# 项目介绍和需求管理

## 学习目标

了解知行教育大数据的项目背景

了解知行教育大数据的看板划分

了解项目技术架构

了解ClouderaManager的应用场景

掌握ClouderaManager架构

能够使用Vmware虚拟机环境

## 项目简介

知行教育大数据分析平台，突出的是“真”，此项目是传智播客联合三方K12教育机构共同研发，并在上线发布后转换为课程，过程真实细致，采用主流的大数据技术和工具，内容特点：

1. 包含了需求分析、设计转换、研发、测试到上线部署维护的完整项目流程。
2. 真实的教育大数据业务逻辑，包括：访问、咨询、意向、线索、报名、考勤等各个阶段，大幅提升学员在教育行业中的竞争力。
3. 深入讲解数据仓库的分层与建模过程。
4. 海量数据场景下的性能优化。
5. 拉链表的具体应用、对变化的更新数据进行存储和分析。
6. 每个公司都会要求使用的版本控制工具。
7. 大公司、核心项目必备的Code Review技能。
8. 可视化部署与运维大数据环境。

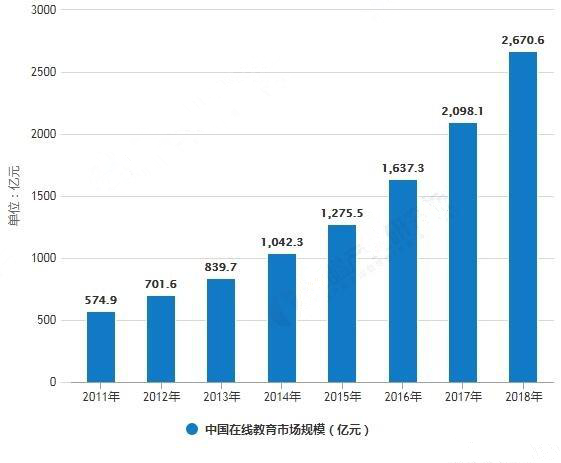
## 项目背景

### 在线教育行业机遇

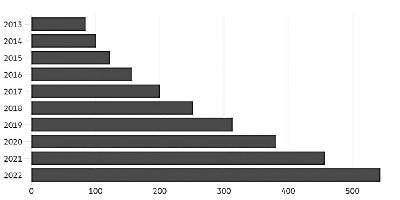
近年来，在线教育产业发展十分迅速。尤其是2018年以来，在线教育平台动作不断，除了洋葱数学、考虫、作业盒子、火花思维、VIPKID、阿卡索等平台纷纷融资外，诸多在线教育平台纷纷上市。新东方在线也在2019年3月成功上市。

在市场规模方面，在线教育很大程度上是随着移动互联网的浪潮发展起来的，在传统的PC时代，虽然出现了网络教育形式，但是真正的所谓在线教育仍然是在2011年左右开始爆发的，这也正是移动互联网发展和普及的时间点。

统计数据显示，**2011-2017年**中国在线教育市场规模呈逐年增长趋势。2011年中国在线教育市场规模已达574.9亿元，2014年中国在线教育市场规模突破了千亿元，截止至2017年中国在线教育市场规模增长至突破2000亿元，达到了2089.1亿元，同比增长28.1%。2018年我国在线教育市场规模达到2670.6亿元左右。中国在线教育市场规模同比增长稳健。



**在近期新冠肺炎引发的疫情期间**，全国数千万学生在线上教育平台迎来新学期。在开学的第一周，全国300多个城市的数十万教师变身主播，通过阿里、腾讯和网易等互联网公司旗下线上教育平台让学生们实现“在家上课”。在线下培训机构全面停课之际，线上教育又快速承接线下需求、聚集流量，在这个特殊时期迎来一次爆发式增长。



**用户需求猛增**

A股自春节开盘后，在线教育板块迎来一波大涨行情。2月12日，在线教育概念股全线飘红，板块涨幅2.91%。在美股，在线教育中概股也表现火热，去年在纽交所上市的网易有道11日上涨39.48%。当日，其他在纳斯达克或纽交所挂牌上市的中概股教育股几乎清一色大涨，“跟谁学”上涨4.56%，好未来上涨3.37%， 51Talk上涨21.06%，流利说上涨15.13%。资本市场对中国在线教育产业持续看好。

这次疫情提供了一次难得的强迫型教育市场的机会。在线教育的优势在本次疫情发生期间再次凸显，解决了教育资源不能异地、分散化使用的问题，证明教育信息化实施的必要性。在线教育迎来快速发展，主要增量将来自于三、四线城市。

2019年中国在线教育用户规模接近4亿，同比增长超过两成。在过去一年里，在线教育以其能够突破地域空间的特性，成为家长和学子的优选项，成为多方角逐的竞技场。

国内多个科技、媒体巨头纷纷开始在线教育领域布局。去年年底，知乎与快手共同发布“快知计划”，拿出百亿流量扶持知识创作者（课程研发和讲师）。字节跳动投资K12大数据精准教学运营商极课大数据，以及投资早教内容提供商HnR新升力。在此之前，腾讯向在线少儿英语教育平台VIPKID投资1.5亿美元，并宣布正式成立“腾讯教育”，向个人、学校、教育机构、教育管理部门，提供智能连接、智能教学、智能科研和智能管理等服务。抖音也公开提到平台对教育领域的支持。

### 行业发展的痛点

受互联网+概念的催化，教育市场发展火热，越来越多的教育机构和平台不断涌现，包括有线上学习和线下培训，K12教育和职业教育等，那些注重用户服务、教育质量的平台会最终胜出。目前的**企业痛点**：

1. 数据量大，现有MySQL业务数据库直接读取模式不能满足业务统计性能、效率需要
2. 系统多、数据分散，缺少从营销、咨询、报名、教学等等完整业务环节的数据贯通
3. 统计分析难度高、工作量大。缺少元数据、数据集合的规范存储，业务部门有数据分析角度需求时，需要程序员、DBA突击查数据、做报表，尤其年底各个部门排队等DBA协助出数据

如何提高用户服务水平，提高教育质量是每个机构都面临的问题。信息的共享和利用不充分，就导致尽管学校多年的信息化应用积累了大量的数据，但信息孤岛的壁垒一直没有打破，对这些数据无法进一步的挖掘、分析、加工、整理，不能给学校教育、教学、研发、总务等各方面管理决策提供科学、有效的数据支撑。

