阿里云大数据实验室

解决方案

目 录

第 1 章 建设背景 1

1.1 政策背景 1

1.2 产业需求 2

1.3 阿里云大数据介绍 3

1.4 阿里云大数据实验室介绍 6

第 2 章 可行性和必要性分析 8

第 3 章 大数据实验室建设内容 12

3.1 后台支撑平台 12

3.2 各种实践课 19

3.3 双创功能 38

3.4 实验室装修参考（院校自建） 39

第 4 章 阿里云大数据平台 42

4.1 数加 42

4.2 阿里巴巴专有云 APSARA STACK 42

第 5 章 开源计算平台 44

5.1 CLOUDERA CDH 44

5.1 TENSORFLOW 47

第 6 章 免费体验：数加体验馆 48

第 7 章 教育部阿里云首批 9 本教材出版 50

第 8 章 成功案例 52

阿里云大数据实验室

目 录

第1章 建设背景

1.1 政策背景

2015 年中央政府工作报告提出制定“互联网+”行动计划，大数据作为“互联网+”行动计划的重要组成部分，已成为新一代信息技术变革的核心。在工作报告中提出要全面鼓励技术创新，保护发明创造，同时还提出企业是技术创新的主体，鼓励和支持企业主导的产学研协同创新，大力发展众创空间，增设国家自主创新示范区，办好国家高新区，发挥集聚创新要素的领头羊作用。

同时，国务院还对外发布了《促进大数据发展行动纲要》，明确指出，信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已成为国家基础性战略资源，大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。目前，我国在大数据发展和应用方面已具备一定基础，拥有市场优势和发展潜力，但也存在创新应用领域不广等问题，亟待解决。

为推动我国大数据产业持续健康发展，实施国家大数据战略，落实国务院《促进大数据发展行动纲要》，按照《中华人民共和国国民经济和社会发展第十三个五年规划纲要》总体部署，2017 年 1 月，工业和信息化部正式发布了《大数据产业发展规划（2016 - 2020 年）》（工信部规〔2016〕412 号，以下简称《规划》）。

2016 年 2 月，教育部公布新增的“数据科学与大数据技术”专业（代码 080910T），北京大学、对外经济贸易大学、中南大学成为首家获批高校。时隔一

年，2017 年 3 月，教育部公布第二批“数据科学与大数据技术”专业获批的 32 所高校。2018 年 3 月，教育部公布第三批“数据科学与大数据技术”专业获批的 250 所高校。截止到当前，我国已有 285 所高校获批该专业。该专业强调培养具有多学科交叉能力的大数据人才，重点培养具有以下三方面素质的人才：一是理论性的，主要是对数据科学中模型的理解和运用；二是实践性的，主要是处理数据的能力。三是应用性的，主要是利用大数据的方法解决具体行业应用问题的能力。

2016 年 9 月，教育部公布新增“大数据技术与应用”专科专业（代码 610215）。 该专业强调培养具有大数据实践能力的大数据人才，重点培养具有以下两方面素质的人才：一是工具的掌握，掌握数据采集和数据分析的基本工具；二是数据分析能力，掌握实用数据分析和初步数据建模能力。

1.2 产业需求

根据 IDC 报告，2019 年全球大数据市场规模将达到 1250 亿美金，中国在全球大数据市场占比将超过 8%（超过 650 亿人民币）。考虑到和大数据相关的行业软件、解决方案、服务和硬件，整个大数据市场规模在 2019 年将达到几千亿的规模，每年都在以非常快的速度递增。

根据麦肯锡全球研究院的报告，中国未来 3-5 年内需要有 **180 万人**从事大数据相关的岗位，目前大约有 **150 万**人才缺口。各个行业都急需大数据人才，特别是熟悉大数据采集加工处理和深度学习建模的大数据工程人才：大数据工程师；熟悉行业知识和大数据的交叉复合型人才：大数据分析师；以及熟悉分布式、多线程和海量数据的大数据 JAVA 全栈工程师。

1.3 阿里云大数据介绍

阿里云创立于 2009 年，为阿里巴巴集团旗下云计算和大数据品牌，是中国最大的云计算平台，为全球 200 多个国家和地区的创新创业企业、政府机构等提供服务。作为全球领先的大数据和云计算基础软件供应商，Alibaba cloud 目前已位列全球第三，与 AWS（亚马逊）、Azure(微软)初步形成了“3A”的竞争格局。目前，阿里云已在国内市场占据领先地位。IDC 发布的数据显示，截止到 **2017 年上半年，阿里云 IaaS 占据了 47.6%的中国市场份额，是市场第二名的 5 倍。**

从 2010 年开始，阿里云正式对外提供云计算商业服务，希望能够帮助更多的中小企业、金融、科研机构、政府部门、中央企业、大型民营企业，实现计算资源的“互联网化”和大数据的广泛应用。目前已经建立遍布全球的超大规模数据中心，如华北、华东、华南、香港、美东、美西、新加坡和日本等多个数据中心群。截至 2016 年 3 月 31 号，阿里云服务的客户数量已超过 230 万，是国内云服务种类最多、用户数最多、商业运营最久的公共云服务提供商。

大数据时代，云计算成为了经济社会发展的基础设施。飞天（Apsara）是阿里云自主研发，服务全球的超大规模通用计算操作系统。它可以将遍布全球的百万级服务器连成一台超级计算机，以在线公共服务的方式为社会提供源源不断的计算能力。从 2009 年诞生第一天起，飞天就希望解决人类计算的规模、效率和安全问题。

飞天研发时市面上没有任何开源软件能够满足它的设计目标，因此也就坚定地走上了自主研发的道路。7 年发展，飞天早已不局限于服务阿里巴巴集团内部业务，而是成长为中国自主研发、服务全球的一个操作系统。正是通过一次次成果的落地与实践，飞天不断进化，演变，也让阿里云变为一家全球领先的云计算、人工智能技术公司，其产品不断地给世界创造价值。

我们在生活中接触过飞天吗？有！比如，12306 将车票查询业务部署在飞天上，让飞天在春运高峰 75%的流量中学会了如何应对大规模并发；《小门神》选择在飞天渲染，让它拥有了同时调度 3 万核计算资源的工作经验。阿里云上百款产品，阿里云云市场上几千款应用都跑在飞天操作系统上。目前，数千名阿里员工致力于研究飞天，研究飞天之上的人工智能，在理论和实践中不断进行修正和优化。2015 年，Sort Benchmark 公布 2015 年排序竞赛最终成绩：飞天用 377 秒就完成了 100TB 的数据排序，打破四项世界纪录。

在计算操作系统层面的突破让阿里云成为一张走出国门的技术新名片。目前，飞天这项中国自主可控的技术服务为全球 200 多个国家和地区的创新创业企业、政府、机构等提供服务，行业遍及政务、游戏、金融、电商、移动、医疗、多媒体、物联网、O2O 等，计算经济的成果初现。

云计算是大数据的基础。面向大数据生态建设，阿里在以往的合作伙伴扶持中总结出一套清晰的数据生态路线图，2015 年已经启动了该计划并且扶持了一批数据合作伙伴和创业公司。目前，阿里云上已汇聚了 50 多个行业解决方案，这些解决方案主要由 200 余家大型合作伙伴提供。未来三年，阿里云计划将这一数字提升到 2000 家，可提供的解决方案也将实现数十倍增长。

2015 年 12 月，阿里启动了全球数据人才认证计划，全面启动大学院校的数据合作，并向天池大数据竞赛的前 20 名选手颁发首批阿里云大数据应用人才认证（ACP）。

阿里云目前在国内云计算市场占有率排名第一，超过亚马逊、微软和 IBM 在中国市场的份额总和。阿里云上众多生态企业，对人才有着极高需求，而阿里云大数据实验室基于阿里云大数据公共云平台“数加”平台，提供了完整的大数据教学实训、科研创新所需的计算能力和工具支持，为大数据人才的培养、科研创新和双创大赛提供了方便、可靠的服务平台。

2016 年 5 月，阿里巴巴集团成为了国家首批双创示范基地。2017 年 2 月，国家发改委正式公布大数据国家工程实验室名单中，由阿里云参与的“工业大数据应用技术国家工程实验室”和“大数据系统软件国家工程实验室”两个大数据实验室均获批复认定，它们分别是工业大数据应用及大数据系统软件领域的唯一国家级工程实验室。

用户通过与阿里云合作，可以打造适应 DT 时代的大数据实验室，全面培养与企业用人需求接轨的大数据人才，打造贯穿产学研全链路的科研创新平台，举办各种创业创新大赛。

1.4 阿里云大数据实验室介绍

阿里云大数据实验室是指：由阿里云开发并拥有知识产权的、旨在帮助大数据从业人员和学生更好地了解和学习大数据产品以及各种大数据真实案例，快速进行大数据科研创新的软件产品包。阿里云大数据实验室作为一站式大数据实训和科研创新平台，为各行业用户提供简单易用的大数据真实环境，让数据价值触手可及。用户可通过简单快捷的可视化操作，对各种大数据进行数据采集、数据加工、数据开发、数据管理、数据分析和机器学习等操作，快速探索各种大数据创新应用。

阿里云大数据实验室主要由面向大数据实训的大数据基础课程体系和行业案例课程体系、面向大数据科研创新的科研组件、大数据实验室后台支撑平台等组成。阿里云大数据实验室依赖阿里云大数据公共云平台“数加”或者阿里巴巴大数据专有云平台。

阿里云大数据实验室主要包含以下功能：

1、各种大数据实训课程（面向教学）：

1）在阿里云大数据真实环境中（非模拟或仿真环境）手把手教学生学习各种大数据产品的基础知识，学习过程中既支持 100%图形化拖拉拽操作，也支持各种 API 开发管理。具体课程包括大数据基础、大数据基础实践、离线分布式平台高级开发和管理、深度学习高级开发和管理等。

