# **物流离线数仓项目详解**

## **自我介绍**

面试官上午好，我叫田辰龙，目前专注于物流类目的离线数仓工作。接下来，我将详细介绍我负责的离线数仓项目。

离线数仓的核心功能是从各类数据源中提取数据，经过清洗、转换等一系列处理后，将其存储起来，为后续的数据分析和数据挖掘提供支持。这样做的重要意义在于，能够实现数据处理与存储的分离，有效减轻单个系统的压力，进而提高数据处理的效率和准确性。

我们选择构建离线数仓，主要基于其具备高性能、低成本、高效率和高质量的优势：

**高性能：**优秀的数据模型能让我们快速查询到所需数据，大大缩短数据检索时间。

**低成本：**良好的数据模型可减少重复计算，实现计算结果的复用，从而降低计算成本。

**高效率：**合理的数据模型能显著改善用户使用数据的体验，提高数据使用效率。

**高质量：**科学的数据模型能解决数据统计口径混乱的问题，降低计算错误的可能性。

## **一、数仓底层环境与数据同步**

本物流数仓项目搭建在 cdh 6.3.2 平台上，采用 hive 3.1.3 作为核心的数据存储和计算工具，有力支撑了物流业务的大规模数据存储与分析工作。

数据同步通过自主开发的脚本完成从 MySQL 到 Hive 的迁移，并将数据同步至 Hive 的 ods 层。这里的自主开发脚本是根据物流业务数据的特点专门设计的，能够确保数据同步的稳定性和准确性。例如，针对 MySQL 中频繁更新的订单数据，脚本会采用增量同步的方式，只同步发生变化的数据，减少数据传输量和同步时间。

## **二、数仓五层架构说明**

### **（一）ods 层（原始数据层）**

ods 层存储从 MySQL 同步过来的原始数据，完全保留数据在业务系统中的原始结构和内容，不做任何处理。这些原始数据是数仓后续所有数据处理的基础来源，包含的主要数据如下：

**订单数据：**包括订单编号、用户 ID、下单时间、订单金额、收货地址等。其中订单编号是唯一标识订单的字段，用户 ID 用于关联用户信息，下单时间精确到秒，方便后续按时间维度进行分析，订单金额则是进行财务相关分析的重要依据。

**物流轨迹数据：**记录货物在各环节的时间、地点、操作人员等信息。比如货物从出库、运输到入库等每个环节的具体时间点，所在的详细地点，以及负责该环节的操作人员姓名和工号等，这些数据能清晰反映货物的流转过程。

**司机与车辆数据：**涵盖司机基本信息（如姓名、身份证号、联系方式等）、车辆型号、牌照、所属机构等。司机的身份证号用于唯一标识司机，车辆牌照是车辆的唯一标识，所属机构信息则有助于按机构维度分析司机和车辆的相关数据。

**机构数据：**包含各物流网点、分公司的编码、名称、所属区域等。机构编码是区分不同机构的唯一标识，所属区域信息能为按区域分析业务提供支持。

### **（二）dim 层（维度表层）**

dim 层对 ods 层的原始数据进行清洗和整合，形成各类维度表，具体如下：

**时间维度表：**包含年、季、月、日、时等不同时间粒度的信息。例如，不仅有具体的年份，如 2023 年，还会细分到该年的第几季度、第几月、第几周、具体哪天以及当天的小时数等，方便后续从不同时间维度对数据进行分析。

**区域维度表：**整合省份、城市、区县等地理信息，明确层级归属关系。比如北京市属于北京市，海淀区属于北京市等，这样的层级关系能让后续按区域分析数据时更加清晰准确。

机构维度表：规范机构信息，补充机构类型、上级机构等字段。机构类型如物流网点、分公司等，上级机构信息能反映出机构之间的隶属关系，有助于分析不同层级机构的业务情况。

**司机维度表：**完善司机信息，增加司机等级、入职时间、所属车队等内容。司机等级可用于评估司机的业务水平，入职时间能分析司机的工作经验与业务表现的关系，所属车队信息方便按车队维度进行统计分析。

**车辆维度表：**补充车辆类型、载重、购置时间等车辆相关信息。车辆类型如货车、卡车等，载重信息影响车辆的运输能力评估，购置时间可用于分析车辆的使用年限与维护成本等关系。

对于区域划分调整、机构名称变更等维度属性的变化，我们通过增加版本号和生效时间进行记录。例如，当某个区域的划分发生变化时，会为该区域信息添加新的版本号，并记录生效时间，这样既能保留历史版本的信息，又能准确反映当前的状态，确保历史数据的可追溯性。

### **（三）dwd 层（事实表层）**

dwd 层用于记录物流业务过程中的具体事件，形成事实表，主要包括：

订单事实表：提取订单创建时间、订单金额、支付状态等信息，并关联用户和机构维度。支付状态如已支付、未支付、退款等，通过关联用户维度可以分析不同用户群体的订单情况，关联机构维度则能了解各机构的订单业务量。

运输事实表：记录货物的出发地、目的地、出发时间、到达时间等信息，关联司机、车辆、区域维度。出发地和目的地通过关联区域维度能明确运输的范围，出发时间和到达时间可用于计算运输时长，关联司机和车辆维度能分析不同司机和车辆的运输效率。

揽收事实表：记录揽收时间、揽收人、揽收机构等揽收相关信息。揽收时间与下单时间对比可分析揽收的及时性，揽收人信息能追溯具体的操作人员，揽收机构则方便统计各机构的揽收业务情况。