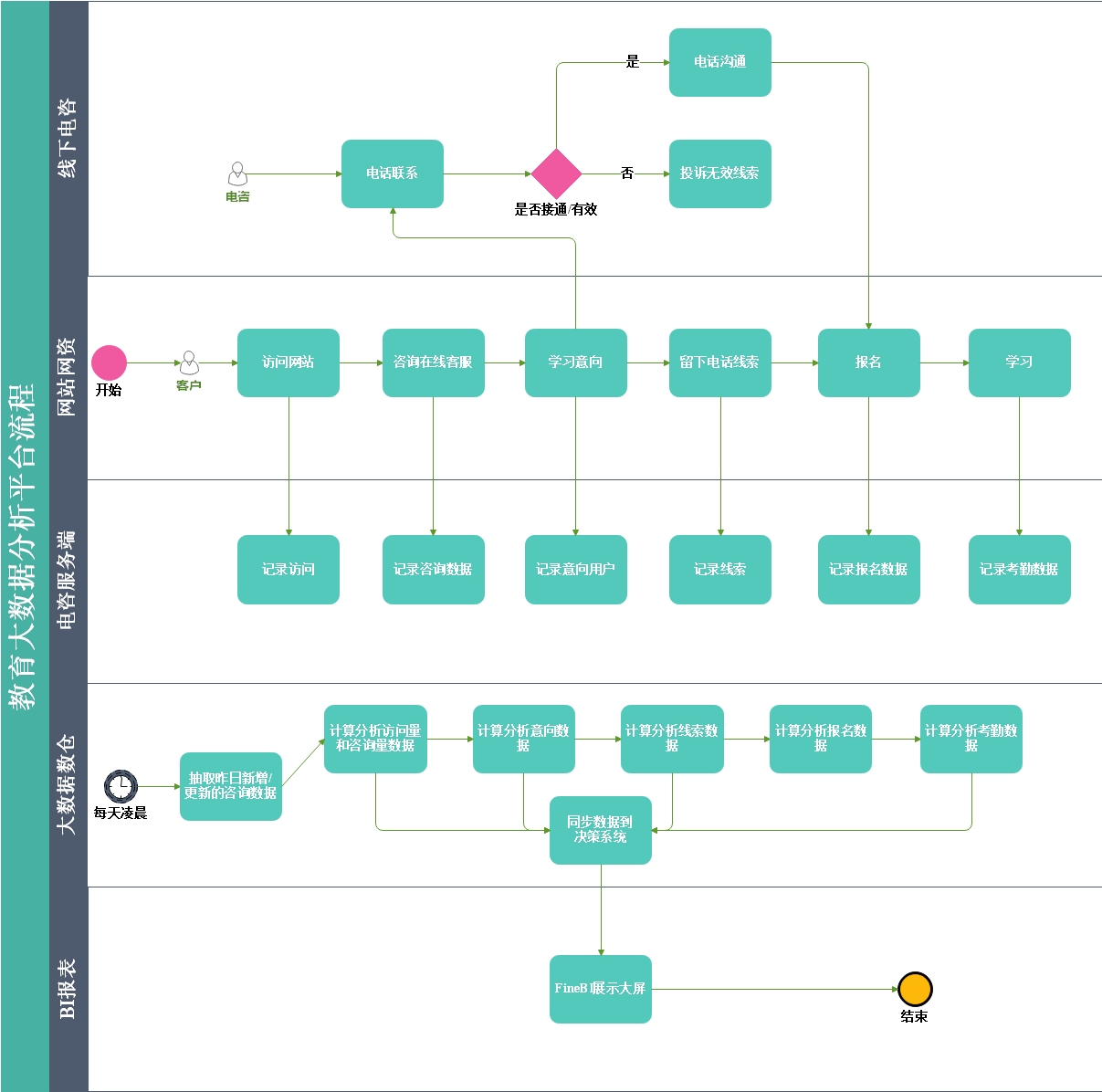
### 大数据技术的应用

大数据技术的应用可以从海量的用户行为数据中进行挖掘分析，根据分析结果优化平台的服务质量，最终满足用户的需求。教育大数据分析平台项目就是将大数据技术应用于教育培训领域，为企业经营提供数据支撑：

1. 建立集团数据仓库，统一集团数据中心，把分散的业务数据进行预先处理和存储
2. 根据业务分析需要，从海量的用户行为数据中进行挖掘分析，定制多维的数据集合，形成数据集市，供各个场景主题使用
3. 前端业务数据展示选择和控制，选取合适的前端数据统计、分析结果展示工具

## 在线教育业务需求





### 访问和咨询用户数据看板

客户访问和咨询主题，顾名思义，分析的数据主要是客户的访问数据和咨询数据。但是经过需求调研，这里的访问数据，实际指的是访问的客户量，而不是客户访问量。原始数据来源于咨询系统的mysql业务数据库。

用户关注的核心指标有：1、总访问客户量、2、地区独立访客热力图、3、访客咨询率趋势、4、客户访问量和访客咨询率双轴趋势、5、时间段访问客户量趋势、6、来源渠道访问量占比、7、搜索来源访问量占比、7、活跃页面排行榜。

#### 总访问客户量

说明：统计指定时间段内，访问客户的总数量。能够下钻到小时数据。

展现：线状图

指标：访问客户量

维度：年、季度、月

粒度：天

条件：年、季度、月

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 地区独立访客热力图

说明：统计指定时间段内，访问客户中各区域人数热力图。能够下钻到小时数据。

展现：地图热力图

指标：按照地区聚合访问的客户数量

维度：年、季度、月

粒度：天

条件：年、季度、月

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 访客咨询率趋势

说明：统计指定时间段内，不同地区（省、市）访问的客户中发起咨询的人数占比；

咨询率=发起咨询的人数/访问客户量；客户与网咨有说一句话的称为有效咨询。

展现：线状图

指标：访客咨询率

维度：年、月、城市

粒度：天

条件：年、季度、月、省、市

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 客户访问量和访客咨询率双轴趋势

说明：统计指定时间段内，每日客户访问量/咨询率双轴趋势图。能够下钻到小时数据。

每日客户访问量可以复用指标1数据；

咨询率可以复用指标3的数据。

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 时间段访问客户量趋势

说明：统计指定时间段内，1-24h之间，每个时间段的访问客户量。

横轴：1-24h，间隔为一小时，纵轴：指定时间段内同一小时内的总访问客户量。

展现：线状图、柱状图、饼状图

指标：某小时的总访问客户量

维度：天

粒度：小时

条件：天

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

#### 来源渠道访问量占比

说明：统计指定时间段内，不同来源渠道的访问客户量占比。能够下钻到小时数据。

展现：饼状图

指标：比值

维度：年、季度、月

粒度：天

条件：年、季度、月

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 搜索来源访问量占比

说明：统计指定时间段内，不同搜索来源的访问客户量占比。能够下钻到小时数据。

展现：饼状图

指标：比值

维度：年、季度、月

粒度：天

条件：年、季度、月

数据来源：咨询系统的web\_chat\_ems\_2019\_12等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

#### 活跃页面排行榜

说明：统计指定时间段内，产生访问客户量最多的页面排行榜TOPN。能够下钻到小时数据。

展现：柱状图

指标：访问客户量

维度：页面、年、季度、月

粒度：天

条件：年、季度、月、Top数量

数据来源：咨询系统的 web\_chat\_text\_ems\_2019\_11等月表

按年：显示指定年范围内每天的客户访问量

按季度：显示指定季度范围内每天的客户访问量

按月：显示指定月份范围内每天的客户访问量

### 意向用户看板

#### 意向学员位置热力图

说明：统计指定时间段内，新增的意向客户，所在城市区域人数热力图。展现：地图热力图

维度：年、月、线上线下

指标：按照地区聚合意向客户id数量

粒度：天，可以下钻到小时数据。

条件：年、月、线上线下

数据来源：客户管理系统的customer(客户静态信息表) 、customer\_relationship（客户意向表）

#### 总意向量

说明：计期内，新增意向客户（包含自己录入的意向客户）总数。

展现：线状图

条件：年、月、线上线下

维度：年、月、线上线下

指标：总意向客户量

粒度：天，可以下钻到小时数据。

数据来源：客户管理系统的customer\_relationship意向表

#### 意向学科排名

说明：统计指定时间段内，新增的意向客户中，意向学科人数排行榜。学科名称要关联查询出来。

展现：柱状图

条件：年、月、线上线下

维度：年、月、线上线下、学科

指标：学科意向客户量

粒度：天，可以下钻到小时数据。

数据来源：客户管理系统的customer\_clue（客户线索表）、customer\_relationship（客户意向表）、itcast\_subject（学科表）