2）手把手教学生学习各个行业大数据应用创新真实案例（脱敏数据），学生在课程中可以学习大数据创新过程中的各种成功经验和失败教训。具体课程包括电商广告精准营销实践、税务纳税评估实践、银行行业风险信用模型实践和电信行业用户流失分析实践等，另外还包含了 11 门免费的行业案例课程。

2、面向科研创新环节的大数据科研创新平台：

在阿里云大数据平台的真实环境中（阿里云大数据平台“数加”），老师和学生可以进行各种大数据应用的创新研究。在研究过程中，可以充分借鉴各个行业的成功经验，加速大数据应用科学研究过程。

第2章 可行性和必要性分析

阿里云大数据实验室全面依托阿里巴巴大数据、人工智能和云计算核心技术和行业经验，底层计算平台全面基于阿里云大数据公共云平台“数加”或者阿里巴巴大数据专有云平台。通过使用阿里云大数据公共云平台“数加”，大量节省了 IT 硬件投入以及运维费用；底层计算平台也可以使用阿里云大数据平台专有云版本，在客户机房中独立部署阿里巴巴飞天操作系统、离线分布式计算平台 MaxCompute、数据集成工具 DataWorks、阿里巴巴机器学习工具 PAI 等。用户还可以采用混合云部署方式，公共云和专有云相结合，既满足了公共云方式节省成本的需求，也满足了搭建硬件服务器存放私有数据的需求。专有云还支持企业级开源 Hadoop 平台（Cloudera CDH）和谷歌机器学习开源工具 TensorFlow。

1. 完善的产品培训让您快速上手大数据和人工智能

1) 在阿里云大数据真实环境中（非模拟或仿真环境）学习大数据

2) 手把手教您大数据采集、加工、处理、BI 报表和深度学习等知识

3) 学习过程中既支持 100%图形化拖拉拽操作，也支持各种 API 开发管理

2. 各个行业真实案例让您深入理解大数据和人工智能

1) 基于全球各个行业大数据真实案例（脱敏数据）进行大数据实训教学

2) 通过案例，学习各种深度学习建模和调优方法

3) 通过案例，学习各个行业的大数据和人工智能创新成功经验

4) 借鉴各种案例内容，快速开展自己的大数据应用科研创新

3. 提供各种课题深度合作、教材联合编写和校外老师培养基地

1）对高校专长的领域进行课题深度合作，一起进行大数据科研创新

2）我司提供各种技术素材，和高校一起进行大数据教材编写

3）为高校提供“校外教师培养基地”，全面提升老师的大数据专业能力

4. 阿里云大数据实验室方案具有明显的成本优势

阿里云大数据实验室软件和服务包具有非常明显的成本优势，学校采用传统方案搭建一个大数据实验室时，不仅要自己采购服务器硬件，还要购买系统软件、离线分布式平台、ETL 工具、数据开发工具、BI 报表工具和深度学习工具，要自己搭建整个基础软件环境，要自己翻译教程/购买第三方教程/自己做课件，自己购买管理软件等一系列流程。整体投资需要 1000 万以上才能起步。而采用阿里云大数据实验室构建方案，学校无需购买服务器硬件，通过一站式的大数据实训和科研平台，直接和学号系统对接，学校可以直接基于平台教学资源进行开课，同时成功地将成本降低到 100 万级别。

5. 公共云资源服务大量节省 IT 投资

目前国际上公认，每 1 元钱的云计算投资能够节省 6 元钱的传统 IT 投资。阿里云大数据实验室通过将底层计算资源部署在云端，帮助用户大量节省了 IT 投资。阿里云大数据实验室依赖底层阿里云大数据公共云平台“数加”平台，客户可以将钱直接充入到自己的阿里云官方企业账号中，客户当年没有使用完的流量费，会自动结转到下一年使用。

6. 阿里云大数据平台在业内持续领先

阿里云大数据实验室底层依托于阿里云大数据公共云平台“数加”平台。阿里云数加“大数据开发平台”是阿里巴巴集团推出的大数据领域平台级产品，提供一站式大数据开发、管理、分析、挖掘、共享、交换等端到端的解决方案，利用MaxCompute（原 ODPS）在几分钟内可将原始数据转变为业务洞察的海量数据处理能力，而无需关心集群的搭建和运维。

1) 拥有完全自主研发的大数据离线分布式计算平台，该平台具有以下功能：

a) 高效处理海量数据

- ✓ 单一物理集群规模可以达到 10000+服务器（保持 80%线性扩展）,并有实际案例
- ✓ 单个集群部署可以支持 100 万服务器以上，支持同城、异地多数据中心模式，并有实际案例

b) 安全性

- ✓ 所有计算在沙箱中运行
- ✓ 多种权限管理方式、灵活数据访问控制策略
- ✓ 数据存储多份拷贝

c) 易用性

- ✓ 开箱即用
- ✓ 支持 SQL、MR、Graph、流计算等多种计算框架
- ✓ 提供丰富的机器学习算法库

d) 整个平台经过实践验证

e) 自主可控，完全自主研发

2) 拥有分布式的分析数据库，该平台具有以下功能：

a) 支持高并发的海量结构化数据实时查询

- ✓ 支持标准 SQL 语法，个性化的统计分析函数和 LBI 相关函数,提供智能的 UDF 和数据结构，满足业务需要
- ✓ 支持十万量级 TPS 数据实时插入，千万量级数据秒级导出

- ✓ 支持千亿级数十 TB 单表的计算
- ✓ 支持对任意字段进行组合查询，支持丰富的逻辑条件

b) 性能

- ✓ 支持 join 表顺序自动选择、索引选择、数据预排序等丰富的优化策略
- ✓ 全字段自动建立索引，高效的索引实现和内存换入换出，最大限度提高查询性能

c) 全面兼容 MySQL 协议，支持 ODPS、RDS、OSS 等数据源

d) 自动化的智能 CBO 和全索引技术最大限度提高查询性能

e) 支持数据实时更新，海量数据快速导出

f) 角色分离，基于 ACL 的权限控制，支持列级授权，保证数据安全

g) 提供图形化的管理控制台，提升用户体验

3) 提供基于公共云的数据处理服务，覆盖数据采集、数据治理、离线分布式计算、BI 分析和机器学习等功能；实验室提供完善的科研功能，科研所使用的公共云资源可以按照实际使用量进行计费。

7. 实验室部署简单易行

1) 支持快速部署

2) 免费提供阿里云大数据实验室的宣传展板和培训教室装修风格

3) 免费提供面向老师的大数据培训

8. 支持混合云架构

专有云环境既可以使用阿里巴巴专有云， 也可以引 入企业级开源 Hadoop 平台（Cloudera CDH）和谷歌机器学习开源工具 TensorFlow。

第3章 大数据实验室建设内容

3.1 后台支撑平台

实验室后台支撑平台主要实现整个实验室的底层架构搭建，完成实训/科研实际操作环境与底层计算集群的衔接，负责整个实验室平台的安全管理、资源调度、权限管控，完成对实训环境中的实验、课程包、软件包和数据包管理等。

3.1.1 实验室管理员管理模块

大数据实验室管理员负责模块，主要是对整个组织及成员、存储信息的保存，具体包括实验室管理（实验室信息、公告管理、banner 信息管理）、人员管理（老师管理和学员管理）、实验课管理（已购买实践课、已分配实践课、结束实践课）、班级管理（班级列表、上传班级）、排课管理（上传/取消/查看排课计划），具体功能包括：

3.1.1.1 实验室管理

1. 实验室信息：可修改实验室信息 logo、介绍（图片上传）实验室名称、实验室描述

2. 公告管理：显示已经发布的所有公告信息（公告内容、发布时间），可以对已经发布的公告进行“删除”和“编辑”，实验室首页公告编辑如：点击“创建公告”按钮，跳出“文本编辑器”窗口，内容填写完成，点击“发布公告”按钮。

3. banner 信息管理：首页 banner 图编辑

3.1.1.2 人员管理

1. 学员管理：

   1) 学员搜索:通过“账号”、“用户名”、“邮箱”“手机”搜索出学员的信息。包含学员个人账号信息及相关的课程信息

   2) 学员列表:包含学员个人账号信息及相关的课程信息

   3) 学员详情:包括个人信息、参与实践课、最优成绩、提交历史

   4) 启用/停用学员账号:启用/停用学员账号

   5) 学员账号删除:具备删除学员账号的能力

2. 教师管理：

   1) 教师搜索:通过“账号”、“用户名”、“邮箱”“手机”搜索出老师的信息。包括教师个人信息，所开实践课，实践课学员等相关概况信息

   2) 教师列表:查看教师个人信息列表

   3) 教师详情:包括教师个人信息，所开实践课，实践课学员等相关概况信息

   4) 教师导入:上传 excel，导入教师清单，邮件发送用户账户信息。

   5) 启用/停用教师账号:停用教师账号即为禁止登陆，启用后才能继续登陆

6) 教师账号删除:具备教师账号删除的能力

7) 重置教师密码：可以重置教师密码

3.1.1.3 实验课管理

1. 已购买实践课：查看目前已经购买的所有实践课程列表（实践编号、实践名称、购买时间、操作）点击“操作”按钮可以查看当前实践课的详细信息（“实践属性”、“实践介绍”、“常见问题”、“实践数据”、“实践阶段”）；“批量分配课程”:将当前实践课程批量分配给教师，可多次分配。

2. 已分配实践课：查看目前所有实践课列表，包括所有已分配给教师的实践课。（实践编号、实践名称、教师、账号、学员数、实践状态、操作）；可以通过“实践课名称”或“教师名称”搜索相关的课程信息。

3. 结束实践课：可结束掉目前已发布的实践课

3.1.1.4 个人中心

1. 个人信息修改：管理员可修改自己相关基本信息（头像、手机号、性别、职称和个人介绍等）

2. 退出：退出当前账号

3.1.2 教师管理模块

本模块主要完成老师在实训过程中所需的各种管理功能，具体包括实验列表、实验数据、教学文档、实验包括、课程测验、实验问题、考勤管理、常见问题、学生管理、统计报表/学生成绩汇总/考勤报表等，具体包括：

