【项目背景】

北京地面交通线网客流特征分析

用户每刷一次交通卡，都会产生一条客流记录。而在北京上海这样的大城市，每天的公交、地铁客流有千万左右。这些海量的客流数据该如何进行分析解读？客流数据的背后反映了一座城市怎样的区域特点和职住格局？如何从乘客刷卡数据算出其出行路径？

公共交通数据通常只有 AFC(记录乘客进出站刷卡信息)和 AVL(记录车次信息)数据，而乘客在公交、地铁系统中的活动轨迹完全未知，相当于一个黑箱。

如果一座城市每一天有 500-700 万名乘客在轨道交通网络中出行，那么出行路径就会变得非常复杂。即便是任意两个站点之间，其出行的路径都可能是非常多元化的。

当我们获取了每一天地铁乘客的进站、出站数据，就会很容易获得总客流的数据。但如何把客流落到具体的每一个“OD 对”（指从起点站到终点站）之间的路径上来？

【实训目标】

一、熟悉大数据项目开发全流程

- 需求：在复杂业务背景下，理解客户需求，提炼项目建设目标
- 业务设计：业务指标、纬度、应用、可视化。提炼应用价值
- 应用设计：数据流转链路、中间表、目标表、作业、调度、监控
- 开发：应用 MYSQL、HADOOP、HIVE、SPARK、AZKABAN、FINEBI 完成项目建设
- 测试：构造模拟数据，设计并执行单元测试、集成测试
- 部署：Linux 系统环境安装和配置、程序的部署

二、掌握项目实训知识要点

1. 业务：智慧交通客流分析，具体涵盖路面交通和轨道交通双融合场景

1.1 乘客、站点、线路、路网概念定义与生态关系
1.2 一卡通在路面交通、轨道交通中的使用场景与数据基础
1.3 路面交通站点、轨道交通站点、刷卡数据 GPS 位置应用
1.4 客流分析在交通领域、城市规划领域中的应用

2. 设计：特征工程与大数据批量离线数据

2.1 源数据系统概念与应用分析
2.2 乘客出行原子特征工程设计
2.3 乘客出行 OD 特征工程设计
2.4 乘客出行 OD 特征矩阵设计
2.5 客流特征分析与可视化

3. 开发：典型客流分析场景

3.1 抽取 MYSQL 客户行为数据至 HDFS 平台
3.2 挂载客户行为明细数据
3.3 整合乘客出行明细数据并形成 OD 出行链
3.4 分析乘客 OD 出行链并加工客流指标数据
3.5 AZKABAN 挂载全链路数据处理作业并形成有效调度
3.6 FINEBI 开发客流衍射图、热力图、气泡图、仪表盘、报表

【项目简介】

智慧交通客流分析系统，是基于***智慧交通大数据平台***，***整合***城市***路面交通***乘客乘车刷卡数据、车辆行驶数据、车站信息，***轨道交通***乘客进出站刷卡数据、车辆行驶数据、车站信息，串联乘客出行链路，生成乘客出行 ***OD***，最终用于分析大型城市***客流潮汐规律***、***出行特征***、***空间分布***、***时间分布***的综合性智慧分析平台。

- **智慧交通大数据平台**，指基于 Hadoop 生态体系建设的综合大数据平台，用于归集路面交通各业务系统数据、轨道交通各业务系统数据。平台涵盖一卡通数据、乘客刷卡数据、车辆数据、车辆行驶数据、车辆维修数据、车辆加能数据、车站数据、线路数据、设备数据、设备维修数据、车辆调度数据、员工数据等全生命周期数据。
- **整合**，又称 ETL(extract,transform,load)，即按照历史时间顺序，将源系统的数据抽取到目标平台的指定数据表的过程。*(本项目案例我们采用 ELT 模式)*
- **路面交通**，公交车系统体系；**轨道交通**，地铁系统体系。
- **出行链路**，乘客完成一次出行目的，如上班、下班、购物等，往往换乘多种交通运输工具，也可能多次换乘相同交通运输工具的不同班次车辆。乘客一次上下车刷卡或一次出入站刷卡形成一次出行链路。
- **出行 OD**，O：origination; D:destination。为达成一次出行目的，乘客串联出行链的第一个 O 和最后一个的 D 为出行 OD。
- **客流潮汐规律**，受城市居民通勤、节假日出行、热点事件等行为事件影响，乘客出行呈现出时空周期往返迁徙特征，最终形成城市客流整体潮汐特征。
- **出行特征**，日常普遍特征和特殊事件特征。
- **空间分布**，客流在特定时间，基于 GIS 地图呈现的空间分布。

【技术规范】

本项目在逻辑上共划分四层，即数据源层、贴源数据层、基础数据层、轻度汇总层、数据集式层。(S、F、A、M)

