1.[**Linux**](http://lib.csdn.net/base/linux)基础和分布式集群技术

**学完此阶段可掌握的核心能力：**

熟练使用[**linux**](http://lib.csdn.net/base/linux)，熟练安装Linux上的软件，了解熟悉负载均衡、高可靠等集群相关概念，搭建互联网高并发、高可靠的服务[**架构**](http://lib.csdn.net/base/architecture)；

**学完此阶段可解决的现实问题：**

搭建负载均衡、高可靠的服务器集群，可以增大网站的并发访问量，保证服务不间断地对外服务；

**学完此阶段可拥有的市场价值：**

具备初级程序员必要具备的Linux服务器运维能力。

1.内容介绍：

在[**大数据**](http://lib.csdn.net/base/hadoop)领域，使用最多的[**操作系统**](http://lib.csdn.net/base/operatingsystem)就是Linux系列，并且几乎都是分布式集群。该课程为大数据的基础课程，主要介绍Linux操作系统、Linux常用命令、Linux常用软件安装、Linux网络、防火墙、Shell编程等。

2.案例：搭建互联网高并发、高可靠的服务架构。

[](http://www.itcast.cn/files/image/201601/20160108132639659.png)

**2.离线计算系统课程阶段**

**1. 离线计算系统课程阶段**

[**Hadoop**](http://lib.csdn.net/base/hadoop)**核心技术框架**

**学完此阶段可掌握的核心能力：**

1、通过对大数据技术产生的背景和行业应用案例了解[**hadoop**](http://lib.csdn.net/base/hadoop)的作用；2、掌握hadoop底层分布式文件系统HDFS的原理、操作和应用开发；3、掌握MAPREDUCE分布式运算系统的工作原理和分布式分析应用开发；4、掌握[**Hive**](http://lib.csdn.net/base/hive)数据仓库工具的工作原理及应用开发。

**学完此阶段可解决的现实问题：**

1、熟练搭建海量数据离线计算平台；2、根据具体业务场景设计、实现海量数据存储方案；3、根据具体数据分析需求实现基于mapreduce的分布式运算程序；

**学完此阶段可拥有的市场价值：**

具备企业数据部初级应用开发人员的能力

**1.1 HADOOP快速入门**

1.1.1 hadoop知识背景

什么是hadoop、hadoop产生背景、hadoop在大数据[**云计算**](http://lib.csdn.net/base/hadoop)中的位置和关系、国内hadoop的就业情况分析及课程大纲介绍

[](http://www.itcast.cn/files/image/201601/20160108132812271.png)

国内外hadoop应用案例介绍

[](http://www.itcast.cn/files/image/201601/20160108133047005.png)

分布式系统概述、hadoop生态圈及各组成部分的简介



1.1.2 [**hive**](http://lib.csdn.net/base/hive)快速入门

hive基本介绍、hive的使用、数据仓库基本知识

1.1.3 数据分析流程案例

web点击流日志数据挖掘的需求分析、数据来源、处理流程、数据分析结果导出、数据展现

1.1.4 hadoop数据分析系统集群搭建

集群简介、服务器介绍、网络环境设置、服务器系统环境设置、JDK环境安装、hadoop集群安装部署、集群启动、集群状态[**测试**](http://lib.csdn.net/base/softwaretest)

HIVE的配置安装、HIVE启动、HIVE使用测试

[](http://www.itcast.cn/files/image/201601/20160108133315808.png)

**1.2 HDFS详解**

1.2.1 HDFS的概念和特性

什么是分布式文件系统、HDFS的设计目标、HDFS与其他分布式存储系统的优劣势比较、HDFS的适用场景

[](http://www.itcast.cn/files/image/201601/20160108132812271.png)

1.2.2 HDFS的shell操作

HDFS命令行客户端启动、HDFS命令行客户端的基本操作、命令行客户端支持的常用命令、常用参数介绍

1.2.3 HDFS的工作机制

HDFS系统的模块架构、HDFS写数据流程、HDFS读数据流程

NAMENODE工作机制、元数据存储机制、元数据手动查看、元数据checkpoint机制、NAMENODE故障恢复、DATANODE工作机制、DATANODE动态增减、全局数据负载均衡

[](http://www.itcast.cn/files/image/201601/20160108141144872.png)

