 作业

HDFS 与 JavaApi 交互，Maven 的简单操作

HDFS 的用户权限管理

HDFS 的配额管理

HDFS 的安全模式

HDFS 的底层原理

6. MapReduce 分布式计算

MapReduce 简介、MapReduce 的 WordCount

Hadoop_count 分析、框架逻辑处理

根据 MapReduce 处理收集流量、根据流量排序、程序分析

温度 MapReduce 序列化、温度排序分组分区、温度程序、程序分析

HadoopMapReduce 分析、HadoopPart 分析、combine 本地节点

combine 程序运行分析、好友关系处理、找出共同好友数，倒排索引

搭建 Hadoop 和 Eclipse 的开发环境

MapReduce 的案例集锦

Shuffle 的过程详解

Hadoop 的集群和 HA

7. **YARN 分布式管理平台，HUE 智能分析管理平台，CDH 大数据平台管理工具，Ooize**

工作流任务调度引擎

YARN，HUE 的简介和背景介绍，MRV1 和 MRV2 的简介
YARN 架构图介绍，架构模型和应用场景
HUE 的搭建和应用
HUE 的架构详解，Job 调度机制
HUE 需要的 rpm 包
HUE 和 HBase 的集成
YRAN 的使用，如何管理 Hadoop
什么是 CM，CDH
CDH,CM 的版本和几种安装方式
利用 CDH 进行开发
CDH 的集群安装和开发
什么是 Ooize
什么是工作流
Ooize 的节点类型
Ooize 的安装和配置,demo 演示
Ooize 的节点，Coordinator 和 Bundle

8. Hive 数据仓库

Hive 基本介绍，结构体系，表，DML，桶表等概念解析
Hive 自定义函数，字符处理，表的 join，数据类型的导入导出，集合两种模式
Hive 客户端操作，Hive 的管理
Hive 的结构体系
Hive 的自定义函数
Hive 的客户端操作：Thrift 客户端
Hive 的客户端操作：JDBC
Hive 的查询
Hive 的数据导入和导出
Hive 的数据模型
Hive 的数据类型

9. Pig 解析大数集高级过程语言

什么是 Pig?
Pig 的体系结构
Pig 的安装和工作模式
Pig 的内置函数
使用 Pig Latin 语句分析数据
Pig 的自定义函数

10. Hbase 分布式的、面向列的开源数据库

什么是 HBase?
HBase 的体系结构
HBase 的表结构
HBase 的安装和部署

-ROOT-和.META.

HBase Shell

HBase 的 Java 编程接口

HBase 上的过滤器

HBase 上的 MapReduce

HBase 的 HA

Hbase 简介、Hbase 内容介绍、Hbase 的体系结构、Hbase 安装和部署、Hbase shell 示范，删除表头 wmv

Hbase Java Api 访问和调用，访问和删除表，过滤器，MR，HA 以及增删改查。Hbase 相关说明

11. Sqoop 高效传输批量数据的工具

什么是 Sqoop？

Sqoop 是如何工作？

使用 Sqoop

12. Flume 日志收集系统

什么是 Flume？

Flume 的体系结构

安装和配置 Flume

使用 Flume 采集日志数据

13. Zookeeper 分布式系统的可靠协调系统

什么是 ZooKeeper?

ZooKeeper 的体系结构

Zookeeper 能帮我们做什么？

安装和配置 Zookeeper

操作 Zookeeper

阶段性项目：Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称 HDFS。HDFS 有高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求，可以以流的形式访问（streaming access）文件系统中的数据。

第二阶段：Storm 实时计算处理模块

1. 组件模块.Redis 缓存中间件

Redis 简介，内容介绍，Reids 的组织和架构图，基于 windows 和 linux 的安装和部署

Redis 的环境搭建，单机版和分布式版本的安装和组成

Redis 的数据操作，和进阶操作，主从配置，哨兵机制以及分票处理架构

Redis 的实际使用，调用 javaApi 等

2. Storm 实时计算简介，图形解释

Strom 基本概念，应用场景
Storm 和 Hadoop,Spark Streaming 应用场景的对比
Storm 的体系结构
Storm 的运行机制
Storm 架构分析
离线计算和流式计算

3. Storm 的内容大纲，技术角度详细讲解

Storm 核心机制-Ack 容错机制
Storm 通信机制（Netty 和 Disruptor）
Storm 并发度概念
Storm 分组的概念
Storm Api 分析
Storm 的编程模型，Topology，Spout,Bolt,Tuple 等概念
Storm 概念之 Stream,StreamGroup
Storm 分组策略
Storm 事物处理
Storm 消息和容错机制
Storm Trident 概念，开发实例
Storm 配置文件详解