#### 意向校区排名

说明：统计指定时间段内，新增的意向客户中，意向校区人数排行榜。

展现：柱状图

条件：年、月、线上线下

维度：年、月、线上线下、校区

指标：校区意向客户量

粒度：天，可以下钻到小时数据。

数据来源：客户管理系统的

注意：学校id，同步时，0和null转换为统一数据，都转换为-1

#### 来源渠道占比

说明：统计指定时间段内，新增的意向客户中，不同来源渠道的意向客户占比。

展现：饼状图

条件：年、月、线上线下

维度：年、月、线上线下、来源渠道

粒度：天，可以下钻到小时数据。

指标：来源渠道意向客户量

数据来源：客户管理系统的customer\_clue（客户线索表）、customer\_relationship（客户意向表）

#### 意向贡献中心占比

说明：统计指定时间段内，新增的意向客户中，各咨询中心产生的意向客户数占比情况。

展现：饼状图

条件：年、月、线上线下

维度：年、月、线上线下、咨询中心

指标：咨询中心意向客户数

粒度：天，可以下钻到小时数据。

数据来源：客户管理系统的customer\_relationship（客户意向表）、employee（员工表）、scrm\_department（部门表）

### 有效线索看板

#### 有效线索转化率

说明：统计期内，访客咨询产生的有效线索的占比。有效线索量/咨询量，有效线索指的是拿到电话且电话有效。

展现：线状图。双轴：有效线索量、有效线索转化率。

条件：年、月、线上线下

维度：年、月、线上线下

指标：访客咨询率=有效线索量/咨询量

粒度：天

数据来源：客户管理系统的customer\_clue线索表、customer\_relationship意向表、customer\_appeal申诉表；咨询系统的web\_chat\_ems访问咨询表

#### 有效线索转化率时间段趋势

说明：统计期内，1-24h之间，每个时间段的有效线索转化率。横轴：1-24h，间隔为1h，纵轴：每个时间段的有效线索转化率。

展现：线状图

条件：年、月、线上线下

维度：年、月、线上线下

指标：某小时的总有效线索转化率

粒度：区间内小时段（区间内同一个时间点的总有效线索转化率）

数据来源：客户管理系统的customer\_clue线索表、customer\_relationship意向表、customer\_appeal申诉表；咨询系统的web\_chat\_ems访问咨询表

#### 有效线索量

说明：统计期内，新增的咨询客户中，有效线索的数量。

展现：线状图。

条件：年、月、线上线下

维度：年、月、线上线下

指标：有效线索的数量

粒度：天

数据来源：客户管理系统的customer\_clue线索表、customer\_relationship意向表、customer\_appeal申诉表

### 报名用户看板

此主题下指标需要能够下钻到小时数据。

#### 校区报名柱状图

说明：统计期内，全部报名客户中，各校区报名人数分布。

展现：柱状图

条件：年、月，校区

维度：天区间，按查询条件来定

指标：报名人数

粒度：天/线上线下/校区

数据来源：客户管理系统的customer\_relationship、itcast\_clazz报名课程表

#### 学科报名柱状图

说明：统计期内，全部报名客户中，各学科报名人数分布。

展现：柱状图

条件：年、月，学科

维度：天区间，按查询条件来定

指标：报名人数

粒度：天/线上线下/学科

数据来源：客户管理系统的customer\_relationship、itcast\_clazz报名课程表

#### 总报名量

说明：统计期内，已经缴费的报名客户总量。

展现：数值。

条件：年、月

维度：年、月

指标：报名客户总量

粒度：天

数据来源：客户管理系统的customer\_relationship表

#### 线上报名量

说明：总报名量中来源渠道为线上访客渠道的报名总量

展现：线状图。

条件：年、月

维度：天区间，按查询条件来定

指标：报名客户总量

粒度：天

数据来源：客户管理系统的customer\_relationship表

#### 意向用户报名转化率

说明：统计期内，新增的意向客户中报名的客户占比。全部报名人数/全部新增的意向人数

展现：线状图。双轴：全部报名人数、报名转化率。

条件：年、月

维度：天/线上线下

指标：报名转化率=全部报名人数/全部新增的意向人数

粒度：天

数据来源：客户管理系统的customer\_relationship表

#### 有效线索报名转化率

说明：线上报名量/线上有效线索量，此处的线索量需要排除已申诉数据。

展现：线状图。双轴：线上报名人数、线上报名转化率。

条件：年、月

维度：天/线上线下

指标：线上报名转化率=线上报名人数/线上有效线索量

粒度：天

数据来源：客户管理系统的customer\_relationship表、customer\_clue表、customer\_appeal表

#### 日报名趋势图

说明：统计期内，每天报名人数的趋势图。

展现：线状图。

条件：年、月

维度：天/线上线下

指标：报名人数

粒度：天

数据来源：客户管理系统的customer\_relationship表

#### 校区学科的报名学员TOP

说明：统计期内，全部报名学员中，校区学科排行榜，topN。A校区b学科第一，B校区a学科第二等等。

展现：柱状图

条件：年、月，校区，学科，数据量N

维度：天/线上线下

指标：报名学员人数

粒度：各校区各学科的报名人数和

数据来源：客户管理系统的customer\_relationship表、itcast\_clazz表

#### 来源渠道占比

说明：统计期内，全部报名学员中，不同来源渠道的报名学员占比情况。

展现：饼状图

条件：年、月

维度：天/线上线下/来源渠道

指标：比值

数据来源：客户管理系统的customer\_relationship表

#### 咨询中心报名贡献

说明：统计期内，全部报名学员中，各咨询中心的报名学员人数占比情况。

展现：饼状图

条件：年、月，咨询中心

维度：天/线上线下/咨询中心

指标：报名学员人数

粒度：天/报名学员人数

数据来源：客户管理系统的customer\_relationship表、employee表、scrm\_department表

### 学生出勤看板

#### 班级出勤人数

说明：统计指定时间段内，不同班级的出勤人数。打卡时间在上课前40分钟(否则认为无效)~上课时间点之内，且未早退，则为正常上课打卡。可以下钻到具体学生的出勤数据。跨天数据直接累加。

指标：出勤人数

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统teach的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表。

#### 班级出勤率

说明：统计指定时间段内，不同班级的学生出勤率。可以下钻到具体学生的出勤数据。出勤率=出勤人数/当日在读学员人数。

指标：出勤率

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表、class\_studying\_student\_count班级在读学生人数。

#### 班级迟到人数

说明：统计指定时间段内，不同班级的迟到人数。上课10分钟后视为迟到。可以下钻到具体学生的迟到数据。跨天数据直接累加。

指标：迟到人数

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表。

#### 班级迟到率

说明：统计指定时间段内，不同班级的学生迟到率。上课10分钟后视为迟到。可以下钻到具体学生的迟到数据。迟到率=迟到人数/当日在读学员人数。

指标：迟到率

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表、class\_studying\_student\_count班级在读学生人数。

#### 班级请假人数

说明：统计指定时间段内，不同班级的请假人数。跨天数据直接累加。

指标：请假人数

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的student\_leave\_apply学生请假申请表、tbh\_class\_time\_table班级作息时间表、course\_table\_upload\_detail班级课表。

#### 班级请假率

说明：统计指定时间段内，不同班级的学生请假率。可以下钻到具体学生的请假数据。请假率=请假人数/当日在读学员人数。

指标：请假率

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的student\_leave\_apply学生请假申请表、class\_studying\_student\_count班级在读学生人数。

#### 班级旷课人数

说明：统计指定时间段内，不同班级的旷课人数。跨天数据直接累加。旷课人数=当日在读学员人数-出勤人数-请假人数。

指标：旷课人数

维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表、student\_leave\_apply学生请假申请表。