1.2.4 HDFS的[**Java**](http://lib.csdn.net/base/java)应用开发

搭建开发环境、获取api中的客户端对象、HDFS的java客户端所具备的常用功能、HDFS客户端对文件的常用操作实现、利用HDFS的JAVA客户端开发数据采集和存储系统

**1.3 MAPREDUCE详解**

1.3.1 MAPREDUCE快速上手

为什么需要MAPREDUCE、MAPREDUCE程序运行演示、MAPREDUCE编程示例及编程规范、MAPREDUCE程序运行模式、MAPREDUCE程序调试debug的几种方式

1.3.2 MAPREDUCE程序的运行机制

MAPREDUCE程序运行流程解析、MAPTASK并发数的决定机制、MAPREDUCE中的combiner组件应用、MAPREDUCE中的序列化框架及应用、MAPREDUCE中的排序、MAPREDUCE中的自定义分区实现、MAPREDUCE的shuffle机制、MAPREDUCE利用数据压缩进行优化、MAPREDUCE程序与YARN之间的关系、MAPREDUCE参数优化

通过以上各组件的详解，深刻理解MAPREDUCE的核心运行机制，从而具备灵活应对各种复杂应用场景的能力

[](http://www.itcast.cn/files/image/201601/20160108141244255.png)

MAPREDUCE实战编程案例：通过一个实战案例来熟悉复杂MAPREDUCE程序的开发。该程序是从nginx服务器产生的访问服务器中计算出每个访客的访问次数及每次访问的时长。原始数据样例如下：

[](http://www.itcast.cn/files/image/201601/20160108141342345.jpg)

通过一系列的MAPREDUCE程序——清洗、过滤、访问次数及时间分析，最终计算出需求所要的结果，用于支撑页面展现：

[](http://www.itcast.cn/files/image/201601/20160108141433571.png)

**1.4 HIVE增强**

1.4.1 HIVE基本概念

HIVE应用场景、HIVE内部架构、HIVE与hadoop的关系、HIVE与传统[**数据库**](http://lib.csdn.net/base/mysql)对比、HIVE的数据存储机制、HIVE的运算执行机制

[](http://www.itcast.cn/files/image/201601/20160108141524699.png)

1.4.2 HIVE基本操作

HIVE中的DDL操作、HIVE中的DML操作、在HIVE中如何实现高效的JOIN查询、HIVE的内置函数应用、HIVE shell的高级使用方式、HIVE常用参数配置、HIVE自定义函数和TRANSFORM的使用技巧、HIVE UDF开发实例

1.4.3 HIVE高级应用

HIVE执行过程分析及优化策略、HIVE在实战中的最佳实践案例、HIVE优化分类详解、HIVE实战案例--数据ETL、HIVE实战案例--用户访问时长统计

HIVE实战案例--级联求和报表实例：



**离线数据挖掘系统**

**学完此阶段可掌握的核心能力：**

1、通过对数据仓库知识的加强初步掌握数据仓库的核心概念和设计流程;2、通过对HADOOP生态圈关键辅助工具的学习掌握hadoop分析系统的整合能力;3、通过电商系统点击流日志数据挖掘系统实战项目，掌握hadoop离线数据挖掘系统从数据采集、入库、分析及报表展现的整套流程

**学完此阶段可解决的现实问题：**

1、可根据企业具体场景设计海量数据分析系统的通用架构2、根据具体场景的特点有针对性地调整数据分析各环节的技术选型;3、根据具体需求搭建起整套离线数据分析系统;4、简单数据仓库模型的设计和架构5、各环节具体功能模块的开发实现

**学完此阶段可拥有的市场价值：**

具备企业数据部中高级应用开发和初级架构师能力

**2.1 数据仓库增强**

2.1.1 数据仓库及数据模型入门

什么是数据仓库、数据仓库的意义、数据仓库核心概念、数据仓库的体系结构

2.1.2 数据仓库设计

建立数据仓库的步骤、数据的抽取、数据的转换、数据的加载、什么是数据模型、数据模型的常见类型、如何设计数据模型、如何选择数据建模的架构

典型数据模型——星型建模实例

2.1.3 数据仓库建模样例

业务建模、领域建模、逻辑建模、物理建模

web点击流日志分析系统数据仓库设计实战：

通过对数据特点和业务需求的分析，关系梳理，设计出一个主题明确、层次合理的数据模型

[](http://www.itcast.cn/files/image/201601/20160108141741623.png)