3.1.2.1 我的实践课

1. 实验列表：显示当前实验课程中的所有实验，每个实验都具有“开始实验”和“结束实验”功能

2. 实验数据：显示当前实践课程中的所有实践数据列表，每个实践数据的具体信息包括“格式”、“实验数据名称”、“下载次数”、“更新时间”、“描述信息”、“点击下载”等内容

3. 教学文档：显示当前实践课程中的所有实验所对应的实验手册列表，每个实验手册的具体信息包括“格式”、“实验手册名称”、“查看次数”、“更新时间”、“描述信息”、“查看文档”等内容

4. 实验报告：显示当前实践课程中的实验报告列表，每个实验报告的具体信息包括“实验报告名称”、“更新时间”等内容

5. 课程测验：显示已经下发的考试课程列表，具备创建题库和下发考试功能。

6. 实验问答：显示当前实验课程中所有学员提交的实验问题，如：点击“我要回答”按钮，跳出“文本编辑器”窗口，内容填写完成，点击“提交”按钮。

7. 考勤管理：教师可以对学员进行考勤管理包含序号 学号 姓名 班级 已到 请假 旷课 迟到 早退等信息，支持对考勤列表的导出功能

8. 常见问题：显示当前实验课程中的所有常见问题。

9. 学员管理：

1) 学员搜索：包含学员个人账号信息及相关的课程信息

2) 查看学员：教师可查看某实践课下所有参与学员信息列表

3) 添加学员：对已经加入排课计划中实践课，可导入学员名单 excel 到该实践课中，邮件发送用户账户信息。

4) 移除学员：教师可对目前存在在实践课里的学员进行移除

5) 学习记录：提供每个学生在学习中心平台内学习课程的学习记录，包含日期、计划、课程、章节、课程完成情况。

6) 学员分组：教师可以根据学员人数进行分组操作

7) 成绩自动打分：教师可以根据学员的作业完成情概况进行自动打分

8) 成绩手动打分：教师可以根据学员的平时表现进行手动打分

9) 平时成绩管理：对每阶段成绩列表可以进行总览查看及细化查看，了解该阶段下学员提交明细

10) 导出成绩管理：教师可进行成绩导出

11) 重置学员密码：将学员的密码进行重置

3.1.2.2 统计报表

部分功能如下：

1. 学生成绩汇总：统计学生成绩支持导出功能

2. 考勤报表：统计学生的考勤状况以报表形式展现

3.1.2.3 个人中心

1. 个人信息修改：教师可修改自己相关基本信息（头像、手机号、性别、职称和个人介绍等）

2. 密码修改：教师在界面上可修改自己密码

3. 我的科研：教师可以在自己的科研环境中进行自己的科研任务开发。（包含添加科研小组成员等）

3.1.3 学生管理模块

主要完成学员在实训过程中所需的各种管理功能，具体包括实验列表、实验数据、教学文档、实验报告、课程测验、小组成员、学习成绩、实验文档、常见问题、个人信息、考勤、消息、密码、统计报表等，具体包括以下功能：

3.1.3.1 我的实践课

1. 实验列表：显示当前实验课程中的所有实验，每个实验都具有“开始实验”和“结束实验”功能

2. 实验数据：显示当前实践课程中的所有实践数据列表，每个实践数据的具体信息包括“格式”、“实验数据名称”、“下载次数”、“更新时间”、“描述信息”、“点击下载”等内容

3. 教学文档：显示当前实践课程中的所有实验所对应的实验手册列表，每个实验手册的具体信息包括“格式”、“实验手册名称”、“查看次数”、“更新时间”、“描述信息”、“查看文档”等内容

4. 实验报告：显示当前实践课程中的实验报告列表，每个实验报告的具体信息包括“实验报告名称”、“更新时间”等内容

5. 课程测验：提供对课程相关的习题管理，题型包含单选、多选、判断题等类型，

6. 小组成员：学员可以查看自己小组成员的详细信息（序号 姓名 学号 ）

7. 学习成绩：显示当前实验课程的学习成绩

8. 实验问答：显示当前实验课程中的所有提交的实验问题，如：点击“我要提问”按钮，跳出“文本编辑器”窗口，内容填写完成，点击“提交”按钮。

9. 常见问题：显示当前实验课程中的所有常见问题。

3.1.3.2 个人中心

1. 个人信息：查看、修改个人信息，包括头像、密码、电话、学号、班级等

2. 我的考勤：学员可以查看自己的考勤情况

3. 查看成绩：序号 姓名 学号 学年 课程性质 课程学分 课程名称 授课教师 课程成绩 平时成绩

4. 我的消息：学员可以查看管理员和教师发布的消息。

5. 修改密码：修改实验室密码、数加密码、专有云密码等

6. 退出：退出当前账号

3.1.4 实验室专有云驱动管理

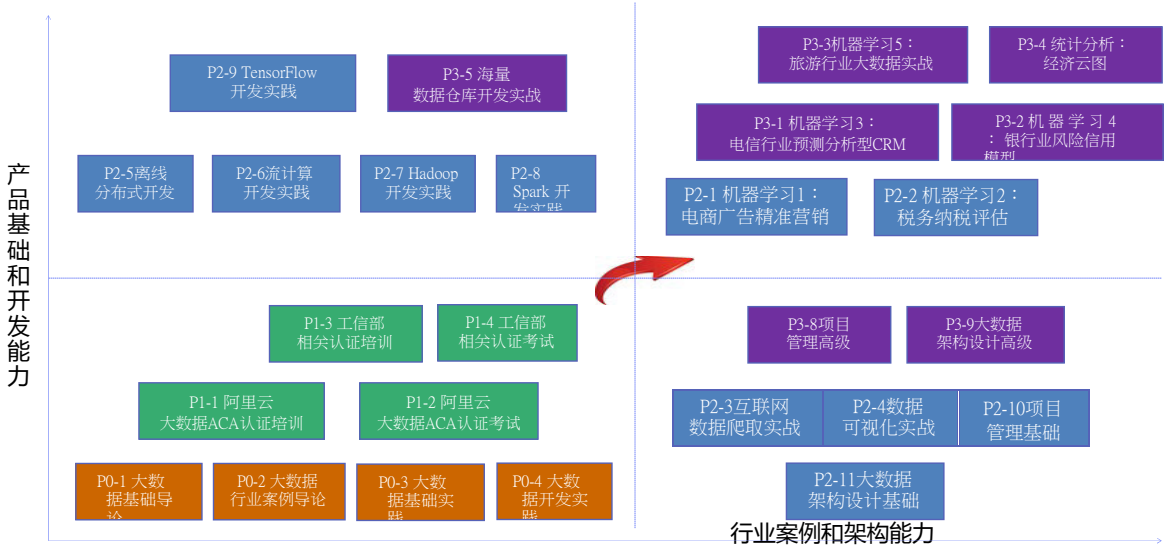
本模块主要完成专有云底层计算资源（具体包括离线分布式计算平台、流计算平台、实时计算平台、图计算平台和机器学习平台等）和大数据实训管理软件的对接，针对每个具体实践，生成多租户实践环境。

1、基于分布式集群管理系统，提供大数据集群管理系统，功能包含云实验应用系统与离线分布式计算系统、实时分布式计算系统、流计算系统、机器学习/深度学习分布式计算系统等对接，可以在每一个实验中打开对应专有云计算平台的逻辑环境，并支持在其中进行各种操作。

2、包含相应专有云平台（离线分布式计算系统、实时分布式计算系统、流计算系统、机器学习/深度学习分布式计算系统等）的实时监控，具体包括实时监控集群的 CPU、内存、硬盘等使用率及相关信息，可以对管理节点、计算节点进行启动、停止等操作管理。

3.2 各种实践课

配套教材为电子版，登录大数据实验室即可查看。



3.2.1 基础课程

产品名称	功能描述	部分目录
大数据基础（32课时）	本课程充分融入产业界的先进理念，结合产业一线实践，全面讲述数据驱动的商业变革，深入地讲解大数据相关理论知识。 在本课程中，学生通过学习可以全面了解大数据概述（基本概念、成功的原则、相关案例）、大数据技术演变历程（数据仓库发展史、分析型框架的演变历程、离线查询的演变历程、相关定理和模型、NoSQL）、Hadoop 2.0介绍（APACHE HADOOP，以及相关组件介绍）、企业Hadoop介绍（IBM BigInsights、Cloudera CDH Hortonworks、华为	第 1 章 大数据学习路线图 1.1 大数据上升为国家战略 1.2 产业需求（就业前景） 1.3 阿里云大数据实验室介绍 1.4 大数据学习路线图 第 2 章 概 述 2.1 什么是数据？ 2.2 什么是大数据 2.3 大数据的特征 2.4 大数据成功的原则 2.4.1 创造业务价值 2.4.2 基于成熟的大数据平台构建解决方案 2.4.3 使用全量数据代替抽样数据 2.4.4 允许不精确，允许混杂性 2.5 阿里云大数据行业案例 2.5.1 内部运维日志管理与分析 2.5.2 阿里云交通行业套牌管理 2.5.3 阿里云实时预判的路况预测系统 2.5.4 人工智能阿里云 ET 解决广州交通拥堵难题 2.6 银行行业示例 2.6.1 信用风险评级 2.6.2 银行反欺诈/发洗钱案例 2.6.3 分析型客户关系管理（ARCM） 第 3 章 大数据技术演变历程 3.1 数据仓库发展史

