一、专业实践训练整体情况

<table>
<tr><td>实践单位名称</td><td colspan="2">杭银消费金融股份有限公司</td></tr>
<tr><td>实践单位地点</td><td colspan="2">杭州市庆春路金龙财富中心</td></tr>
<tr><td>实践岗位名称</td><td colspan="2">大数据开发工程师</td></tr>
<tr><td rowspan="2">专业实践训练时间</td><td rowspan="2">集中进行</td><td>2021 年 05 月 01 日开始　至　2021 年 12 月 31 日结束</td></tr>
<tr><td>专业实践训练累计　244　天（单位考核前），其中项目研究天数　130　天（单位考核前）</td></tr>
<tr><td colspan="3">（1）基本概况（含实践单位简介、实习实践内容等）

杭银消费金融股份有限公司于 2015 年 12 月正式开业，作为中国银监会批复筹建并监管的持牌金融机构，是浙江省内获批组建并开业的消费金融公司，可在全国范围内开展消费金融业务（全国仅 25 家）。由杭州银行作为主发起人，联合滴滴、中国银泰等知名企业组建的持牌消费金融机构，注册资本 25.61 亿元。 公司秉承“数字普惠金融”初心，支持服务传统金融覆盖不充分的、具有消费信贷需求的客户群体，以“数据、场景、风控、技术”为核心，始终不懈探索消费金融新模式，为全国消费者提供专业、高效、便捷、可信赖的金融服务。

随着大数据时代的到来，人们对数据资产更加重视，数据赋能业务能做的事情变得越来越多。
本次实践内容主要包括：建设企业级数据仓库、指标体系建设、数据治理。</td></tr>
<tr><td colspan="3">（2）项目研究概述（含项目名称、项目来源、项目经费、主要研究目标和技术难点等）

项目名称：企业级数据仓库建设
项目来源：杭银消费金融股份有限公司内部需求
项目经费：杭银消费金融股份有限公司提供
主要研究目标：建立统一规范的企业级数据仓库，全面涵盖个业务系统数据，统一数据规范，规范指标体系建设，同时开展数据治理工作，提高数据质量。
技术难点：实现实时数仓建设；复杂脚本代码优化</td></tr>
</table>

（3）项目开展情况（含项目研究内容、研究方案及技术路线，研究团队分工、本人承担任务及完成情况，存在问题与改进建议等，不少于500字。）

杭银消费金融股份有限公司于2015 年12 月正式开业，作为中国银监会批复筹建并监管的持牌金融机构，是浙江省内获批组建并开业的消费金融公司，可在全国范围内开展消费金融业务（全国仅25 家）。由杭州银行作为主发起人，联合滴滴、中国银泰等知名企业组建的持牌消费金融机构，注册资本25.61 亿元。 公司秉承“数字普惠金融”初心，支持服务传统金融覆盖不充分的、具有消费信贷需求的客户群体，以“数据、场景、风控、技术”为核心，始终不懈探索消费金融新模式，为全国消费者提供专业、高效、便捷、可信赖的金融服务。

随着大数据时代的到来，人们对数据资产更加重视，数据赋能业务能做的事情变得越来越多。

本次实践内容主要包括：建设企业级数据仓库、指标体系建设、数据治理。

二、专业实践训练收获

（一）围绕考核评价指标体系，举例说明以下收获（不少于 800 字）

在此次项目实践中，收获颇多。对数仓建设进行了系统的学习与实践，对金融行业的知识也有了一定的了解，对了解了数仓建设中运用到的各项技术。

数据仓库的建设：本次项目是金融行业的数据仓库建设，按照业务范围对现实实体进行抽象，主要划分为以下几个主题：当事人主题、事件主题、协议主题、财务主题、风控主题等。本次数据建模主要使用维度建模方法，面向分析场景构建数仓模型，重点关注快速、灵活的解决分析需求，同时大规模数据的快速响应性能。数仓模型主要分为：ODS 层、DW 层（DWD 层和 DWS 层）、DM 层、DIM 层。ODS 层存放原始数据，主要是业务操作系统的数据。DWD 层是构建数仓明细表（事实表），以业务过程作为建模驱动，基于每个具体的业务过程特点，构建最细粒度的明细事实表。DWS 层是数仓轻度汇总层，DM 是数据集市层，DIM 层是公共维度层。

金融行业的知识：按照业务流程，用户从申请借款到还款结束，主要会经历：申请授信、授信成功、借款、还款、逾期等过程。而这些重点业务过程也是建模的重点表，业务方最为关注的信息。

各项技术：在开发过程中，我们会用到 sqoop 来抽取数据，用 airflow 来进行配置任务调度和依赖，用 tableau 和帆软（finereport）来进行报表数据的展现。实时数仓用 flink 和 iceberg。

能力提升：数仓建设能力和 hive 代码优化能力得以快速提升。hive 代码优化主要是对数据倾斜进行调优，发生数据倾斜主要有以下几种场景：count distinct、group by 、join on，主要思想就是限制行和列，让尽可能少的数据进行计算。一般有以下几种方法：1、使用 map join，让小表全部加载到内存中。2、使用 group by 代替 count distinct。3、数据分布不均匀，对行列进行剪裁和 filter 操作，让 join 的表数据量尽量小，过滤掉不需要的列和分区。