**2.2 离线辅助系统**

2.2.1 数据采集系统

数据采集概念介绍

FLUME日志采集框架介绍、FLUME工作机制、FLUME核心组件、FLUME参数配置说明、FLUME采集nginx日志实战案例

[](http://www.itcast.cn/files/image/201601/20160108141846380.png)

2.2.2 任务调度系统

任务调度系统概念介绍、常用任务调度工具比较、OOZIE介绍、OOZIE核心概念、OOZIE的配置说明、OOIZE实现mapreduce/hive等任务调度实战案例

[](http://www.itcast.cn/files/image/201601/20160108141937827.png)

2.2.3 数据导出

数据导出概念介绍、SQOOP基础知识、SQOOP原理及配置说明、SQOOP数据导入实战、SQOOP数据导出实战、SQOOP批量作业操作

[](http://www.itcast.cn/files/image/201601/20160108142046952.png)

**2.3 web点击流日志分析系统实战项目**

2.3.1 项目介绍

1. 在PC时代，营销的核心是购买，在移动互联网时代，其核心是如何实现用户个性化互动，对用户传播更为精准化的内容，而实现这一核心的基础就是对数据的管理和分析——数据驱动型商业模型。

2. 各类互联网服务产品(如网站、APP)都可以通过前端技术获取用户的详细行为数据(如访问的页面，点击的区域、登陆的频次、注册行为、购买的行为等)，将这些点击流日志数据与后台商业数据综合起来，就可以挖掘对公司运营决策意义非凡的商业价值。

3. 本项目则是一个用大数据技术平台实现的点击流日志分析数据挖掘系统，项目内容涵盖一个典型数据挖掘系统中，包括需求分析、数据采集、数据存储管理、数据清洗、数据仓库设计、ETL、业务模型统计分析、数据可视化的全部流程。

2.3.2 需求分析

什么是点击流日志、点击流日志的商业价值、点击流日志分析需求

业务模型指标体系设计——流量分析、来源分析、受访分析、访客分析、转化率分析

2.3.3 系统设计及开发

1. 系统架构设计

[](http://www.itcast.cn/files/image/201601/20160108142232029.png)

2. 数据采集设计及开发——数据格式、数据内容分析、数据生成规律、采集系统技术选型解析、FLUME采集系统实现

3. 数据存储设计及开发——存储技术选型、存储业务流程解析、存储目录规划及文件命名规则、小文件合并实现

4. 数据统计设计及开发——数据预处理、数据加载、原始数据表的创建、数据入库、数据ETL

5. 报表统计设计——数据模型设计、事实表设计、维度表梳理

6. 业务指标设计及开发——PV统计(时间维度、终端维度、地域维度)、来访次数统计(时间维度、地域维度、终端维度)、独立访客统计(时间维度、终端维度、地域维度)、受访页面统计(时间维度、栏目维度)、页面热点图、转化率分析、来源关键词分析、来源[**搜索引擎**](http://lib.csdn.net/base/searchengine)分析、来源广告推广分析

2.3.4 任务调度系统设计实现

任务调度单元实现、各环节任务运行频次及依赖关系梳理、工作流设计及实现、工作流定义配置上传部署、工作流启动即状态监控

[](http://www.itcast.cn/files/image/201601/20160108142351100.png)

2.3.5 数据可视化——结果报表展现

1. hive分析结果使用sqoop导出到msyql数据库

2. 报表展现系统技术选型：

后台使用spingmvc + [**spring**](http://lib.csdn.net/base/javaee) + mybatis

前端页面使用全静态异步刷新技术[**jQuery**](http://lib.csdn.net/base/jquery) + Echarts

3. web展现程序架构搭建，使用maven构建项目工程

4. web展现程序页面设计开发：原型页面设计、[**js**](http://lib.csdn.net/base/javascript)代码开发

5. 最终实现以下数据可视化效果：

(1)流量概况可视化效果：

[](http://www.itcast.cn/files/image/201601/20160108142442663.png)