	FusionInsight）、SPARK介绍、TensorFlow以及相关组件、阿里云大数据平台介绍（数加、大数据云平台交换、MaxCompute、Analytic DB等）。	3.1.1 什么是数据仓库 3.1.2 数据仓库的发展史 3.1.3 企业内容管理（ECM） 3.2 大数据中心的兴起 3.3 大数据技术演变历程 3.3.1 分析型框架的演变历程 3.3.2 离线查询的演变历程 3.3.3 复杂事件处理 (CEP)和流计算 3.4 相关定理和模型 3.4.1 CAP 定理 3.4.2 一致性 3.4.3 ACID 模型 3.4.4 BASE 模型 3.5 NOSQL 3.6 人工智能的发展历程 第 4 章 APACHE HADOOP 以及相关组件 4.1 HADOOP 的优势 4.2 HDFS 4.3 ZOOKEEPER 4.4 AVRO 4.5 APACHE OOZIE WORKFLOW SCHEDULER FOR HADOOP 4.6 LUCENE 4.7 FLUME 4.8 SQOOP 4.9 HCATALOG 4.1 HBASE 4.11 PIG 4.12 HIVE 4.13 JAQL 4.14 ZEPPELIN 4.15 TEZ 4.16 MAHOUT 4.16.1 Apache Mahout 的推荐引擎 4.16.2 Apache Mahout 中的聚类分析框架 4.16.3 Mahout 的分类计算框架 4.17 FLINK 4.17.1 Flink 的特点 4.17.2 Flink 与 SparkStreaming 的差异对比 第 5 章 企业 HADOOP 介绍 5.1 IBM BIGINSIGHTS 5.1.1 Big SQL 3.0 介绍 5.1.2 GPFS File Place Optimizer（GPFS-FPO） 5.2 CLOUDERA CDH 5.2.1 CDH 的优势:

		5.2.2 ClouderaManager 5.2.3 Impala 5.2.4 Hue 5.2.5 Cloudera Search 5.2.6 Solr 5.3 HORTONWORKS 5.4 华为 FUSIONINSIGHT 第 6 章 SPARK 6.1 SPARK 的特点 6.2 SPARK RDD&DATAFRAME 6.2.1 Spark RDD 6.2.2 DataFrame 6.3 SPARK SQL 6.4 SPARK STREAMING 6.5 SPARK MLLIB 6.5.1 Spark MLlib 的特点 6.6 SPARK GRAPHX 6.6.1 Spark Graphx 的特点 第 7 章 TENSORFLOW 7.1 TENSORFLOW 的概述 7.1.1 数据流图 7.2 TENSORFLOW 的发展 7.3 TENSORFLOW 的应用场景 7.3.1 单机多卡 7.3.2 分布式 TensorFlow 7.3.3 Tensorflow On Spark 7.4 TENSORFLOW 的优势 第 8 章 阿里云大数据平台概述 8.1 大数据云平台交换 8.2 数加介绍 8.3 MAXCOMPUTE 8.3.1 概述 8.3.2 大数据开发套件 8.3.3 产品的主要应用场景 8.3.4 产品优势 8.3.5 相关支撑组件 8.4 ANALYTICDB 介绍 8.4.1 概述 8.4.2 主要使用场景 8.4.3 主要功能 第 9 章 参考和引用资料
大数据基础实践（32	手把手地教授学生熟悉阿里云大数据平台的各种基础操作，	第 1 章 概 述 1.1 MAXCOMPUTE 简介 1.2 什么是大数据开发套件 第 2 章 大数据开发套件

课时）	基于一个个的用例（use cases）从数据采集、加工和治理，到数据开发、管理、BI报表和机器学习等全方位培训学生，全面提升学生基于图形化方式构建BI报表和算法建模能力。	2.1 项目空间 2.2 表 2.3 视图 第 3 章 SQL 基 础 3.1 插入数据 3.2 查询数据 第 4 章 数据开发管理示例 4.1 使用向导创建商品目录表（作业） 4.2 使用 DDL 语句创建商品目录表（作业） 4.3 导入本地数据（作业） 4.4 数据表管理：商品目录表 4.5 数据同步 第 5 章 机器学习实践：贷款拖欠预测 5.1 创建基础数据表 5.2 导入基础数据 5.3 贷款拖欠预测建模 第 6 章 机器学习实践：超市商品购买关联分析 6.1 创建基础数据表 6.2 导入基础数据 6.3 超市商品购买关联分析建模 第 7 章 机器学习实践：学生考试成绩预测 7.1 创建基础数据表 7.2 导入基础数据 7.3 学生考试成绩预测建模 第 8 章 机器学习实践：统计分析 8.1 创建基础数据表 8.2 导入基础数据 8.3 统计分析建模 第 9 章 机器学习实践：文本分析 9.1 创建基础数据表 9.2 导入基础数据 9.3 文本分析实验建模 第 10 章 机器学习实践：葡萄酒品质预测 10.1 数据准备 10.2 实验建模 第 11 章 BI 报表实践：居民收入感受指数和物价满意指数 BI 报表（附件） 11.1 创建居民收入感受指数和物价满意指数表 11.2 导入本地数据 11.3 导入 MAXCOMPUTE 数据源 第 12 章 BI 报表制作练习：国内生产总值统计 BI 报表(附件) 12.1 创建表 12.2 从本地文件导入数据 12.3 创建数据集

		12.4　构建 BI 报表
大数据行业案例导论	经过学习，学生基本掌握大数据在真实案例中如何应用，了解人工智能在真实案例中的作用，从而基本了解大数据的核心价值是分析和预测。	第 1 章 概述 一、培养目标和内容 二、培养要求 三、产业需求（就业前景） 第 2 章 大数据行业导论实验指导书 案例一：交通行业套牌管理 案例二：实时预判的路况预测系统 案例三：银行反欺诈/反洗钱案例 案例四：分析型客户关系管理 案例五：心脏病预测案例 案例六：雾霾天气预测（二分类） 案例七：学生考试成绩预测（二分类） 案例八：评分卡信用评分（二分类） 案例九：商家作弊行为检测（二分类） 案例十：用户购买行为预测（二分类） 案例十一：相似标签自动归类（文本分析） 案例十二：新闻分类案例（文本分析） 案例十三：生存预测（多分类） 案例十四：信用风险预测（多分类） 案例十五：聚类与分类 案例十六：精细化营销（聚类） 案例十七：农业贷款发放预测（回归） 案例十八：协同过滤做商品分析（关联推荐） 案例十九：使用时间序列分解模型预测商品销量（时间序列）
大数据开发实践	本模块主要讲述并带领用户学习如何基于 Java 和 Python 进行分布式开发，深入学习阿里云大数据离线分布式平台 Maxcompute SQL 的高阶内容，并进一步学习 MapReduce、图计算等相关开发。	第 1 章 TUNNEL 开发 1.1　环境部署 1.2　用 TUNNEL 命令行上传下载数据 1.3　TUNNEL SDK 上传下载数据 第 2 章 分区 2.1 分 区 第 3 章 SQL 进阶 3.1　插入数据 3.2　查询数据 第 4 章 函数 4.1　创建数据表 4.2　日期函数 4.3　数学函数 4.4　聚合函数 4.5　窗口函数 4.6　字符串函数 4.7　用户自定义函数 4.8　其他函数 第 5 章 MAPREDUCE 开发

		5.1 基本介绍 5.2 测试准备 5.3 代码演示 5.4 资源测试 第 6 章 GRAPH 开发 6.1 前言 6.2 图加载阶段 6.3 编程接口 6.4 举例 第 7 章 JAVA SDK 开发 7.1 介绍 7.2 使用 ODPS JDBC 接入 ODPS 7.3 利用 ODPS 计算圆周率 PI 第 8 章 PYTHON SDK 开发 8.1 介绍 8.2 安装 8.3 快速开始

3.2.2 产品培训课程

产品名称	功能描述	
互联网数据爬取实战	手把手教学生学习互联网数据爬取的各种方法和技巧，并基于实战项目加深技能的掌握，为学生下一步进行更多的大数据分析和建模打下基础。只有获取到足够的数据后，才能做进一步的业务分析和机器学习/深度学习建模。	第 1 章 网络爬虫概述 1.1 概 述 1.2 网络爬虫结构 1.3 网络爬虫的类型 1.4 网络爬虫的实现原理 1.5 网络爬虫的实现技术 1.6 开发环境的配置 1.7 在 ECLIPE 中集成 PYTHON 开发环境 1.8 PYTHON 项目搭建 第 2 章 正则表达式 2.1 正则表达式概念 2.2 正则表达式的语法规则 2.3 PYTHON 中的正则 第 3 章 BEAUTIFULSOUP 3.1 BEAUTIFUL SOUP 的概述 3.2 BEAUTIFUL SOUP 的使用 第 4 章 REQUESTS 库基本使用 4.1 REQUEST 库概念 4.2 REQUEST 库的使用 第 5 章 REQUESTS+正则表达式爬取电影 TOP100 榜 5.1 流程框架 5.2 爬虫实战 5.3 项目源代码

		第 6 章 模拟登陆爬虫项目 6.1 模拟登录爬虫项目功能分析 6.2 模拟登陆某电商网站 6.3 模拟登录爬虫项目流程 6.4 模拟登录爬虫项目编写实战 6.5 利用 COOKIE 模拟浏览器登录 第 7 章 使用 SELENIUM 模拟浏览器抓取某电商网站服装信息 7.1 流程框架 7.2 爬虫实战 7.3 项目源代码 第 8 章 SCRAPY 框架 8.1 SCRAPY 架构图 8.2 SCRAPY 框架及依赖的安装 8.3 SCRAPY 框架的使用 第 9 章 JAVA 实现某电商网站数据爬取 9.1 数据库配置 9.2 代码演示 9.3 结果展示 第 10 章 EXCEL 表格数据爬取 10.1 基本用法 10.2 常见画图 10.3 多图合并
数据可视化开发	经过学习，学生基本掌握大数据背景下基于 Quick BI 和开源组件进行数据可视化开发，学习各种 dashboard 的制作。	第 1 章 数据可视化介绍 1.1 概 述 1.2 详细说明 1.3 主要应用 1.4 相关分析 第 2 章 EXCEL 数据可视化 2.1 概 述 2.2 柱 形 图 2.3 百分比柱形堆积图 2.4 折线图 2.5 饼 图 2.6 散 点 图 第 3 章 ECHARTS 数据可视化 3.1 概 述 3.2 开发环境 3.3 画图原理 3.4 准备工作 3.5 ECHARTS 常用图 3.6 ECHARTS 其他图 3.7 ECHARTS 高级使用 第 4 章 PYTHON 数据可视化 4.1 安装部署