#### 班级旷课率

说明：统计指定时间段内，不同班级的学生旷课率。旷课率=旷课人数/当日在读学员人数。

指标：旷课率

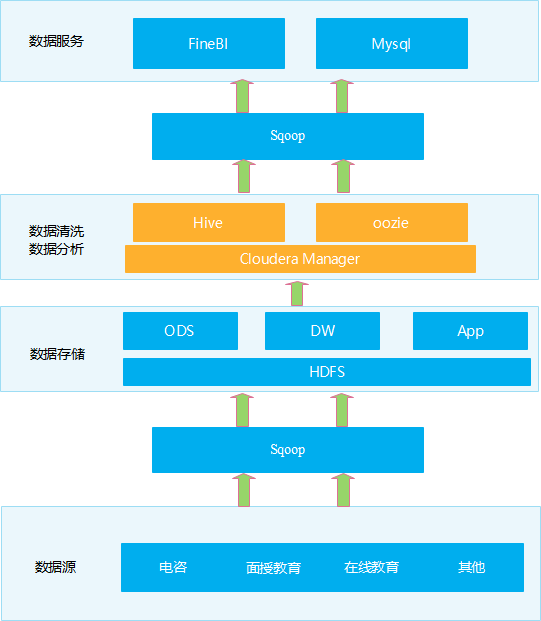
维度：年、月、天

粒度：上午、下午、晚自习

条件：年、月

数据来源：教学实施与保障系统的course\_table\_upload\_detail班级课表、tbh\_student\_signin\_record学生打卡记录表、tbh\_class\_time\_table班级作息时间表、student\_leave\_apply学生请假申请表、class\_studying\_student\_count班级在读学生人数。

## 项目技术架构



数据源：电咨、线下面授、在线教育等OLTP系统的数据，大多存储于mysql。

数据抽取：使用sqoop实现关系型数据库和大数据集群的双向同步。

数据存储：HDFS

数据清洗：数据的清洗、转换、统计分析等都是使用基于CM管理的Hive来进行的。

数据分析：数据的清洗、转换、统计分析等都是使用基于CM管理的Hive来进行的。

数据同步：使用sqoop实现关系型数据库和大数据集群的双向同步。

OLAP数据服务：采用常用的Mysql数据库。

## 理解Cloudera Manager

### 产生的背景

#### Apache版本的大数据组件

Hadoop的原始版本为Apache的开源版本，在国内的使用非常多。

##### 优点

* 完全开源，更新速度很快
* 大数据组件在部署过程中可以深刻了解其底层原理
* 可以了解各个组件的依赖关系

##### 缺点

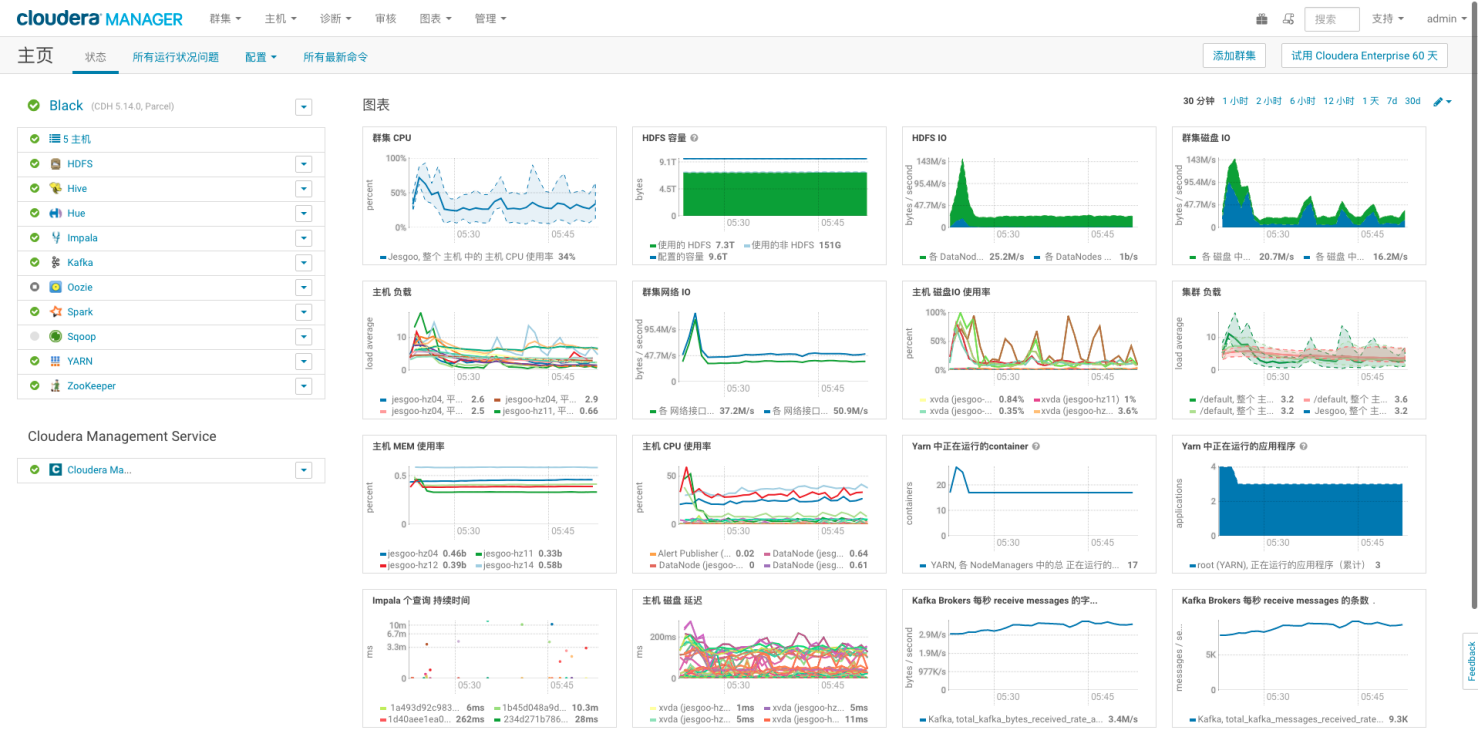
* 部署过程极其复杂，超过20个节点的时候，手动部署已经超级累
* 各个组件部署完成后，各个为政，没有统一化管理界面
* 组件和组件之间的依赖关系很复杂，一环扣一环，部署过程心累
* 各个组件之间没有统一的metric可视化界面，比如说hdfs总共占用的磁盘空间、IO、运行状况等
* 优化等需要用户自己根据业务场景进行调整（需要手工的对每个节点添加更改配置，效率极低，我们希望的是一个配置能够自动的分发到所有的节点上）

#### CDH版本大数据组件

正是为了解决Apache原生版本的各种缺陷，诞生了可以使用Cloudera Manager进行管理的CDH版本。CDH是Apache Hadoop和相关项目中最完整、最稳定的、经过测试和最流行的发行版。Cloudera Manager是用于管理CDH群集的B/S应用程序。

##### 优点

* 统一化的可视化界面 自动部署和配置，大数据各类组件(hadoop、hive、hue、kudu、impala、zookeeper等)安装、调优极其便捷 零停机维护（免费版本不具有弹性升级）
* 多用户管理（权限控制）
* 稳定性极好（部分优化措施都已经调整好）



##### 缺点

* server和agent需要占用额外的内存和cpu(server占用内存为2G，agent占用内存1G，总共cpu为0.5核)
* 对linux常用命令需要了解颇深
* 对hadoop的apache版本有一定的安装经验和调优经验

### ClouderaManager介绍

Cloudera Manager是用于管理CDH群集的B/S应用程序。Cloudera Manager通过对CDH集群的每个部分提供细粒度的可视性和控制来设置企业部署的标准，使运营商能够提高性能，提高服务质量，提高合规性并降低管理成本。