(2)来源地域分析可视化效果：

[](http://www.itcast.cn/files/image/201601/20160108142530729.png)

(3)来源类型分析可视化效果：

[](http://www.itcast.cn/files/image/201601/20160108142644768.png)

**3.Storm实时计算部分阶段**

**实时课程分为两个部分：流式计算核心技术和流式计算计算案例实战。**

**1.流式计算核心技术**

流式计算核心技术主要分为两个核心技术点：Storm和Kafka，学完此阶段能够掌握Storm开发及底层原理、Kafka的开发及底层原理、Kafka与Storm集成使用。具备开发基于storm实时计算程序的技术能力。

**学完此阶段可掌握的核心能力：**

(1)、理解实时计算及应用场景

(2)、掌握Storm程序的开发及底层原理、掌握Kafka消息队列的开发及底层原理

(3)、具备Kafka与Storm集成使用的能力

**学完此阶段可解决的现实问题：**

具备开发基于storm的实时计算程序的能力

**学完此阶段可拥有的市场价值：**

具备实时计算开发的技术能力、但理解企业业务的能力不足

**1.1、流式计算一般结构**

2011年在海量数据处理领域，Hadoop是人们津津乐道的技术，Hadoop不仅可以用来存储海量数据，还以用来计算海量数据。因为其高吞吐、高可靠等特点，很多互联网公司都已经使用Hadoop来构建数据仓库，高频使用并促进了Hadoop生态圈的各项技术的发展。一般来讲，根据业务需求，数据的处理可以分为离线处理和实时处理，在离线处理方面Hadoop提供了很好的解决方案，但是针对海量数据的实时处理却一直没有比较好的解决方案。就在人们翘首以待的时间节点，storm横空出世，与生俱来的分布式、高可靠、高吞吐的特性，横扫市面上的一些流式计算框架，渐渐的成为了流式计算的首选框架。如果庞麦郎在的话，他一定会说，这就是我要的滑板鞋!

[](http://www.itcast.cn/files/image/201601/20160108142854719.png)

上图是流式分析的一般架构图，抽象出四个步骤就是数据采集、数据缓冲、数据处理、数据输出。一般情况下，我们采用Flume+kafka+Storm+[**Redis**](http://lib.csdn.net/base/redis)的结构来进行流式数据分析。实时部分的课程主要是针对Kafka、Storm进行学习

**1.2、流式计算可以用来干什么**

一淘-实时分析系统：实时分析用户的属性，并反馈给搜索引擎。最初，用户属性分析是通过每天在云梯上定时运行的MR job来完成的。为了满足实时性的要求，希望能够实时分析用户的行为日志，将最新的用户属性反馈给搜索引擎，能够为用户展现最贴近其当前需求的结果。

携程-网站性能监控：实时分析系统监控携程网的网站性能。利用[**HTML5**](http://lib.csdn.net/base/html5)提供的performance标准获得可用的指标，并记录日志。Storm集群实时分析日志和入库。使用DRPC聚合成报表，通过历史数据对比等判断规则，触发预警事件。

一个游戏新版本上线，有一个实时分析系统，收集游戏中的数据，运营或者开发者可以在上线后几秒钟得到持续不断更新的游戏监控报告和分析结果，然后马上针对游戏的参数和平衡性进行调整。这样就能够大大缩短游戏迭代周期，加强游戏的生命力。

实时计算在腾讯的运用：精准推荐(广点通广告推荐、新闻推荐、视频推荐、游戏道具推荐);实时分析([**微信**](http://lib.csdn.net/base/wechat)运营数据门户、效果统计、订单画像分析);实时监控(实时监控平台、游戏内接口调用)

为了更加精准投放广告，阿里妈妈后台计算引擎需要维护每个用户的兴趣点(理想状态是，你对什么感兴趣，就向你投放哪类广告)。用户兴趣主要基于用户的历史行为、用户的实时查询、用户的实时点击、用户的地理信息而得，其中实时查询、实时点击等用户行为都是实时数据。考虑到系统的实时性，阿里妈妈使用Storm维护用户兴趣数据，并在此基础上进行受众定向的广告投放。

**1.3、Storm核心技术点**

基础技术点

linux环境准备、zookeeper集群搭建、Storm集群搭建、Storm配置文件配置项讲解、集群搭建常见问题解决。

[](http://www.itcast.cn/files/image/201601/20160108142945307.png)

Storm练习案例

根据蚂蚁金服提供的最新数据，今年双十一的交易峰值为8.59万笔/秒，是去年3.85万笔/秒的2.23倍。这一数据也超过了6万笔/秒的预估。如何实时的计算订单金额，让公司领导层看到呢?