		4.2 数据导入
离线分布式开发实践	经过学习，学生基本掌握阿里云大数据的分布式开发技能，并通过多个业务场景如基于电影评论数据的采集、基于 LBS 的热点店铺搜索、搭建社交好友推荐系统和海量电力设备监测数据存储分析等进一步强化学生的实际动手能力。	第 1 章 使用 MAXCOMPUTE STUDIO 开发大数据应用 1.1 工具安装 1.1.1 安装 JDK 1.1.2 安装 IntelliJ IDEA 1.1.3 安装 Studio 插件 1.2 项目空间连接 1.3 本地数据上传下载 1.3.1 本地数据导入 1.3.2 数据导出本地 1.4 编译 SQL 脚本 1.4.1 语法检查与错误提示 1.4.2 潜在风险提示 1.4.3 本地编译与错误定位 1.5 可视化分析作业运行 1.6 JAVA UDF 开发 第 2 章 基于电影评论数据的采集 2.1 数据采集概述 2.2 实现技术 2.3 项目案例 2.3.1 业务场景 2.3.2 远程桌面连接 ECS 服务器 2.3.3 安装 MySQL 数据库 2.3.4 创建电影数据表 2.3.5 创建 Java 项目 2.3.6 编码实现 2.3.7 测试运行 2.3.8 课后作业 2.3.9 应用推广 第 3 章 基于 LBS 的热点店铺搜索 3.1 LBS 含义 3.2 LBS 应用领域 3.3 应用介绍 3.3.1 场景介绍 3.3.2 数据集介绍 3.4 解决方案 3.4.1 经纬度 3.4.2 计算地球上两点间的距离 3.4.3 使用 GeoHash 算法 3.5 项目开发 3.5.1 开发环境 3.5.2 建表并上传数据 3.5.3 创建 MaxCompute Java 项目 3.5.4 编写代码

		3.5.5 本地测试 3.5.6 MaxCompute 上测试 第 4 章 搭建社交好友推荐系统 4.1 介绍 4.2 实现方式 4.2.1 好友推荐的实现 4.2.2 数据描述 4.2.3 期望推荐结果 4.2.4 MR 介绍 4.2.5 分析实现 4.2.6 在阿里云上的实现方式 4.3 搭建好友推荐系统 4.3.1 任务说明 4.3.2 本地开发测试 4.3.3 云端部署 第 5 章 基于数加实现在线交易分析 5.1 业务场景 5.2 数据源 5.3 业务需求及分析 5.4 准备工作 5.4.1 进入 MaxCompute 项目 5.4.2 创建在线交易表 5.4.3 创建用户信息表 5.4.4 创建交易分析存储表 5.4.5 创建用户数汇总表 5.4.6 创建工作流任务 5.5 数据导入 5.5.1 导入本地用户信息表数据 5.5.2 导入本地在线交易表数据 5.6 数据加工 5.6.1 数据加工“交易分析” 5.6.2 数据加工“用户数” 5.6.3 配置任务调度 5.6.4 测试工作流 5.7 数据展现 5.7.1 添加数据源 5.7.2 创建数据集 5.7.3 在线交易分析可视化
流计算开发实践	本课程主要带领学生学习和理解当前开源流计算工具（Flink、Storm、Spark Streaming）和企业级流计算工具（IBM Streams 和 Aliyun StreamSQL），基于 Flink 和	第 1 章 流计算概述 1.1 当前计算模式面临着挑战 1.2 提供实时计算分析平台 1.3 CEP 平台介绍 1.4 IBM STREAMS 流计算 1.5 STORM 平台介绍 1.6 SPARK STREAMING

	Aliyun SteamSQL 进行流计算开发和管理，学习 Flink 的安装部署以及最佳实践等。	1.7 APACHE FLINK 第 2 章 流计算场景用例 2.1 概述 2.2 方案介绍 2.3 某银行交易日志实时监控系统 2.4 某移动网络质量实时分析系统 2.5 某移动实时精确营销系统 第 3 章 FLINK 技术架构 3.1 核心架构 3.2 基本概念 3.3 关键技术 3.4 程序框架 3.5 功能支持列表 第 4 章 FLINK 环境部署 4.1 安装启动测试 4.2 代码结构 4.3 IDEA 开发环境部署 4.4 FLINK ON YARN 第 5 章 FLINK 流计算开发 5.1 FLINK 数据类型和序列化 5.2 FLINK DATASTREAM API 编程指南 5.3 FLINK DATASET API 编程指南 5.4 TABLE API &SQL 编程指南 第 6 章 FLINK 流计算管理 6.1 启动 6.2 关闭 6.3 运行样例程序 6.4 启动 SCALA 环境 6.5 监控 6.6 提交任务 6.7 执行环境 6.8 调试 第 7 章 最佳实践 7.1 运行 WORDCOUNT 程序 7.2 编写第一个 FLINK 程序 7.3 出租车乘客热点分析 第 8 章 平台比较 8.1 FLINK 和 STORM 的功能对比 8.2 FLINK 和 STORM，SPARK STREAMING 的功能对比 8.3 FLINK 和 STORM 的性能对比 第 9 章 参考文献

TensoFlow 开发实践	本模块主要带领学生学习如何基于 Google TensorFlow 进行机器学习和深度学习建模，具体包括： – 手写数字识别 (MNIST) 初级教程，主要学习多分类 (multiclass classification) 的相关知识； – 手写数字识别 (MNIST) 高级教程 ； – TensorFlow 基础，主要学习如何使用 TensorFlow 架构训练大规模模型 – 卷积神经网络实践，主要学习如何使用 TensorFlow 在 CIFAR-10 数据集上训练卷积神经网络。卷积神经网络是为图像识别量身定做的一个模型。 相比其它模型，该模型利用了平移不变性(translation invariance)，从而能够更更简洁有效地表示视觉内容。 – 循环神经网络 (Recurrent Neural Network, 简称 RNN)实践 – 序列到序列模型 (Sequence-to-Sequence Model)实践，主要学习如何采用序列到序列模型进行机器翻译，学员将学会构建一个完全基于机器学习,端到端的 英语-法语 翻译器。
Hadoop 开发实践	本模块主要带领学生学习如何基于 Java 和 Python 进行开源 Hadoop 的离线分布式开发，Hadoop 集群环境采用 Cloudera CDH，通过手把手的教学，让学生熟悉如何基于 HDFS 进行非结构化数据存储、如何基于 Hbase 进行需要快速响应 APP 的开发、如何基于 Impala、Hive 进行结构化数据管理等。
Spark 开发实践	本模块主要带领学生学习如何基于 Java 和 Python 进行开源 Spark 的离线分布式开发，Spark 集群环境采用 Cloudera CDH 中部署的企业 Spark 版本，通过手把手的教学，让学生熟悉如何基于 Spark SQL 进行结构化数据管理，如何基于 Spark Streaming 进行流计算的开发、如何基于 Spark GraphX 进行图并行计算以及如何基于 Spark Mllib 进行机器学习等。

3.2.3 行业案例实践课

产品名称	功能描述	部分目录
税务纳税评估实践	本课程全面基于税务行业已经实施的各种预测分析型大数据应用实践（重点是纳税评估），带领用户深入学习大数据背景下的深度学习建模全流程（以图形化托拉拽方式为主），在本课程中，学生通过学习可以全面了解税	第 1 章 税务行业实践：纳税评估 1.1 业务理解 1.2 原始数据数据文件 1.2.1 纳税评估数据文件 1.2.2 新增纳税人数据文件 1.2.3 纳税客户细分数据文件 1.3 创建基础数据表 1.3.1 使用向导创建纳税评估表 1.3.2 使用 DDL 语句创建纳税评估表 1.4 导入基础数据 1.4.1 从纳税评估数据文件中导入数据 1.4.2 从新增纳税人数据文件中导数据 第 2 章 税务行业建模分析 2.1 纳税评估建模

	务纳税评估的全流程（从业务理解、数据准备开始，到数据特征工程，最后到深度学习建模的各种调优方法）。	2.1.1 数据特征工程 2.1.2 模型训练 1：逻辑回归多分类模型 2.1.3 模型训练 2：逻辑回归多分类模型调整 2.1.4 模型训练 3：切分比例调整 2.1.5 模型训练 4：最小收敛误差调整 2.1.6 模型训练 5：朴素贝叶斯模型 2.1.7 模型训练 6：切分比例参数调整 2.1.8 模型训练 7：随机森林模型 2.1.9 模型训练 8：森林中树的个数调整 2.1.10 模型训练 9：叶节点数调整 2.1.11 模型训练 10：切分比例参数调整 2.1.12 模型结束语 2.1.13 对新增纳税人进行监控等级预测 第 3 章 税务行业纳税人客户细分 3.1 创建基础数据表 3.1.1 创建纳税客户细分表 3.2 导入基础数据 3.2.1 导入数据到纳税客户细分表 3.3 纳税客户细分建模 3.3.1 模型训练结束语
电商广告精准营销实践	本模块主要讲述并带领用户学习第三代电商广告精准营销的相关技术，在课程中，学员将会学习如何对每个用户进行360度画像；对某用户个性化推荐，预测客户打开推荐链接的点击通过率（CTR）；预测打开推荐链接的客户购买该商品转化率是多少（CVR）；根据预测结果，对高购买倾向客户进行精准营销推荐。	第1章 电商领域实践：精准营销概述 1.1 业务理解 1.2 原始数据文件 1.2.1 用户信息数据文件 1.2.2 推荐分析数据文件 1.2.3 用户行为数据文件 1.2.4 待预测用户信息数据文件 第 2 章 电商领域实践：精准营销数据分析处理 2.1 创建基础数据表 2.1.1 创建用户信息表 2.1.2 创建推荐分析表 2.1.3 创建用户行为表 2.1.4 创建待预测用户信息表 2.2 导入基础数据 2.2.1 导入数据到用户信息表 2.2.2 导入数据到推荐分析表 2.2.3 导入数据到用户行为表 2.2.4 导入数据到待预测用户信息表 2.3 更新用户信息 2.3.1 创建用户信息临时表 2.3.2 向用户信息临时表中插入数据 2.3.3 查询导入到“用户信息临时表”数据 2.4 更新推荐状态 2.4.1 创建推荐分析临时表 2.4.2 向推荐分析临时表中插入数据