4. Storm 工程部署，单机和集群开发

Storm 单机版测试和使用
Storm 之 zookeeper 集群搭建和使用
Storm 集群的安装和搭建，linux 环境准备
Storm wordCount 实例
Storm 程序本地模式 debug、Storm 程序远程 debug
Storm 实时计算模块设计以及开发
Storm 定时加载每个 worker

5. Storm 补充知识，案例讲解

Storm 实际项目实战解析，难点解析
Storm 源码管理和分析

6. Storm 与其他中间件集成 Api

与 JDBC 集成
与 Redis 集成
与 HDFS 集成
与 HBase 集成
与 Kafka 集成
与 Hive 集成

与 JMS 集成

Storm： Storm 用于流式计算范式。应用范围很广，对于增量信息的实时处理，都应该能应用到，比如：
1) 对用户上传图片/文本实时打标签（用预先训练的模型）
2) 在线学习(on-line learning)，实时对模型进行增量学习。比如广告点击率、推荐预测率之类的模型。

第三阶段：Spark 内存计算处理模块

1. Kafka 消息队列模块

Kafka 基本介绍，JavaApi 和 ShellApi
Kafka 单机，集群部署
Kafka 常见指令解析配置文件说明，Partition，Segment，consumerGroup 等说明和展示
Kafka 分布式集群部署实战，HA 设置

2. Spark 技术简介，架构图设计，技术背景，Scala 语言教程。

Spark 语言简介，与 Hadoop 对比，Scala 应用场景
Scala 语言简介，使用处理流程等
Scala 相关软件介绍，基础语法
Scala 方法和函数，函数式编程特点
Scala 数组，映射，元组，集合
Scala 定义类和构造器
Scala 单例对象
Scala Apply 方法
Scala 的模式匹配
Scala 的 Option 类型和偏函数
Scala 的隐式转换，科里化，Actor 并发模型
Scala 编程练习（单机版 WordCount）
Scala 精通，深入和加强，源码导读

3. Spark 内存计算模型详解

Akka 与 Rpc 简介，并发编程框架的理解和通信的小例子,RPC 编程实战
Spark 的入门：安装和部署，搭建单机版的 standalone 模式，spark->单 master 和多 master
park-Shell 的单机执行和集群执行模式、shell 版本的 wordcount
在 Spark-shell 中运行集群 wordcount.
spark-shell 从 hdfs 中读取数据
通过 Scala 的 Api 写 wordcount，将结果输出到不同的数据源
在集群提交 Spark 任务
Python 写 Spark,在集群上运行
Java 写 Spark，并在集群上运行

R 写 Spark，并运行在集群上
提前预习 RDD 相关概念，总结和复习
SparkRDD 的概念，转换方式和详细解释
Spark 的各种算子 transformation,Action，Map 等详解
Spark 的宽依赖和窄依赖+stage 划分+集群运行原理
Spark 如何设置 cache，如何设置 checkpoint
Spark-sql 简介和入门
Spark datafrom 的简单操作
Spark-sql 的命令行创建表
Spark-sql 的 api 操作的两种模式
Spark-sql 之 hive on Spark 简介
Spark-hive 操作，Spark on yarn 讲解

4. Spark Streaming 技术详解

Spark 实时计算之 Spark streaming 概述和图解
Spark streaming 实时计算、数据源和 Dstream
Spark stremaing 术语定义，离散流，批数据，时间片或批处理时间间隔等
Spark streaming 的编程模型，以及如何使用
Spark streaming Dstream 操作，和算子转换操作
Spark streaming window 窗口操作
Spark streaming updateStateByKey 的讲解
Spark streaming 有无状态的 transformation
Spark streaming sql 编程实战
Spark streaming flume 结合实战操作
Spark streaming kakfa 中读取数据
Spark streaming 结合 ELK 进行实战
Spark streaming 多语言操作，新版本特性
Spark streaming 持久化，性能调优

Spark： Spark 是一个基于内存的开源计算框架，于 2009 年诞生于加州大学伯克利分校 AMPLab（AMP：Algorithms，Machines，People），它最初属于伯克利大学的研究性项目，后来在 2010 年正式开源，并于 2013 年成为了 Apache 基金项目，到 2014 年便成为 Apache 基金的顶级项目，该项目整个发展历程刚过六年时间，但其发展速度非常惊人。

项目一：

项目标题：大数据网站离线日志分析系统

Web 日志分析概述 Web 日志由 Web 服务器产生，可能是 Nginx, Apache, Tomcat 等。从 Web 日志中，我们可以获取网站每类页面的 PV 值（PageView，页面访问量）、独立 IP 数；稍微复杂一些的，可以计算