[](http://www.itcast.cn/files/image/201601/20160108143035755.png)

(图为双十一支付宝成交金额)

Storm基础及原理

Storm常用组件和编程API：Topology、 Spout、Bolt、Storm分组策略(stream groupings)、Storm项目maven环境搭建、使用Strom开发一个WordCount例子、Storm程序本地模式debug、Storm消息可靠性及容错原理、Storm任务提交流程、Strom消息容错机制。

[](http://www.itcast.cn/files/image/201601/20160108143144078.png)

(图为storm组件)

**1.4、Kafka核心技术点**

Storm结合消息队列Kafka：消息队列基本概念(Producer、Consumer、Topic、Broker等)、消息队列Kafka使用场景、Storm结合Kafka编程API、Kafka负载均衡、Kafka消息存储原理等。

[](http://www.itcast.cn/files/image/201601/20160108143250209.png)

(图为Kafka消息队列原理)

**2.流式计算案例实战**

实战案例部分主要有三个企业实战案列，分别是基于点击流的日志分析系统、基于系统日志的监控告警系统、基于订单系统的交易风控系统，三个案列是企业中的典型项目。学完此阶段能够独立根据企业的业务性质开发相关的storm程序。

**学完此阶段可掌握的核心能力：**

1、掌握企业核心业务需求

2、掌握实时系统常见的开发流程及运营经验

**学完此阶段可解决的现实问题：**

可以独立开发storm程序来满足业务需求

**学完此阶段可拥有的市场价值：**

熟练学习和掌握后，可满足企业开发的初级需求，根据市场反馈数据看，薪资普遍在 15000-18000元/月。

**2.1、案例：流量日志分析**

流量日志分析之漏斗模型：大型电商网站，上亿规模的用户，千万级别的活跃用户，如何评估一个商品专题页面的效果好不好呢?比如：浏览次数、加入购物车次数、下单次数、支付次数、完成。

[](http://www.itcast.cn/files/image/201601/20160108143355933.png)

(图为日志分析漏斗模型-数据部必备)

流量日志分析之基础数据分析：电商网上商品数量在千万级别，店铺数量在百万级别，如何实时的计算一个每个商品页的访问数、用户数、来源信息等基础信息呢?如何实时的计算每个店铺的访问数、用户数、来源信息等基础数据呢?

[](http://www.itcast.cn/files/image/201601/20160108143445781.png)

(图为页面浏览分析-数据部必备)

**2.2、案例：统一监控告警系统**

随着公司业务发展，支撑公司业务的各种系统越来越多，为了保证公司的业务正常发展，急需要对这些线上系统的运行进行监控，做到问题的及时发现和处理，最大程度减少对业务的影响。不同业务的会有几十上百台服务器去支撑，大型企业可能是成千上万台服务器，那么每台服务器的硬件状态、业务应用状态如何实时的监控，做到及时发现，快速解决问题呢?

[](http://www.itcast.cn/files/image/201601/20160108143549577.png)

(图为企业产生日志的系统清单)

统一监控系统触发的短信告警

[](http://www.itcast.cn/files/image/201601/20160108143646105.png)

统一监控系统触发的邮件告警

[](http://www.itcast.cn/files/image/201601/20160108143734458.png)

**2.3、案例：交易风控系统**

电子商务是以互联网络为平台的贸易新模式，它的一个最大特点是强调参加交易的各方和所合作的伙伴都要通过Internet密切结合起来，共同从事在网络环境下的商业电子化应用。用户信息容易受到计算机病毒、黑客的攻击，商业信息和数据易于搭截侦听、口令试探和窃取，为了防止用户信息异常给商家和用户带来不必要的损失，企业期望针对用户的订单进行分析，对触发规则的订单进行风险预警，在必要情况下进行拦截及锁定订单。

[](http://www.itcast.cn/files/image/201601/20160108143832155.png)

(图为订单异常拦截)