派送事实表：记录派送时间、签收人、派送状态等派送相关信息。派送状态如已签收、未签收、派送中等，签收人信息用于确认货物已送达，派送时间可用于分析派送的及时性。

事实表以事务为单位记录数据，每一条记录都对应一个具体的业务事件，为后续的指标计算提供了详细且准确的基础数据。

### **（四）dws 层（业务主题层）**

dws 层根据业务主题，对 dwd 层的事实表和 dim 层的维度表进行聚合关联，形成面向特定业务场景的汇总表，例如：

订单主题表：按日聚合各机构、各区域的下单数量和金额。通过该表可以快速了解不同机构和区域每天的订单业务量和营收情况，为业务调度和资源分配提供参考。

运输主题表：按司机和车辆类型聚合每日的运输次数、运输里程、运输时长。能直观反映不同司机的工作效率和不同类型车辆的运输表现，有助于对司机进行绩效考核和车辆的合理调配。

揽派主题表：按城市聚合每日的揽收量、派送成功量、派送失败量。可以分析不同城市的揽收和派送业务情况，对于派送失败量较高的城市，可进一步排查原因并进行优化。

这些汇总表可直接用于业务分析，减少了后续的计算量，提高了分析效率。

### **（五）ads 层（指标结果层）**

ads 层存储各类指标的统计结果，直接服务于业务决策。统计结果按不同维度和周期（日、周、月）存储，如各区域的下单金额、各司机的运输完成率等。这些结果通过 Tableau、FineBI 等工具进行可视化展示，以图表等直观的形式呈现给管理层，为其提供清晰的业务数据参考，帮助管理层做出更科学的决策。

## **三、指标体系说明**

### **（一）原子指标**

原子指标是直接来源于业务系统原始数据的基础指标，包括：

**下单数：**指用户成功提交订单的数量。

**下单金额：**用户下单时所涉及的总金额。

**运输完成次数：**货物成功从出发地运送到目的地的次数。

**揽收数：**物流机构成功揽收的货物数量。

**派送数：**成功送达收件人的货物数量。

**运输里程：**每次运输过程中货物的行驶距离。

**运输时长：**货物从出发到到达目的地的时间间隔。

### **（二）衍生指标**

衍生指标是通过对原子指标进行简单计算得到的指标，包括：

**平均每次运输里程：**总运输里程除以运输完成次数。

**平均每次运输时长：**总运输时长除以运输完成次数。

**去重金额：**去除重复计算的订单金额，例如同一订单多次修改金额时，只计算最终的金额。

**订单取消率：**订单取消数与下单数的比值，反映订单取消的比例。

**揽收及时率：**规定时间内完成揽收的数量与总揽收数的比值，体现揽收的及时性。

**派送及时率：**规定时间内完成派送的数量与总派送数的比值，衡量派送的及时性。

**平均每单金额：**订单总金额与下单数的比值，反映每笔订单的平均价值。

### **（三）派生指标**

派生指标是结合原子指标、衍生指标和业务逻辑，按不同维度计算得到的复合指标。

**机构维度**

某机构某周期的订单取消率：该机构在某周期内的订单取消数与下单数的比值。通过该指标可以了解该机构在特定周期内订单的稳定性，若取消率过高，需分析原因并进行改进。

某机构某周期的平均揽收时长：该机构在某周期内从下单到揽收的平均时间。该指标能反映机构的揽收效率，平均揽收时长越短，说明揽收效率越高。

**城市 / 省份维度**

某城市某周期的派送及时率：该城市在某周期内按时派送的数量与总派送数的比值。反映该城市的派送服务质量，及时率高说明派送服务更可靠。

某省份某周期的平均运输里程：该省份在某周期内的总运输里程与运输次数的比值。可了解该省份内运输的平均距离，为车辆调度和路线规划提供参考。

**司机 / 卡车维度**

某司机某周期的运输准点率：该司机在某周期内按时完成运输的次数与总运输次数的比值。是评估司机工作表现的重要指标，准点率高说明司机的时间管理和驾驶能力较好。

某卡车类型某周期的平均载重：该类型卡车在某周期内的总载重与运输次数的比值。能反映该类型卡车的实际载重能力利用情况，为卡车的选型和装载规划提供依据。

通过这些指标，可以从多个角度全面分析物流业务的运营状况，为业务优化提供有力的数据支持。

## **四、报表展示**

报表使用的是 FineReport，以下是部分报表的详细说明：

**第一张图：**ads\_truck\_stats（卡车分析）：根据卡车类别名称进行分组，展示每种卡车类型完成运输的总次数。例如，厢式货车完成运输 1000 次，平板卡车完成 800 次等。通过该报表可以清晰了解不同类型卡车的运输工作量，为卡车的调配和维护计划提供参考。

**第二张图：**ads\_order\_cargo\_type\_stats（各类型货物运单统计）：按照货物类型名称进行分组，展示每种货物类型的下单数。比如，电子产品下单 500 单，生活用品下单 300 单等。该报表有助于了解不同类型货物的需求情况，便于调整仓储和运输资源。

**第三张图：**ads\_city\_stats（城市分析）：依据城市名称进行分组，展示平均每次运输里程占完成运输里程的百分比。通过该报表可以对比不同城市的运输里程分布情况，了解各城市运输业务的特点。

**第四张图：**ads\_express\_province\_stats（各省份快递统计）：根据省份名称进行分组，展示每个省份的揽收金额。例如，广东省揽收金额 50 万元，浙江省 30 万元等。该报表能反映各省份的快递业务营收情况，为业务拓展和资源投入提供方向。