得出用户所检索的关键词排行榜、用户停留时间最高的页面等；更复杂的，构建广告点击模型、分析用户行为特征等等。

这个项目将来可以做什么（举个例子）：

某电子商务网站，在线团购业务。每日 PV 数 100w，独立 IP 数 5w。用户通常在工作日上午 10:00-12:00 和下午 15:00-18:00 访问量最大。日间主要是通过 PC 端浏览器访问，休息日及夜间通过移动设备访问较多。网站搜索浏量占整个网站的 80%，PC 用户不足 1%的用户会消费，移动用户有 5%会消费。

通过简短的描述，我们可以粗略地看出，这家电商网站的经营状况，并认识到愿意消费的用户从哪里来，有哪些潜在的用户可以挖掘，网站是否存在倒闭风险等。

功能描述：

1. Seo 网站日志功能介绍
2. 网站服务器日志数据分析
3. 网站 apache 服务器数据获取对接
4. 网站数据仓库搭建，数据模型设计
5. 服务器日志数据仓库设计，可视化接口对接
6. 可视化使用，优化，仪表盘使用。

使用哪些技术：Hadoop Spark Hive Hdfs Elkstack Flume HUE OOZIE Kafka

效果图：

项目二：

项目标题：互联网+在线电商平台行业分析系统

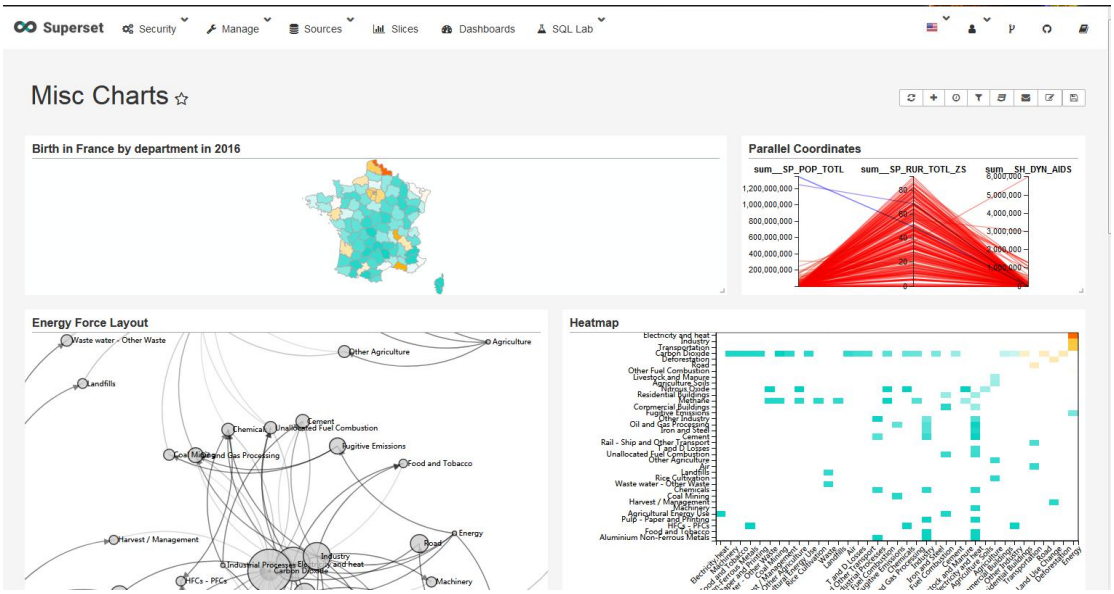
在众多的互联网细分行业中，电商行业起步早，发展时间长，行业特征显著：
1）商品品类及 SKU 多，用户覆盖面广，运营难度大；
2）总体上客单价低（除旅游、奢侈品等外），强调留存与复购；
3）电商产品设计相对成熟，优化运营是重中之重；
4）电商行业竞争白热化，精细化运营是冲出重围的必备技能。
基于现在市场竞争压力，所以我们需要一套项目和设计方案来解决我们对海量数据获取和分析的需求系统。

功能描述：

1. 项目概述重点，难点描述。
2. 电商网站用户商品趋势走向分析和归类。
3. 卖家和买家商品的数据留存模型分析和设计。
4. 用户偏好策略指定，制定用户行为轨迹流程。
5. 需求确定，数据库模型和方案建设，技术选型(spark streaming storm)。
6. 整体方案架构设计，kafka 到 redis 到 spark straming 到数据服务到可视化。
7. 项目隐患排查，知识梳理，重点掌握及了解。

使用哪些技术：Kafka SparkStreaming Redis Hue Superset

效果图：



课程设计的软件基本为最新版本，如需指导安装请联系助教老师。