**4.Spark内存计算阶段**

**学完此阶段可掌握的核心能力：**

1.掌握[**Scala**](http://lib.csdn.net/base/scala)函数式编程特性，熟练使用[**scala**](http://lib.csdn.net/base/scala)开发程序，可以看懂其他用Scala编写源码。

2.搭建[**Spark**](http://lib.csdn.net/base/spark)集群、使用Scala编写Spark计算程序，熟练掌握Spark原理，可以阅读Spark源码。

3.理解DataFrame和RDD之间的关系，熟练使用DataFrame的API，熟练使用Spark SQL处理结构化数据，通过Spark SQL对接各种数据源，并将处理后结果写回到存储介质中。

4.理解Spark Streaming的核心DStream，掌握DStream的编程API并编写实时计算程序。

**学完此阶段可解决的现实问题：**

熟练使用Scala快速开发Spark大数据应用，通过计算分析大量数据，挖掘出其中有价值的数据，为企业提供决策依据。

**学完此阶段可拥有的市场价值：**

学习完spark并掌握其内容，将具备中级大数据工程师能力，薪水可以达到 20K~25K。

**1.Scala函数式编程**

介绍：Scala是一门集面向对象和函数式编程与一身的编程语言，其强大的表达能力、优雅的API、高效的性能等优点受到越来越多程序员的青睐。Spark底层就是用Scala语言编写，如果想彻底掌握Spark，就必须学好Scala。

案例：Scala编程实战，基于Akka框架，编写一个简单的分布式RPC通信框架

[](http://www.itcast.cn/files/image/201601/20160108144053407.png)

**2.使用Spark处理离线数据**

介绍：Spark是基于内存计算的大数据并行计算框架，具有高容错性和高可伸缩性，可以在大量廉价硬件之上部署大规模集群，在同等条件下要比Hadoop快10到100倍。

[](http://www.itcast.cn/files/image/201601/20160108144143829.png)

**3.使用Spark SQL处理结构化数据**

介绍：Spark SQL的前身是Shark，专门用来处理结构化的数据，类似Hive，是将SQL转换成一系列RDD任务提交到Spark集群中运行，由于是在内存中完成计算，要比hive的性能高很多，并且简化了开发Spark程序的难度同时提高了开发效率。

[](http://www.itcast.cn/files/image/201601/20160108144233444.png)

**4.使用Spark Streaming完成实时计算**

介绍：Spark Streaming类似于Apache Storm，用于流式数据的处理。根据其官方文档介绍，Spark Streaming有高吞吐量和容错能力强等特点。Spark Streaming支持的数据输入源很多，例如：Kafka、Flume、Twitter、ZeroMQ和简单的TCP套接字等等。数据输入后可以用Spark的高度抽象原语如：map、reduce、join、window等进行运算。而结果也能保存在很多地方，如HDFS，数据库等。另外Spark Streaming也能和MLlib([**机器学习**](http://lib.csdn.net/base/machinelearning))以及Graphx完美融合

[](http://www.itcast.cn/files/image/201601/20160108144342459.png)

**5.Spark综合项目：**

介绍：该项目使用了Spark SQL和Spark Streaming对游戏整个生命周期产生的数据进行了分析，从玩家第一次登录游戏到每天产生的游戏日志，通过大量的数据来分析该游戏的运营情况和玩家的各种行为：如活跃用户、用户留存、充值比例、游戏收人、外挂分析等。

通过玩家登录游戏产生的数据分析一天之内各个时间段进入游戏的情况

[](http://www.itcast.cn/files/image/201601/20160108144501524.png)

通过玩家登陆游戏产生的数据分析玩家在全国地区的分步情况，调整广告投放策略

[](http://www.itcast.cn/files/image/201601/20160108144911588.png)

用户留存指标可以分析游戏对玩家的吸引力，分析用户流失原因

[](http://www.itcast.cn/files/image/201601/20160108145032474.png)

用户等级信息可以分析玩家等等级分布情况、调整装备爆率和游戏难度

[](http://www.itcast.cn/files/image/201601/20160108145315182.png)

通过上面游戏各个数据指标的分析，可以让游戏运维者了解游戏的运维情况，为运维者提供各种个性化的调整策略，从而保证游戏健康、稳定的运营。