素质养成：一方面主要是代码开发规范的习惯养成，另一方面是自我学习探索的能力。代码开发规范主要涉及：命名方式：如表命名、作业名命名、脚本命名、层级命名、主题域命名、指标命名规范。自我学习探索能力：在开发过程中，对于一些大任务，跑批耗时较长，在此过程中，需要结合业务特点、数据特点以及技术进行优化，探索 mapreduce 的执行过程特点，发现问题所在进行优化。

（二）取得成效

目前各大公司的产品需求和内部决策对于数据实时性的要求越来越迫切，需要实时数仓的能力来赋能。传统离线数仓的数据时效性是 T+1，调度频率以天为单位，无法支撑实时场景的数据需求。即使能将调度频率设置成小时，也只能解决部分时效性要求不高的场景，对于实效性要求很高的场景还是无法优雅的支撑。因此实时使用数据的问题必须得到有效解决。

通过对实时数仓相关技术的调研，发现实时计算框架目前已经经历了三代发

展，分别是：Storm、SparkStreaming、Flink。目前实时计算框架越来越成熟。一方面，实时任务的开发已经能通过编写 SQL 的方式来完成，在技术层面能很好地继承离线数仓的架构设计思想；另一方面，在线数据开发平台所提供的功能对实时任务开发、调试、运维的支持也日渐趋于成熟，开发成本逐步降低，有助于在企业中落地这件事。我们尝试把公司内实时数仓建设的目的定位为，以数仓建设理论和实时技术，解决由于当前离线数仓数据时效性低解决不了的问题。在数据湖的建设中，我们使用了 Flink+Iceberg 方式来实时同步业务系统的数据，构建实时的 Data Pipeline。业务端产生大量的业务数据，被导入到 Kafka 的消息队列。运用 Flink 流计算引擎执行 ETL 后，导入到 Apache Iceberg 原始表中。在计算业务指标时，我们会再起一个 Flink 作业从 Iceberg 中消费增量数据。实时计算主要用于监控授信及放款业务，观察业务是否正常，是否会有卡单的情况。以及大屏监控授信的区域，防止大批量授信于同一地点，造成欺诈风险。

本次数仓建设，取得的成就非常显著。一方面为数据分析师以及业务方提供了清晰明了的数据指标，方便业务分析各类具体业务过程问题，找出问题所在并及时解决。另一方面数据质量以及数据的准确性也有了很大的提高，对于各类监管报表的上报提供了良好的基础，为公司创造了良好的社会形象。数据作为公司的重要资产之一，本次实践将分散在业务系统的数据整合在一起，并使之结构化存储，为后期分析客户风险以及营销都打下了坚实的基础。

数仓建设与论文相关度一般，主要是数据的处理以及大数据架构会对论文的数据框架有帮助。数据仓库的建设的主要目的就是清洗数据，最大化体现数据的价值，使之成为数据资产。我的论文题目为“企业金融事件舆情检测系统的设计与实现“，数据将从各网站以及各论坛上获取，获取的数据会经过多次清洗而沉淀，用以分析。金融知识的学习，业务的学习，各种专业名词的认识等使我对消费金融行业有了一个较为全面的认识。因此，此次实践对论文中数据处理方面有着一定的帮助。

3.在校期间主要研究成果【含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等】

成果名称	类别[含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	学校排名/总参与单位数

本人承诺

在专业实践训练及考核报告撰写过程中，如实提供材料，严守

学术道德、遵循学术规范。

签字：王新华　　2022年 6月 6日

三、考核评价

校外合作导师(或现场导师)评价	重点对研究生项目研究开展情况、职业素养、行业知识掌握、环境和岗位适应能力、工程实践能力、团队协作能力，以及通过技术应用创新、成果转化、解决工程实际问题等取得的经济和社会效益等方面的评价： 该项目完成得较高，对基地建设做出了贡献。 校外合作导师（或现场导师）签字：周慧迪 2022年 6 月 6 日
校内导师评价	重点对研究生科学素质、基础及专业知识掌握、技术应用创新能力、取得的研究成果、项目研究与学位论文撰写的相关程度等方面的评价： 校外实践内容充分利用了所学技术。 校内导师签字：[illegible] 2022 年 6 月 6 日

实践单位过程考核意见	实际实践开始时间：2021年 5月 1日　实际实践结束时间：2021年 12月 31日 专业实践训练累计天数：　244　其中项目研究天数：130 实践单位过程考核结果：☑优秀　☐良好　☐合格　☐不合格 审核签字并盖公章：周其进　2022年 6月 6日
最终考核结果审核备案	考核总成绩（由现场答辩考核成绩 90%+单位过程考核成绩 10%组成）： 是否重修：☐是　☐否 教学管理部（或相关分院）审核签字（公章）：　年　月　日

四、相关支撑材料

在校期间主要研究成果【含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等】证明材料原件扫描件，具体提交要求如下：

1.产品与样机扫描件包含企业证明材料（含产品与样机功能及创新性介绍、社会经济效益、个人贡献说明及相关照片等）。

2.授权专利扫描件包含专利证书授权页；未授权专利扫描件包含专利受理书扫描件和专利请求书扫描件。

3.著作扫描件包含封面、封底和版权页。

4.软件著作权扫描件包含著作权证书和登记申请表。

5.论文扫描件包含封面、封底、目录和论文全文（含收录证明）。

6.标准扫描件包含封面、版权页、发布公告、前言和目次。

7.获奖扫描件包含显示单位和个人排名的获奖证书。

8.成果转化扫描件包含企业证明材料（含成果技术说明、社会经济效益、个人贡献说明及相关照片等）。