使用Cloudera Manager，可以轻松部署和集中操作完整的CDH堆栈和其他托管服务（Hadoop、Hive、Spark、Kudu）。其特点：应用程序的安装过程自动化，将部署时间从几周缩短到几分钟; 并提供运行主机和服务的集群范围的实时监控视图; 提供单个中央控制台，以在整个群集中实施配置更改; 并集成了全套的报告和诊断工具，可帮助优化性能和利用率。

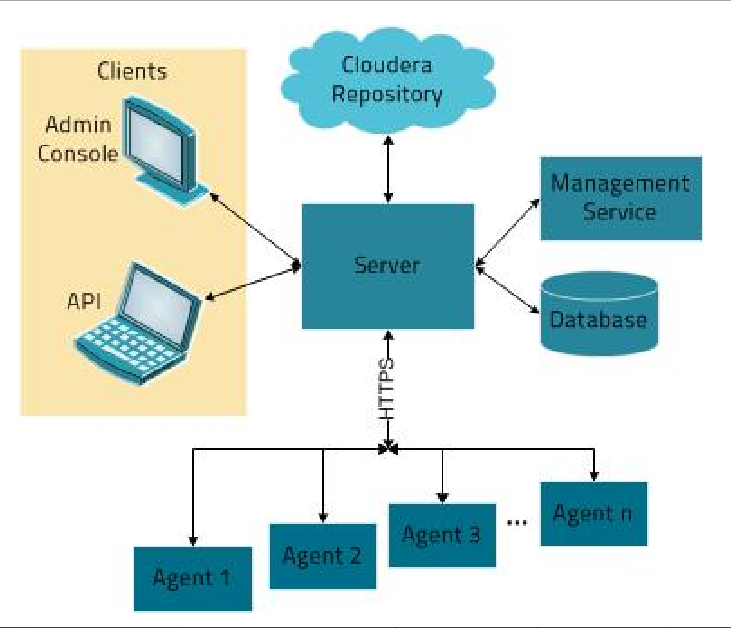
### ClouderaManager应用场景

* 适用于节点在5个以上的集群，小公司用到的服务较少时，为了节省服务器等资源，不需要部署cm。
* 适用于所有的专业大数据公司，这类企业的硬件资源一般都比较充足。
* 适用于运维工作较频繁的场景，使用apache版本的运维人员，对某一个组件进行调优配置，需要消耗半天的时间进行调整，效率极低；该平台安装好以后，维护工作相对来将就轻松许多。

补充：

* cm在国内用户量很大，戴尔、一号店等知名公司都在使用
* cm在主流的大数据平台框架中，用户量比例很高
* cm的免费版本不支持弹性升级。

### ClouderaManager架构



1. Server：Cloudera Manager的核心是Cloudera Manager Server。提供了统一的UI和API方便用户和集群上的CDH以及其它服务进行交互，能够安装配置CDH和其相关的服务软件，启动停止服务，维护集群中各个节点服务器以及上面运行的进程。
2. Agent：安装在每台主机上的代理服务。它负责启动和停止进程，解压缩配置，触发安装和监控主机
3. Management Service：执行各种监控、报警和报告功能的一组角色的服务
4. Database：CM自身使用的数据库，存储配置和监控信息
5. Cloudera Repository：云端存储库，提供可供Cloudera Manager分配的软件
6. Client：用于与服务器进行交互的接口，包含Admin Console和API
   1. Admin Console：管理员可视化控制台
   2. API：开发人员使用API可以创建自定义的Cloudera Manager应用程序