		2.4.3 查询导入到“推荐分析临时表”的数据 2.5 用户信息推荐分析宽表状态 2.5.1 创建用户推荐分析宽表 2.5.2 向用户信息推荐分析宽表中插入数据 2.5.3 查询导入到“用户信息推荐分析宽表”的数据 第 3 章 电商领域实践：精准营销模型分析 3.1 精准营销建模 3.1.1 数据特征工程 3.1.2 模型训练 1：逻辑回归二分类模型 3.1.3 模型训练 2：切分比例参数调整 3.1.4 模型训练 3：最小收敛误差调整 3.1.5 模型训练 4：随机森林模型 3.1.6 模型训练 5：GBDT 二分类模型 3.1.7 模型训练 6：最大叶子数调整 3.1.8 模型训练 7：叶节点最少样本数调整 3.1.9 模型训练 8：线性支持向量机 3.1.10 模型训练 9：PS-SMART 二分类 3.1.11 模型训练结束语 3.1.12 进行推荐预测 3.2 MAXCOMPUTE 存储资源使用说明
银行行业风险信用模型实践	本模块主要讲述并带领用户学习如何进行风险信用模型建模和应用等	
电信行业用户流失分析实践	本模块主要讲述并带领用户学习如何进行用户流失分析建模和分析实践等	
统计分析案例：宏观经济分析	经过学习，学生基本掌握大数据背景下的统计分析全流程，学习如何基于阿里云统计分析工具以及开源工具进行各种统计分析工作。	
医疗行业大数据实践	主要目的是带领学生学习如何基于阿里云大数据和机器学习平台进行医疗大数据的建模和创新，具体包括以下内容： ● 脑外伤急救后迟发性颅脑损伤营销因素分析实践 ● 偏态分布的激素水平影响因素分析实践 ● 药物选择实践	

3.2.4 所含数据资源列表、数据量大小和主要元数据信息（部分）

数据资源列表	数据量大小	主要元数据信息
推荐分析	5M 到 1T（默认 5M，学校可以根据需要自己选择大小）	推荐编码、推荐活动编码、发送推荐时间、商品编 码、推荐状态、操作时间、用户编码、性别、年龄、消费等级、数码达人、靓丽女士、潇洒男士、美食一族、总共浏览次数、近 6 个月浏览次数、总共推荐次数、总共推荐成功次数、近 6 个月推荐次数、近 6 个月推荐成功次数、总共购物次数、近 6 个月购物次数、总共评论次数、近 6 个月评论次数
用户行为	3M 到 1T（默认 3M，学校可以根据需要自己选择大小）	操作时间、用户编码、行为编码、行为对象编码
人口普查数据	4M 到 1G（默认 4M，学校可以根据需要自己选择大小）	年龄、工作类型、序号、教育程度、受教育时间、婚姻状况、职业、关系、种族、性别、资本收益、资本损失、每周工作小时数、原籍、 收入
用户信息	3M 到 100G（默认 3M，学校可以根据需要自己选择大小）	用户编码、性别、年龄、消费等级、数码达人、靓丽女士、潇洒男士、美食一族、总共浏览次数、近 6 个月浏览次数、总共推荐次数、总共推荐成功次数、近 6 个月推荐次数、近 6 个月推荐成功次数、总共购物次数、近 6 个月购物次数、总共评论次数、近 6 个

		月评论次数
二季度用户购买行为数据	3M 到 1G（默认 3M，学校可以根据需要自己选择大小）	用户编号、物品编号、购物行为、购物时间
三季度用户购买行为数据	2M 到 1G（默认 2M，学校可以根据需要自己选择大小）	用户编号、物品编号、购物行为、购物时间
红楼梦全集节选	2M 到 1G（默认 2M，学校可以根据需要自己选择大小）	编号，内容
待预测用户	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	用户编码、性别、年龄、消费等级、数码达人、靓丽女士、潇洒男士、美食一族、总共浏览次数、近 6 个月浏览次数、总共推荐次数、总共推荐成功次数、近 6 个月推荐次数、近 6 个月推荐成功次数、总共购物次数、近 6 个月购物次数、总共评论次数、近 6 个月评论次数
当期收入感受指数	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	统计时间,当期收入感受指数,未来收入信心指数,当期物价满意指数,未来物价预期指数
商品目录	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	名称,规格,主条码
酒	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	固酸、挥发酸、柠檬酸、糖分、氯化物、二氧化物、总亚硫酸、密度、酸碱性、硫酸盐、酒精含量、质量
分区表插入数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	日期、序号、信息
数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	Ip、用户名

国内生产总值	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	统计年度、国内生产总值(亿元、人均国内生产总值(元)、国民生产总值(亿元)、第一产业(亿元)、第二产业(亿元)、工业(亿元)、建筑业(亿元)、第三产业(亿元)、交通运输仓储邮电通信业(亿元)、批发零售贸易及餐饮业(亿元)
贷款拖欠预测	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	年龄,教育水平,工作年数,现居地年数,家庭收入,是否有过拖欠历史,贷款占比,信用卡欠款,其他债务
学生成绩预测数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	性别、住址、家庭成员数、是否与父母住在一起、母亲的文化水平、父亲的文化水平、母亲的工作、父亲的工 作、学生的监护人、从家到学校需要的时间、每周学习时 间、挂科数、是否有额外的学习辅助、是否有家教、是否有相关考试学科的辅助、是否有课外兴趣班、是否有向上求学意愿、家里是否联网、家庭关系、课余时间量、跟朋友出去玩的频率、日饮酒量、周饮酒量、健康状况、出勤量、期末成绩
学生成绩	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	学号,姓名,性别,所属班级编号,入学成绩
学生信息表	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	学号,姓名,性别,所属班级编号,入学日期
学生身高统计表	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	学号、姓名、班级编号、身高

时间	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	时间
中文分词词典	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	词典项
中文分词模式	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	模式项
院系信息	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	院系编号,院系名称
班级信息表	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	班级编号,班级名称,入学日期,所属院系中文名
纳税评估表	4M 到 1G（默认 1M，学校 可以根据需要自己选择大小）	入库、营业收入、成本、营业费用、管理费用、财务费用、利润总额、登记注册类型、行业、登记月份数、监控等级
新增纳税人表	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	入库、营业收入、成本、营业费用、管理费用、财务费用、利润总额、登记注册类型、行业、登记月份数、监控等级（重点监控 3 一般监控 2 正常 1）
纳税客户细分表	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	入库、营业收入、成本、营业费用、管理费用、财务费用、利润总额、登记注册类型、行业、登记月份数
公司部门管理费用	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	部门、管理费用比例
公司某年任务完成情况汇总	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	季度、目标值、完成值
某班级学生成绩	1M 到 1G（默认 1M，学校	姓名、专业课成绩、公共

	可以根据需要自己选择大小）	课成绩、总成绩
某专业学生成绩	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	姓名、专业课成绩、数政成绩、外语成绩、计算机成绩、公共课成绩、总成绩
商品房销售额月度数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	月份、销售额
网购金额	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	年龄、网购金额
学生入选名单	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	学号、姓名、性别、原学专业、系别、录取学校、录取专业、专业课成绩、数政成绩、外语成绩、计算机成绩、公共课成绩、总成绩
06 非分区表插入数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	日期、序号、数量
出版社信息	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	出版社编号、出版编号、出版社名称、注册时间、商品名称
日期	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	id、name、sex、hight、date
商品信息	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	出版编号、商品名称、商品类别、商品价格、评价等级
学生行为数据	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	行为时间、学号、行为地点、行为编码、备注说明
05 大数据开发套件基础	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	字段、含义、类型、说明
friend_in.csv	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	uid、friends
online_trading.csv	1M 到 1G（默认 1M，学校	create_time、good_cate、brandbuyer_id、

	可以根据需要自己选择大小）	trans_num、trans_amount、click_cnt、addcart_cnt、collect_cnt
poi.csv	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	
user.csv	1M 到 1G（默认 1M，学校可以根据需要自己选择大小）	buyer_id、buyer_prov、gender age_range、zodiac

