- 数据开发专业级技能认证考试
  - 大数据简介（作为了解）
    - 大数据是什么
      - 历史
        - 1. 2008年2009年世界范围才开始提出这个概念
        - 2. 2011年麦肯锡咨询首先投入资源进行研究
      - 4V定义
        - Volume：数据量超大
        - Variety：数据类型多样，组成庞大数据集，有结构化和非结构化的数据
        - Velocity：数据增长量很快
        - Value：数据的价值密度低，从数据集中挖掘有价值数据，但是具有巨大的商业价值
        - 如果多加一个V，真实性 — 即是，不像以前对样本抽样，而是全部处理——《大数据时代》
    - 大数据的应用场景
      - 杀熟，对产品有依赖度，有打车实验，孙教授带队样本实验
      - 推荐算法，购物车算法，啤酒尿不湿
      - 金融领域，贷款大数据
      - 电商，用户画像
    - 大数据的发展前景
      - 目前依然在应用初期
        - 1.新兴技术带来新兴行业，创造新的商业价值
        - 2. 推动其他行业领域的科技发展，助力科研
        - 3. 上下游产业链形成了
        - 4. 国家政策支持
    - 企业中大数据的一般的处理流程
      - 1. 数据源
        - 来自传统关系型数据库
        - 日志文件
        - 第三方数据，
        - 数据爬取，爬虫
      - 2. 数据采集
        - sqoop — 与关系型数据库之间的导入与导出
        - flume
          - 采集日志文件，数据流
          - 一部分进HDFS，离线分析，一部分给kafka，实时处理
        - kafka
          - flume和kafka都有一个类似消息队列的机制，缓存大数据环境处理不了的数据
          - 做实时数据流处理
      - 3. 数据存储 — HDFS、HBase、ES — 其实Hbase也是基于HDFS的
      - 4. 数据清洗 — mapreduce 、hive(ETL)、SparkCore、Sparksql
      - 5. 数据分析 — MapReduce、Hive、SparkSQL、impala（impa:le）、kylin
      - 6. 数据展示 — 可视化，做报表
    - 数据部门的组织架构
      - 运维组（工资第三）
        - 集群性能监控
        - Hadoop、flume、kafka、Hbase、spark等框架平台的搭建
      - 大数据开发工程师
        - 数仓组（工资其次）
          - ETL工程师、数据清洗
          - HIVE工程师，数据分析、数仓建模
        - 数据研发组（工资最高）
          - 推荐系统工程师
          - 算法工程师
          - 用户画像工程师
      - 可视化组
        - 可视化工程师（做报表、图标）
  - 主题