### ClouderaManager功能

#### 信号检测

默认情况下，Agent 每隔 15 秒向 Cloudera Manager Server 发送一次检测信号。但是，为了减少用户延迟，在状态变化时会提高频率。

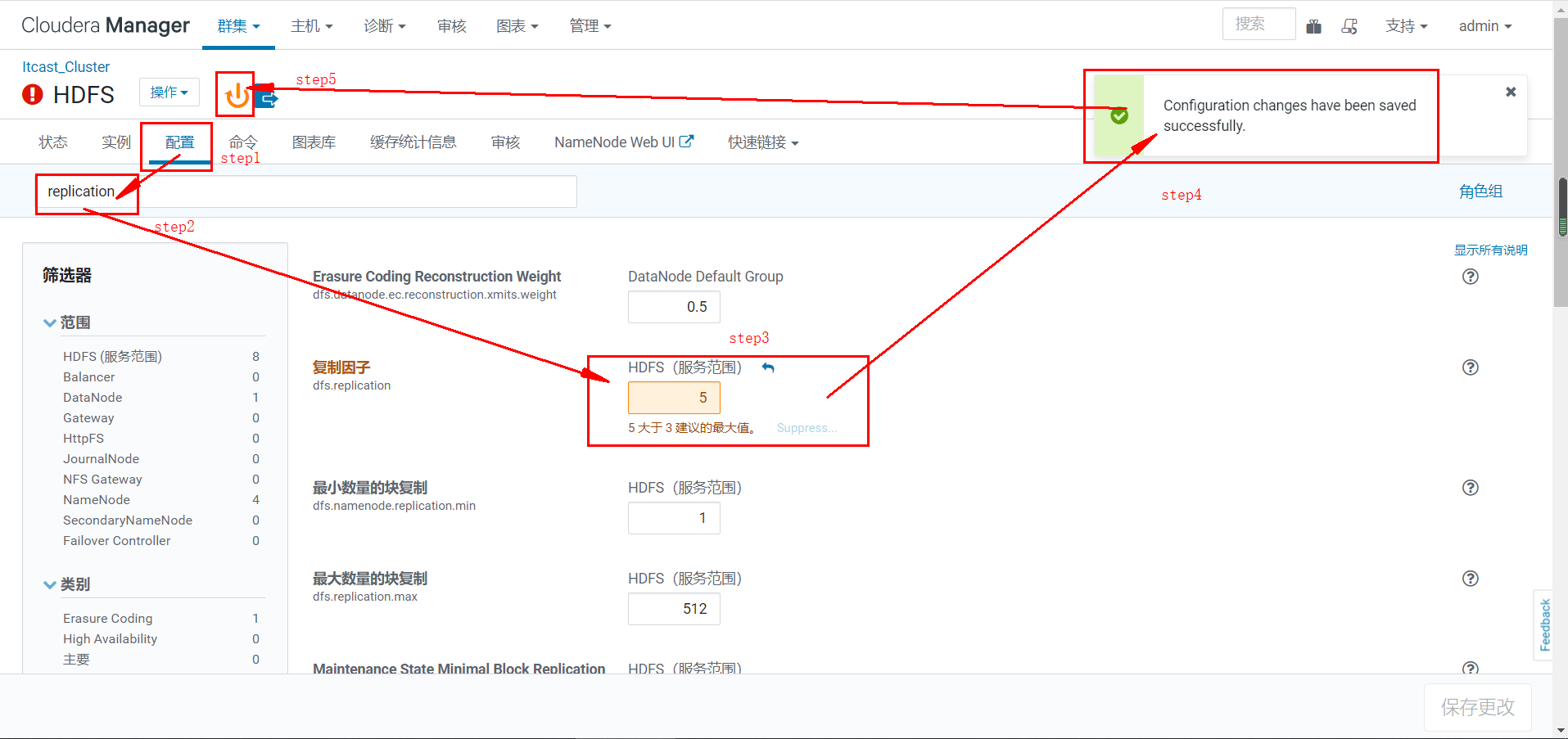


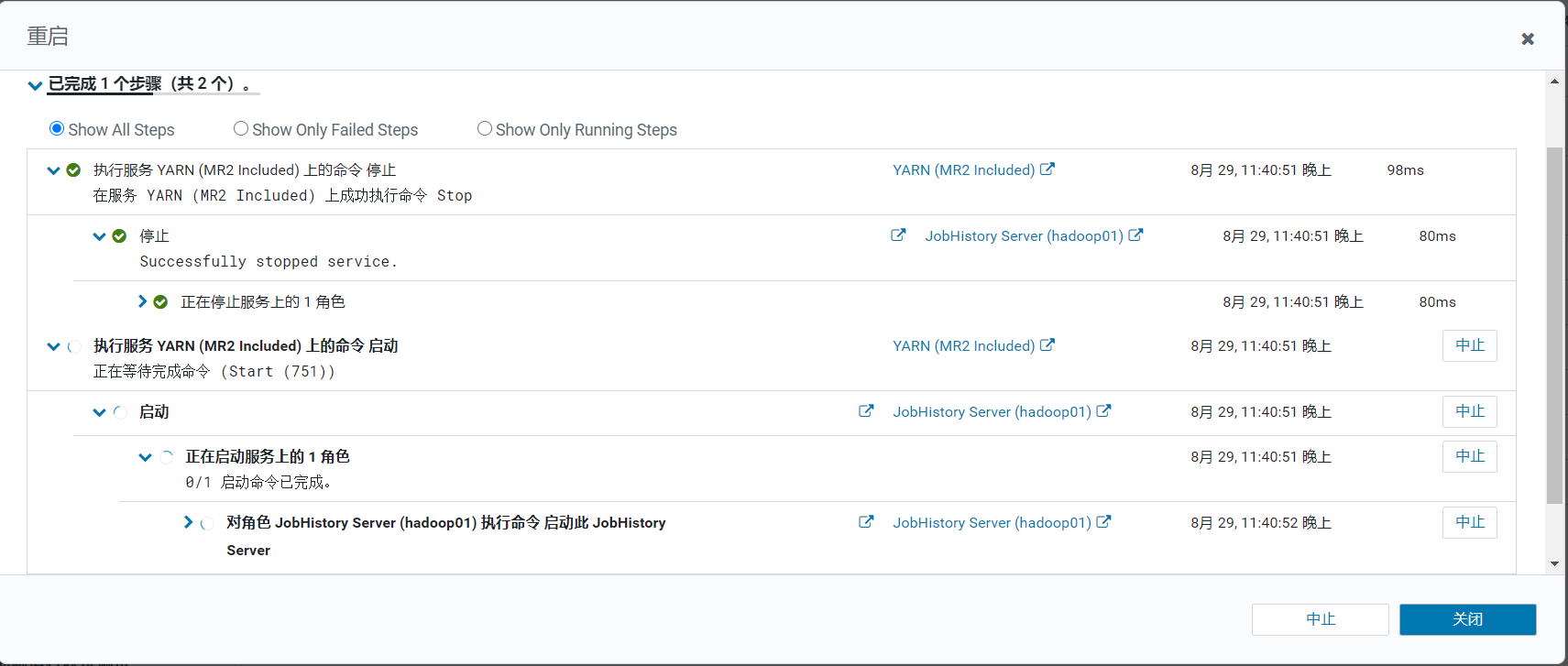
#### 状态管理

模型状态捕获什么进程应在何处运行以及具有什么配置。

运行时状态是哪些进程正在何处运行以及正在执行哪些命令（例如：重新平衡HDFS或执行备份/灾难恢复计划或集群升级、停止）。

当更新配置（例如Hue Server Web 端口）时，相当于更新了模型状态。但是，如果 Hue 在更新时正在运行，则它仍将使用旧端口。当出现这种不匹配情况时，角色服务会标记为“过时的配置”。要重新同步，需重启角色服务（这会触发重新生成配置和重启进程）。





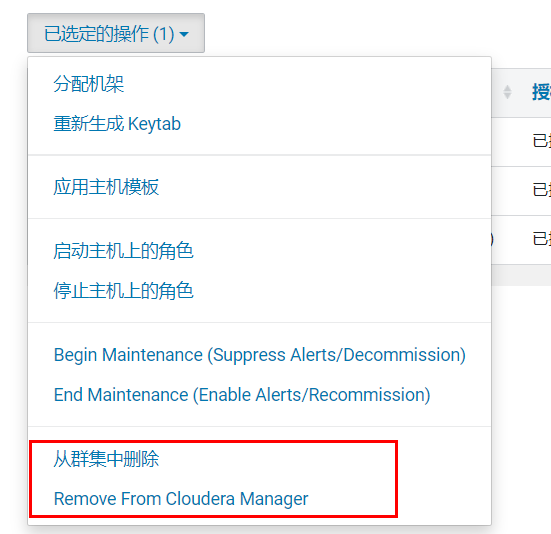
#### 主机管理

Cloudera Manager 作为群集中的托管主机身份，可对JDK、Cloudera Manager Agent、CDH、Impala、Solr等所有软件角色的主机进行管理。

Cloudera Manager 提供添加和删除主机的操作。

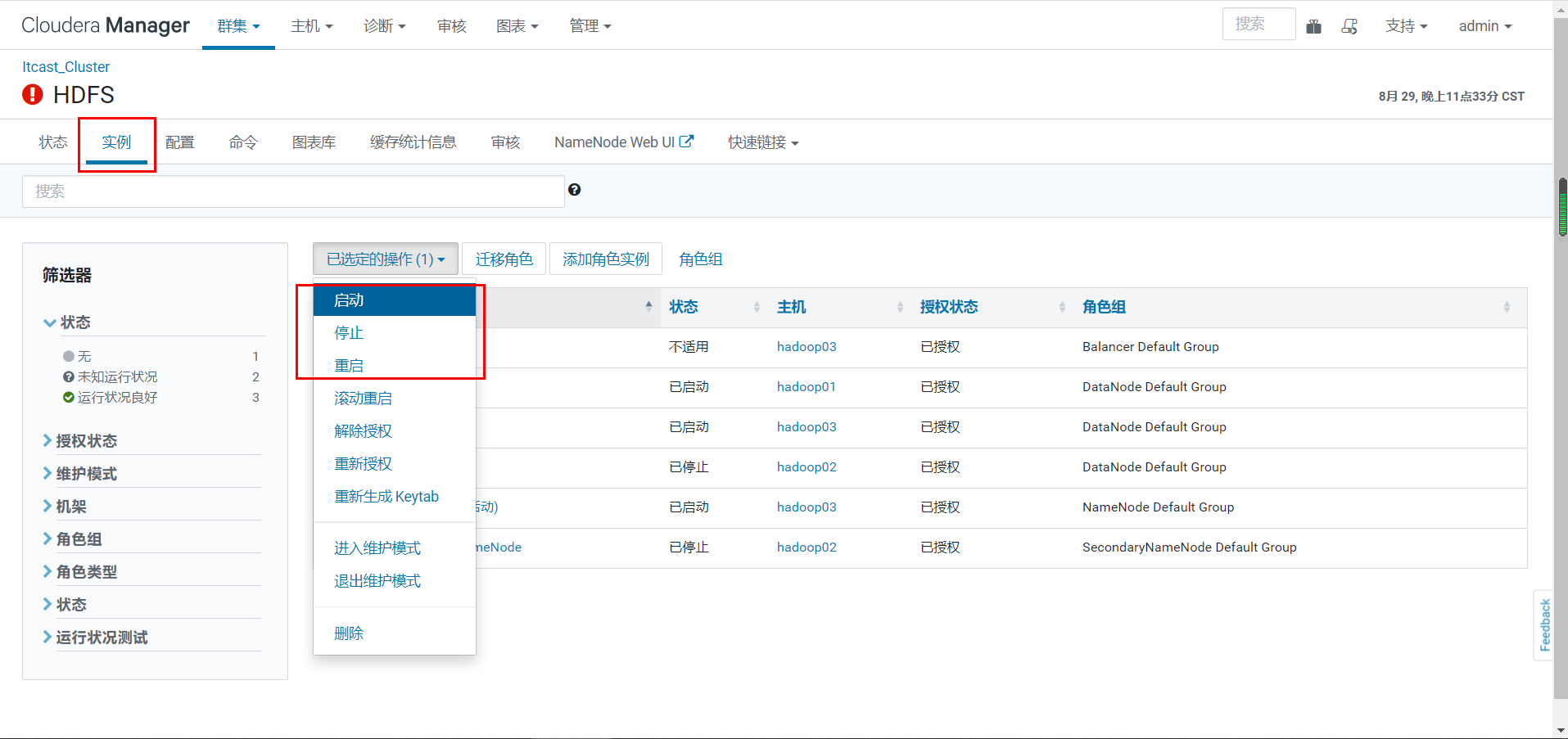
Cloudera Management Service Host Monitor 角色执行状况检查并收集主机度量，可以监控主机的运行状况和性能。