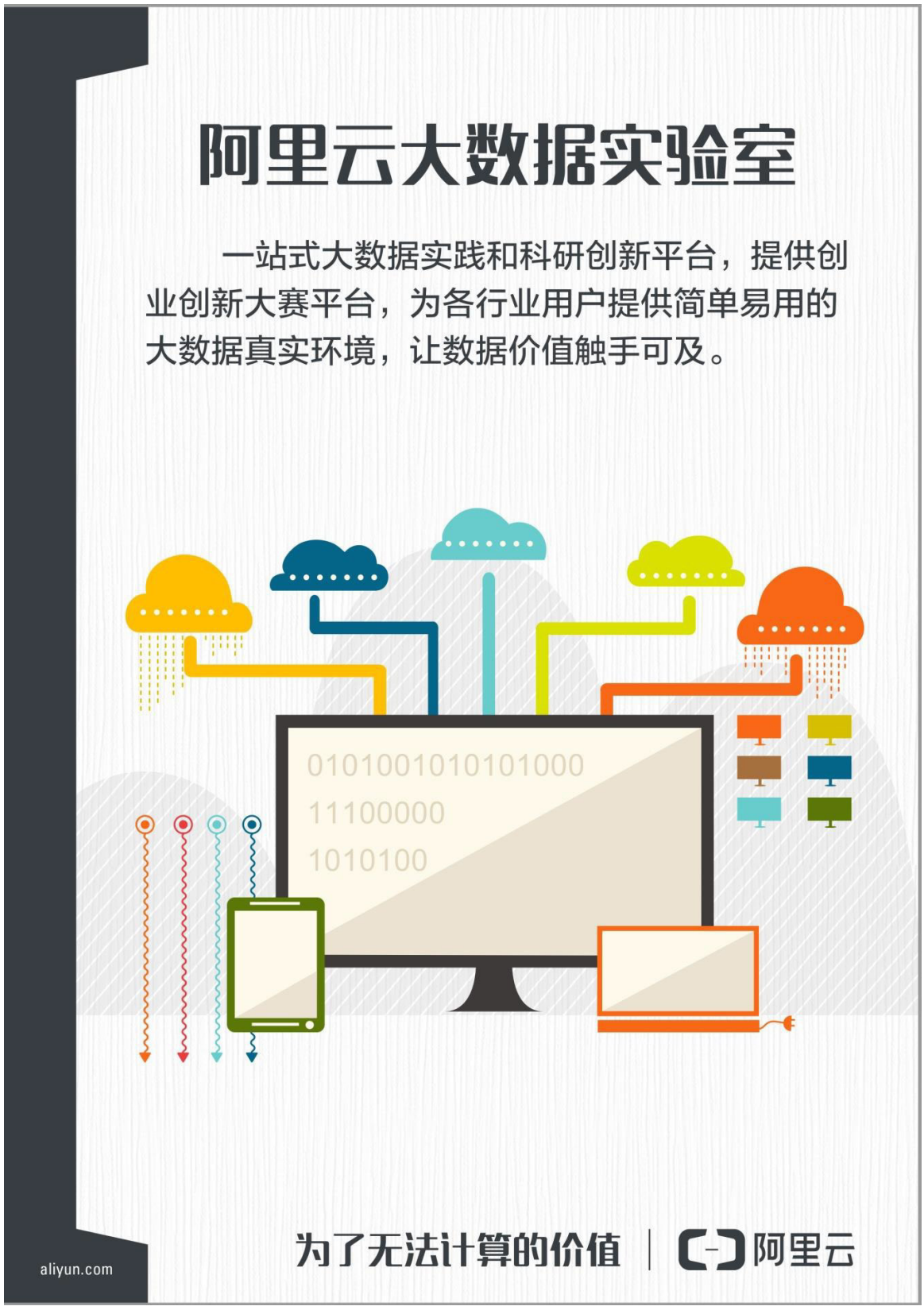
3.3 双创功能

基于阿里云大数据实验室，高校的老师和学生可以参加阿里云天池大赛（每年有几十场比赛，面向学生为主）和工信部全国移动互联创新大赛。阿里云天池大赛每场比赛的奖金从几万到几十万不等。目前，阿里云天池大赛已经面向全球 91 个国家，有超过 19 万人参与。天池可以带领学生学习和探索，一路成长为领域数据科学家。

工信部全国移动互联创新大赛自 2015 年成功举办以来，已服务近万名双创青年、百家双创企业，形成中国双创的一股重要力量。大赛以创新精神为火种，点燃社会创新创业热情;以大赛为平台打造创新创业的生态系统，推动双创人才培育、创新技术成果落地。大赛致力于挖掘全国创新创业优秀项目与人才，构筑创新创业生态环境，推动产创融合，推动全国的科技创新创业热潮不断发挥作用。大赛为后续科技创新项目的产业化落地奠定坚实基础，真正实现了高端平台与地方政策、科技创业与转型升级、资本与人才、高校与企业以及大企业之间的相互融合发展，为淄博乃至山东打造全国性科技创业示范基地，推动“新旧动能转换”实践，引领科技创新与地方产业升级的跨越式发展。

3.4 实验室装修参考（院校自建）

大数据实验室培训教室的装修是为学生提供一个能够实际操作大数据的真实环境，为高校师生提供场地、配套耗材、基础设施等硬件，主要包括实验区、办公区、成功展示区，以及工位设计、形象墙、休息区和台式电脑等。阿里云提供相应的企业文化，产品宣传相关资料，具体装修方案由学校根据教学要求自行确定。


阿里云大数据实验室
一站式大数据实践和科研创新平台，提供创业创新大赛平台，为各行业用户提供简单易用的大数据真实环境，让数据价值触手可及。
0101001010101000
11100000
1010100
aliyun.com
为了无法计算的价值 | 阿里云


一站式
大数据实践和
科研创新平台
01
数据采集
02
计算引擎
03
数据加工
04
数据分析
05
机器学习
06
数据应用
aliyun.com
为了无法计算的价值 | 阿里云

第4章 阿里云大数据平台

4.1 数加

阿里云大数据实验室底层依托于阿里云大数据公共云平台：“数加”平台。阿里云数加“大数据开发平台”是阿里巴巴集团推出的大数据领域平台级产品，提供一站式大数据开发、管理、分析、挖掘、共享、交换等端到端的解决方案，利 用 MaxCompute（原 ODPS）在几分钟内可将原始数据转变为业务洞察的海量数据处理能力，而无需关心集群的搭建和运维。产品架构如下图所示：

4.2 阿里巴巴专有云 Apsara Stack

Apsara Stack 专有云，阿里巴巴将其定位为第一个大规模商用的“专有公共云”。2015 年 7 月，阿里云就发布了专有云 1.0 版，如今已经是 3.0 版。Apsara Stack 基于与阿里云公共云同源的飞天大规模操作系统及 API，具有“超大规模”、“一应俱全”、“安全可控”、“稳定可靠”四大核心优势。

例如，Apsara Stack 在每个区域中可部署多达 6000 个节点；具备与公共云同源的大数据、人工智能、Aliware 互联网业务中台及安全产品等 55 款丰富的全方位云服务；多租户隔离，符合等保四级、可信云等安全认证；具有金融级别容灾方案，系统持续高可用。

同时，Apsara Stack 提供了自主可控的系统框架，包括统一账号、统一资源管理、统一权限管理、统一交付、统一监控运维以及统一故障定位，为企业提供了统一规划管理，提供“货真价实”的专有云平台。

Apsara Stack 在为企业实现自主可控的同时，还支持 Windows Server、SQL Server 等商业引擎，以及 Docker、Linux、MySQL、K8S 等开源引擎，并提供兼容生态的接口和模型并未客户提供定制的开源、商业、自研引擎。

第5章 开源计算平台

5.1 Cloudera CDH

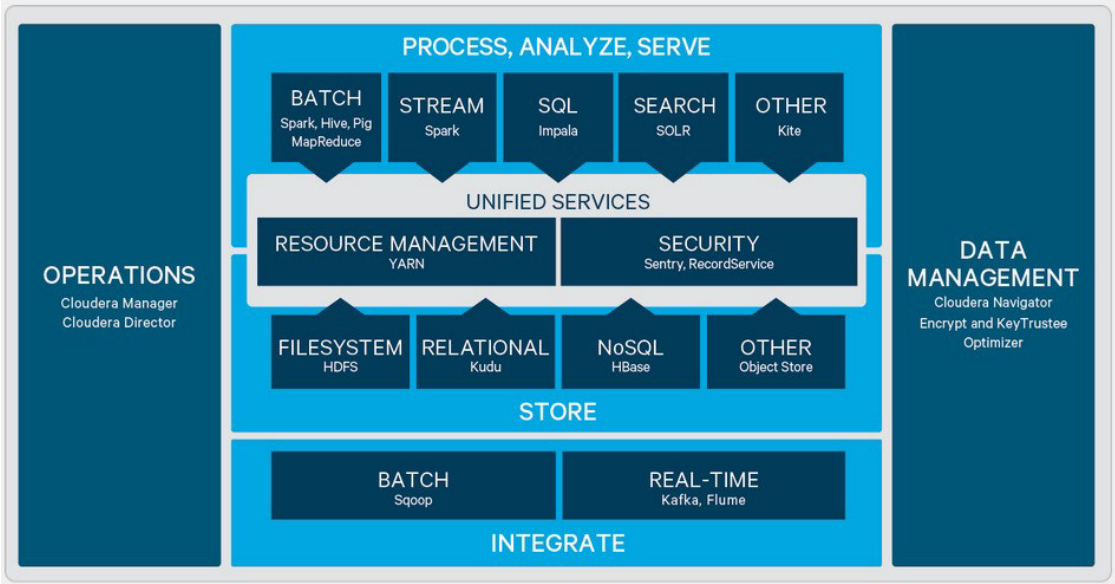
Cloudera Distribution Hadoop(CDH)是 Cloudera 基于 Apache Hadoop 和其相关 project 发行的 Hadoop 企业版，可以做批量处理，交互式 sql 查询和及时查询，基于角色的权限控制。在企业中使用最广的 Hadoop 分发版本。CDH 提供了 Hadoop 的核心功能，可扩展存储、分布式计算、基于 Web 的用户界面和重要的企业能力。CDH 是 Apache 授权许可的开放源代码，提供一整套的 Hdoop 解决方案进行统一的批处理、交互式 SQL 和交互式搜索， 以及基于角色的访问控制。截至目前为止，CDH 共有 5 个版本，其中，前两个已经不再更新， 最近的两个，分别是 CDH4（在 Apache Hadoop 2.0.0 版本基础上演化而来的），CDH5，它们每隔一段时间便会更新一次。

Cloudera 版本层次十分清晰，且它提供了适用于各种操作系统的 Hadoop 安装包，可直接使用 apt-get 或者 yum 命令进行安装，更加省事。就目前而言，不收费的 Hadoop 版本主要有三个（均是国外厂商），分别是：Apache（最原始的版本，所有发行版均基于这个版本进行改进）、Cloudera 版本（Cloudera’ s Distribution Including Apache Hadoop，简称 CDH）、Hortonworks 版本(Hortonworks Data Platform，简称“HDP” ），对于国内而言，绝大多数选择 CDH 版。

CDH 提供以下功能:

- 灵活性：存储任何类型的数据，并可以通过多种不同的计算框架（批处理、交互式 SQL、文本搜索、机器学习和统计计算）操作它们。
- 集成性：基于广泛的硬件和软件解决方案，快速实现完整的 Hadoop 平台搭建和运行。
- 安全性：处理和控制敏感数据信息。
- 可扩展性：使大量应用可以弹性扩展，以满足你的实际需求。

- 高可用性：保障关键任务稳定运行。
- 通用性：保护已有 IT 基础设施和投入。



作为世界上最流行的 Hadoop 及相关项目的发行版，CDH 融了当下几乎所有的最流行大数据处理技术。

1. 数据迁移：

实现与现有系统或应用程序快速集成，是通过批量加载处理（Apache Sqoop）或流式传输（Apache Flume，Apache Kafka）将数据移入和移出 Hadoop。

2. 数据存储：

基于 HDFS 分布式文件系统的 Hadoop 无限扩展的灵活架构允许企业存储、分析无限数量和类型的数据。

HBase 是一个高性能的分布式数据存储，与 Cloudera 的平台集成，提供安全且易于管理的 NoSQL 数据库。对存储的所有数据进行快速，随机的读写操作，并与其他组件（如 Apache Kafka 或 Apache Spark Streaming）集成，以在单个平台中构建完整的端到端工作流程。

Apache Kudu 是一个介于 HDFS 和 HBase 的性能特点之间的一个系统，在随机读写和批量扫描之间找到一个平衡点，并保障稳定可预测的响应延迟。从而应对快速变化数据的快速分

析型数据仓库，靠系统自身能力，支撑起同时需要高吞吐率的顺序和随机读写的应用场景，提供可与 MapReduce， Spark 和其它 hadoop 生态系统集成。

3. 工作负载管理：

通过 YARN 提供的开源资源管理，使我们可以超越批处理，并将数据打开到各种工作负载，包括交互式 SQL，高级建模和实时流。同时其细粒度配置可实现更好的集群利用率，在不中断最关键的操作的同时，通过在业务中的优先级工作负载和基于组的策略中启用工作负载 SLA， 实现以更多方式处理更多数据。

4. 安全与管制：

通过 Sentry 的开源授权模型保证了用户数据访问和 BI 应用程序的充分安全。在 Sentry 中，所有的权限都只能授予角色，当角色被挂载到用户组的时候，该组内的用户才具有相应的权限。即：权限→角色→用户组→用户，这一条线的映射关系在 Sentry 中显得尤为清晰。因此业界将 sentry 成为 Hadoop 的“哨兵”。

5. 多样化的分析平台：

使用多个数据访问选项（Apache Hive，Apache Pig）进行大规模复杂数据转换用于大型数据集批处理（MR2）或快速内存（Apache Spark）处理。

利用 Spark 的分布式内存存储，实现各案例的高性能处理，包括批处理，实时流式处理和高级建模与分析。与 MapReduce 相比，Spark 具有显着的性能改进，是将数据转化为实际结果的首选工具。

为了提供更为直接的查询和响应，并且无需编写 MapReduce 应用程序，您可以使用 Apache Impala。作为大规模并行处理（MPP）引擎，Impala 实现了数量级的性能提升，并且可以支持高并发工作负载，以秒为单位返回结果，为业务分析人员提供广泛的访问，从而获得 最快的时间见解。

对于不了解 SQL 的用户，可以使用 Cloudera Search。Cloudera Search 是 Apache Solr 完全集成在 Cloudera 平台中。我们可以通过简单的自然语言访问存储在 Hadoop、Hbase 或云存储中的数据。最终用户和其他 Web 服务可以通过全文查询和各方面的深入研究来探索文本，半结构化和结构化数据，从而快速过滤和聚合它，以获得业务洞察。

5.1 Tensorflow

近几年，信息时代的快速发展产生了海量数据，诞生了无数前沿的大数据技术与应用。在当今大数据时代的产业界，商业决策日益基于数据的分析作出。当数据膨胀到一定规模时，基于机器学习对海量复杂数据的分析更能产生较好的价值，而深度学习在大数据场景下更能揭示数据内部的逻辑关系。本文就以大数据作为场景，通过自底向上的教程详述在大数据架构体系中如何应用深度学习这一技术。大数据架构中采用的是 hadoop 系统以及 Kerberos 安全认证，深度学习采用的是分布式的 Tensorflow 架构，hadoop 解决了大数据的存储问题，而分布式 Tensorflow 解决了大数据训练的问题。

首先，TensorFlow 的一大亮点是支持异构设备分布式（heterogeneousdistributedcomputing）。其次，TensorFlow 支持卷积神经网络（convolutional neural network，CNN）和循环神经网络 recurrent neural network，RNN），以及 RNN 的一个特例长短期记忆网络（long short-term memory，LSTM），这些都是目前在计算机视觉、语音识别、自然语言处理方面最流行的深度神经网络模型。

TensorFlow 是一个采用数据流图（data flow graphs），用于数值计算的开源软件库。节点（Nodes）在图中表示数学操作，图中的线（edges）则表示在节点间相互联系的多维数据 数组，即张量（tensor）。它灵活的架构让你可以在多种平台上展开计算，例如台式计算机中的一个或多个 CPU（或 GPU），服务器，移动设备等等。TensorFlow 最初由 Google 大脑小组（隶属于 Google 机器智能研究机构）的研究员和工程师们开发出来，用于机器学习和深度神经网络方面的研究，但这个系统的通用性使其也可广泛用于其他计算领域。

第6章 免费体验：数加体验馆

数加体验馆：免费体验 + 教程 = 大数据零距离！

地址：https://data.aliyun.com/experience?spm=a21gt.99266.501311.2.rnhFAv

- DataIDE 示例教程
- BI 报表示例教程
- 机器学习示例教程
- 推荐引擎示例教程
- 人口普查统计案例：结合人口普查数据搭建实验，统计学历和收入的关系。
- 心脏病预测案例：包括数据预处理、特征工程、模型训练和预测等一套机器学习流程。
- 【图算法】金融风控实验：利用图算法，针对个人信用，解决金融行业的风控问题。
- 协同过滤做商品推荐：本实验利用协同过滤算法搭建了一套购物推荐流程。
- 农业贷款预测的回归算法实现：通过回归算法建立模型，预测农业贷款的发放
- 【文本分析】新闻分类：通过主题模型实现了整个文本分类的流程。
- 【在线预测】中学生成绩预测：本实验主要是展示平台在线预测能力，通过中学生的在校园行为预测期

第7章 教育部阿里云首批 9 本教材出版

一方面是工科毕业生的人数居世界第一，另一方面人才缺口仍然很大。这一鸿沟正在由产业界和教育界联手弥补。

6 月 9 日，中国高等教育学会与阿里云联合表示，为响应教育部“新工科”建设号召，云计算-大数据-云安全高等教育改革示范性教材正式出版，包括《云安全原理与实践》、《云计算原理与实践》等 9 本，这是首套该行业内由产业界与教育界联合编撰的新型教材，为中国的云计算、人工智能等行业的长远发展奠定基础。

这套教材自 2015 年底开始编写，由教育部高等教育计算机类专业教学指导委员会与阿里云联合成立云计算教学专家组，由阿里云派驻资深专家与清华大学、复旦大学、上海交通大学、中国科学技术大学、北京航空航天大学等高校的一线教师共同编写，博采众长，融合了产业界的一线实践优势与教育界的理论优势。

首批教材分别是《云安全原理与实践》、《云计算原理与实践》、《云上运维及应用实践教程（基础篇）》、《云上运维及应用实践教程（提高篇）》、《企业迁云实战》、《互联网大数据处理技术与应用》、《大数据基础及应用》、《大数据挖掘与应用》、《互联网大数据处理技术与应用》。

历时 18 个月的时间和在试点院校的试用积累，这 9 本以企业人才需求为导向，将学以致用、场景化案例教学为宗旨，服务于高校云计算、大数据和安全技术领域人才培养的教材正式获批出版，将被应用于各大高校的日常教学中。

第8章 成功案例

截止到 2018 年，数据在全国合作院校超过 500 所，合作内容包括： 大数据/人工智能实训、双创、科研、大数据/人工智能人才培养、大数据/人工智能认证培训和大数据/人工智能信息化等。**其中阿里云大数据实验室在全国合作院校超过 200 家，阿里云大数据实验室成功案例具体如下：**

(一) **实训：截止到目前，数据在大数据实训层面合作院校和培训机构超过 200 个，其中付费用户超过 22 个，具体包括：**

a) **部分合作院校如：**北京理工大学（已落地）、山东财经大学（已落地）、太原理工大学（已落地）、华北电力大学（已落地）、河南财经大学（战略合作）、天津财经大学（战略合作）、南京工业大学（战略合作）、临沂大学（战略合作）、青岛黄海学院（战略合作）、南京信息工程大学（战略合作）、南京农业大学（战略合作）、南京邮电大学（战略合作）、江苏信息职业技术学院（已落地）、江苏经贸战略合作协议（战略合作）、江苏建筑职业技术学院（战略合作）、郑州铁路职业技术学院（战略合作）、邢台职业技术学院（战略合作）；

b) **已落地培训机构有：** 大连云工场（已落地）、浙江杰夫兄弟（已落地）、山西未来大智（已落地）、南京诚立功（已落地）、河南凯创教育（已落地）、南京永营（已落地）、大连永营（已落地）、中科创大（战略合作）等。

(二) **科研：截止到目前，数据在科研层面合作院校超过 20 所，部分院校**如北京理工大学、中山大学、河南财经大学、天津财经大学、山东财经大学、临沂大学、青岛黄海学院、江苏信息职业技术学院、郑州铁路职业技术学院等。

(三) **双创：截止到目前，我们在双创方面正在落地的院校超过 160 所**，比如通过与中科创大战略合作，将各种资源引入到其与高校联合运营的 151 所高校双创学院，其他正在合作院校如临沂大学、青岛黄海学院、河南财经大学、东北财经大学、中北大学、江苏信息职业技术学院等。