- **数据源层**：指智慧交通大数据平台的数据来源系统。本项目中默认包含两大系统，即公交系统和轨道交通系统。数据源系统中的数据库均采用 MYSQL，版本为 MYSQL 。
- **贴源数据层**：指智慧交通大数据平台与数据源系统的缓冲地带，一般用于存放由数据源系统抽取至大数据平台的原始数据。本层所存放数据遵循不加工或少加工原则。贴源数据层的表名遵循 S_系统名_源系统表名。
- **基础数据层**：指经由贴源数据层的数据清洗、规范、整合后，形成的基础明细数据，为大数据平台的潜在分析提供直接数据来源。基础数据层的表名遵循 F_主题名_表名 。
- **轻度汇总层**：为快速应用数据进行数据分析，节省数据加工次数，提升数据加工效率，特设计轻度汇总层，用于缓冲频繁往复使用的中间结果数据。基础数据层的表名遵循 A_主题名_源系统表名 。
- **数据集市层**：存放特定分析主题集市的结果指标数据、纬度数据，为数据分析可视化，数据应用提供结果数据。基础数据层的表名遵循 M_集市名_表名 。

智慧交通大数据平台应用 HIVE 进行数据分层与结构管理。安装 HIVE 前需先安装 JDK、HADOOP。指定 HADOOP 版本为 V3.2.2，HIVE 版本为 V3.1.2，JDK 版本为 V1.8.251。

数据处理作业分别采用 SPARK。指定 SPARK 版本为 V3.1.2。数据作业命名规范如下：

- 数据源系统到贴源数据层的加工作业命名为 S_系统名_作业名 。
- 贴源数据层到基础数据层的加工作业名为为 F_系统名_作业名 。
- 基础数据层到轻度汇总层的加工作业名为为 A_系统名_作业名 。
- 轻度汇总层到数据集市层的加工作业名为为 M_系统名_作业名 。

作业调度系统采用 AZKABAN，指定版本为 V0.0.1。

可视化工具采用 FINEBI，指定版本为 V0.0.1。

【数据字典】

1、公交 IC 卡刷卡交易 数据

公交 IC 卡刷卡交易数据				
序号	字段名	中文名称	数据类型	备注
1	CARD_ID	卡号	CHAR(8)	卡唯一识别号
2	CARD_TYPE	卡类型	CHAR(2)	00:其它 01:普通卡 06:纪念卡 07:员工卡 10:老年人卡 12:中小学生卡 13:其他学生卡 51:残疾人卡 52:见义勇为卡
3	TRADE_TYPE	交易类型	CHAR(2)	06:储值卡正常扣款记录 08:储值卡补票扣款记录 B4:计次卡正常扣次记录 B5:计次卡补票扣次记录 B9:员工卡乘车记录 F0:分段票制上车刷卡 F6:投币记录
4	TRADE_TIME	交易时间	CHAR(14)	通常为下车时间
5	MARK_TIME	标注时间	CHAR(14)	通常为上车时间
6	TRADE_STATION	交易站	NUMBER(2)	通常为下车站
7	MARK_STATION	标注站	NUMBER(2)	通常为上车站
8	LINE_ID	线路号	CHAR(5)	扣款的线路号
9	BUS_ID	车辆号	CHAR(8)	扣款的车辆号
10	MARK_LINE_ID	标注线路号	CHAR5)	上车时的线路号
11	MARK_BUS_ID	标注车辆号	CHAR(8)	上车时的车辆号

2、地铁一卡通交易明细

地铁一卡通交易明细				
序号	字段名	中文名称	数据类型	备注
1	CARD_ID	卡号	CHAR(8)	卡唯一识别号
2	CARD_TYPE	卡类	CHAR(2)	"00:其它 01:普通卡 06:纪念卡 07:员工卡 10:老年人卡 12:中小学生卡 13:其他学生卡 51:残疾人卡 52:见义勇为卡"
3	LINE_ID	线路号	CHAR(5)	扣款的线路号
4	STATION_ID	车站号	CHAR(3)	交易车站号
5	DEVICE_ID	设备号	CHAR(8)	刷卡闸机编号
6	OPERATING_DATE	运营日	CHAR(8)	格式：YYYYMMDD
7	TRADE_TIME	交易时间	CHAR(14)	格式：YYYYMMDDHHMMSS
8	IO_TYPE	进出站标志位	CHAR(1)	0：进站；1：出站

【任务一】

1、实现 Spark 与 Hive 的整合；

2、开发 Spark ELT 作业，实现从源数据文件（公交刷卡数据集和轨道交通刷卡数据集）到 Hive ODS 层的抽取和加载；

3、开发 Spark ETL 作业，实现从数据库 mysql（维度表，公交站点信息和地铁站点信息）到 Hive ODS 层的抽取和加载。

注：因公交站点信息和地铁站点信息以.csv 文件形式提供，所以要求：

1）先将数据文件加载到 MySQL；

2）再通过 Spark ETL 实现 MySQL -> Hive 的 ELT 过程。

要求：在下次课前一天提交任务完成代码。