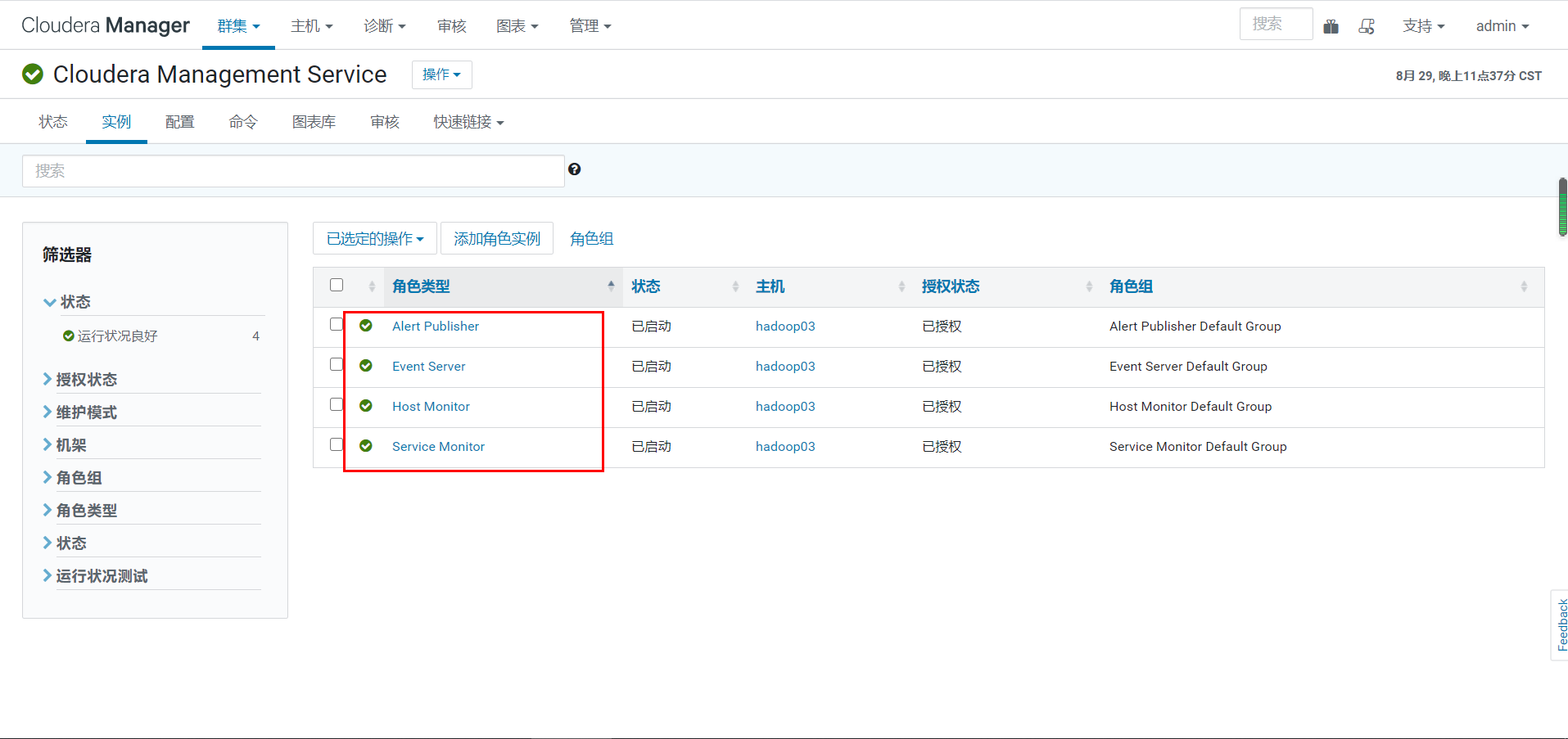
#### 进程启停

在Cloudera Manager管理的群集中，只能通过 Cloudera Manager 启动或停止服务。Cloudera Manager 支持自动重启崩溃进程。如果一个角色实例在启动后反复失败，Cloudera Manager 还会用不良状态标记该实例。



特别需要注意的是，停止 Cloudera Manager 和 Cloudera Manager Agent 不会停止群集；所有正在运行的实例都将保持运行。

#### 监控管理



Activity Monitor：收集关于MapReduce服务运行的活动的信息。默认情况下不添加此角色。

Host Monitor：收集有关主机的运行状况和指标信息。

Service Monitor：从YARN服务中收集关于服务和活动信息的健康和度量信息。

Event Server：聚合组件的事件并将其用于警报和搜索。

Alert Publisher ：为特定类型的事件生成和提供警报。

Reports Manager：生成图表报告。

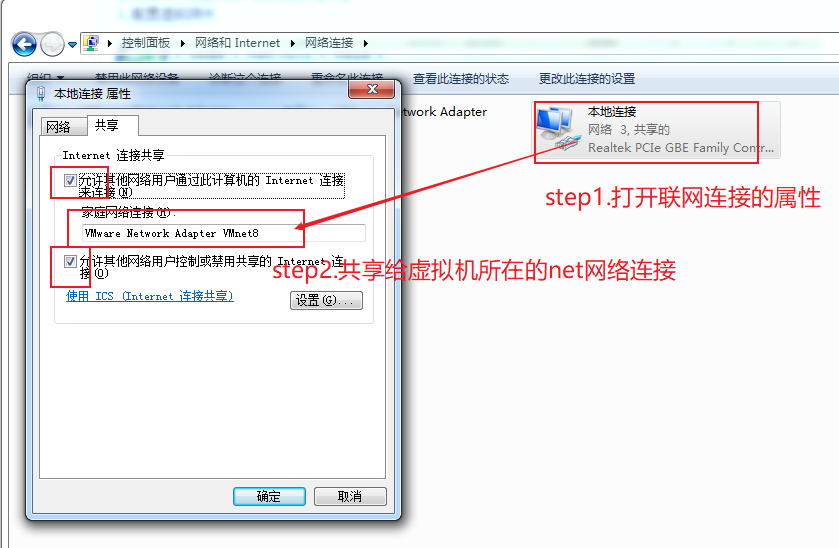
## 使用Cloudera Manager虚拟机

|  |  |  |  |
| --- | --- | --- | --- |
| 节点 | 内存 | CPU | 硬盘 |
| hadoop01 | 12G | 4 | 50 |
| hadoop02 | 3G | 2 | 50 |
| hadoop03 | 1G | 1 | 50 |

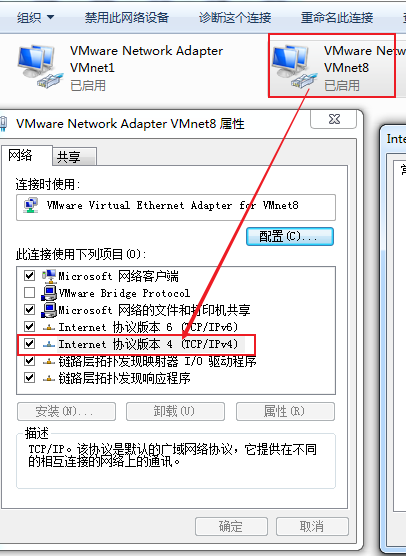
压缩包在【Home\资料\三台环境】目录下，使用步骤：

### 先设置win10网络配置

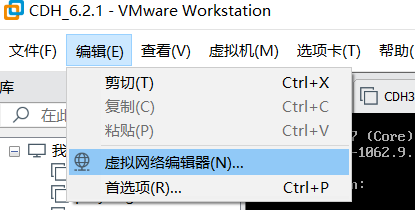
#### 设置网络共享

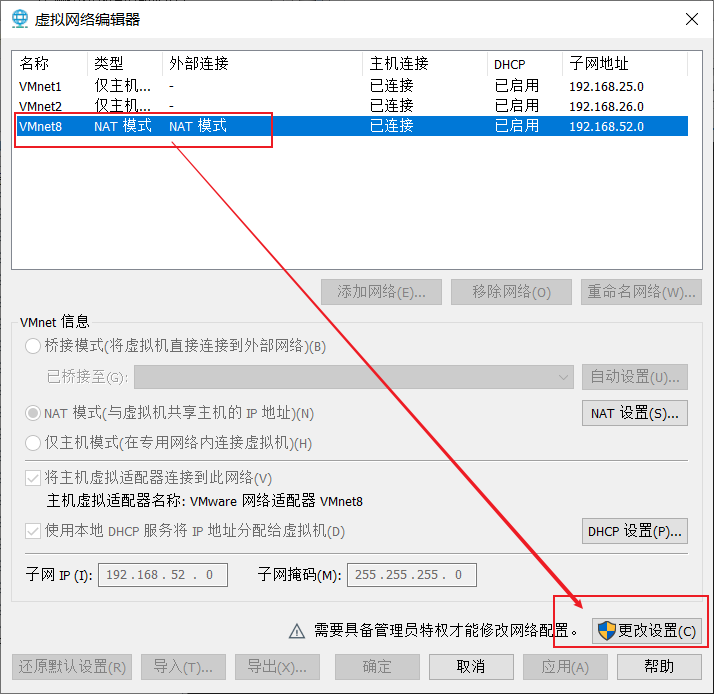


#### 设置VM8网络连接的IP



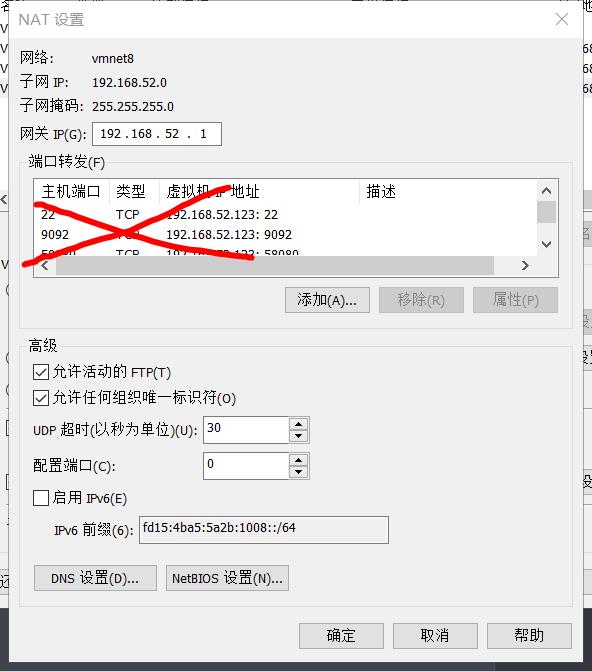
### 设置VMware虚拟机网络











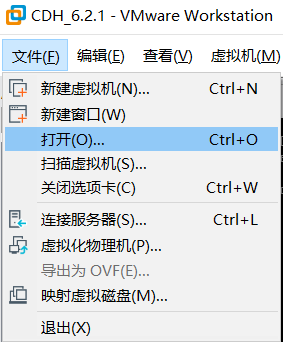
### 设置时钟同步

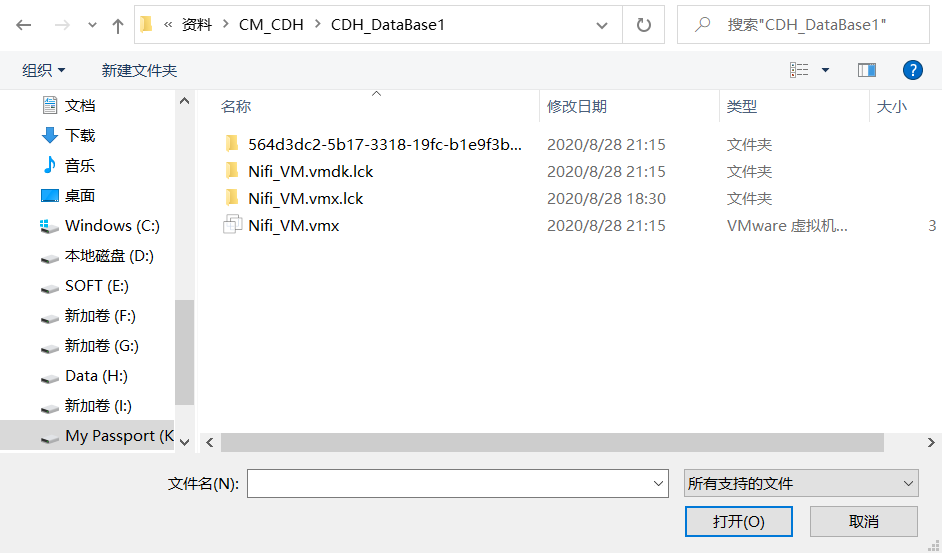
详见【Home\讲义\第1章 项目介绍\时钟同步】



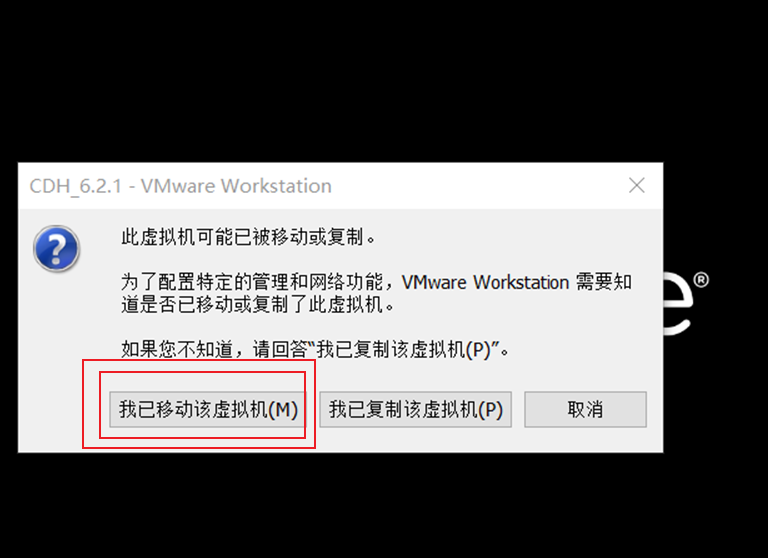
### 打开虚拟机

解压后，直接使用VMware打开三台虚拟机即可。





**注意**如果出现**【我已移动】**或【我已复制】，不能默认，必须选择【我已移动】，否则需要重新解压并启动。



### 关机

使用完毕后，通过[**shutdown -h now**]命令来关闭服务器，不要挂起或强行断电。

