大数据项目之电商数仓

（电商数据仓库系统）

## 数据仓库概念

数据仓库是一个为数据分析而设计的企业级数据管理系统。数据仓库可集中、整合多个信息源的大量数据，借助数据仓库的分析能力，企业可从数据中获得宝贵的信息进而改进决策。同时，随着时间的推移，数据仓库中积累的大量历史数据对于数据科学家和业务分析师也是十分宝贵的。

## 数据仓库核心架构





电商业务流程常见数据流

订单流程

用户下单 → 订单表（MySQL） →Seatunnel采集 → Hive ODS.订单数据

↓

清洗（剔除测试数据、转换时间戳） → DWD.订单明细

↓

按天/商品/用户 聚合 → DWS.订单宽表

↓

生成KPI指标如：下单人数、下单金额 → ADS.订单统计

↓

接入报表/可视化平台（如 Superset、FineBI）

用户行为流程

埋点日志 → Kafka → Seatunnel采集 → HDFS原始日志 → ODS层解析日志

↓

行为类型分表（点击、浏览、加购） → DWD层行为表

↓

用户活跃、跳出、转化率计算 → DWS层汇总

↓

每日UV、活跃用户数等指标 → ADS层

整体架构概览（业务流程）

数据采集 → 数仓分层建模 → 数据加工→ 数据可视化

数据采集

本项目的数据采集模块基于定时调度执行的 JAR 包程序，在 Linux 服务器环境中运行。该程序负责模拟真实业务场景数据的生成与采集，并将数据以结构化格式（如 CSV、JSON、数据库表等）输出至指定存储路径，为离线数仓的 ODS 层提供数据源支撑。

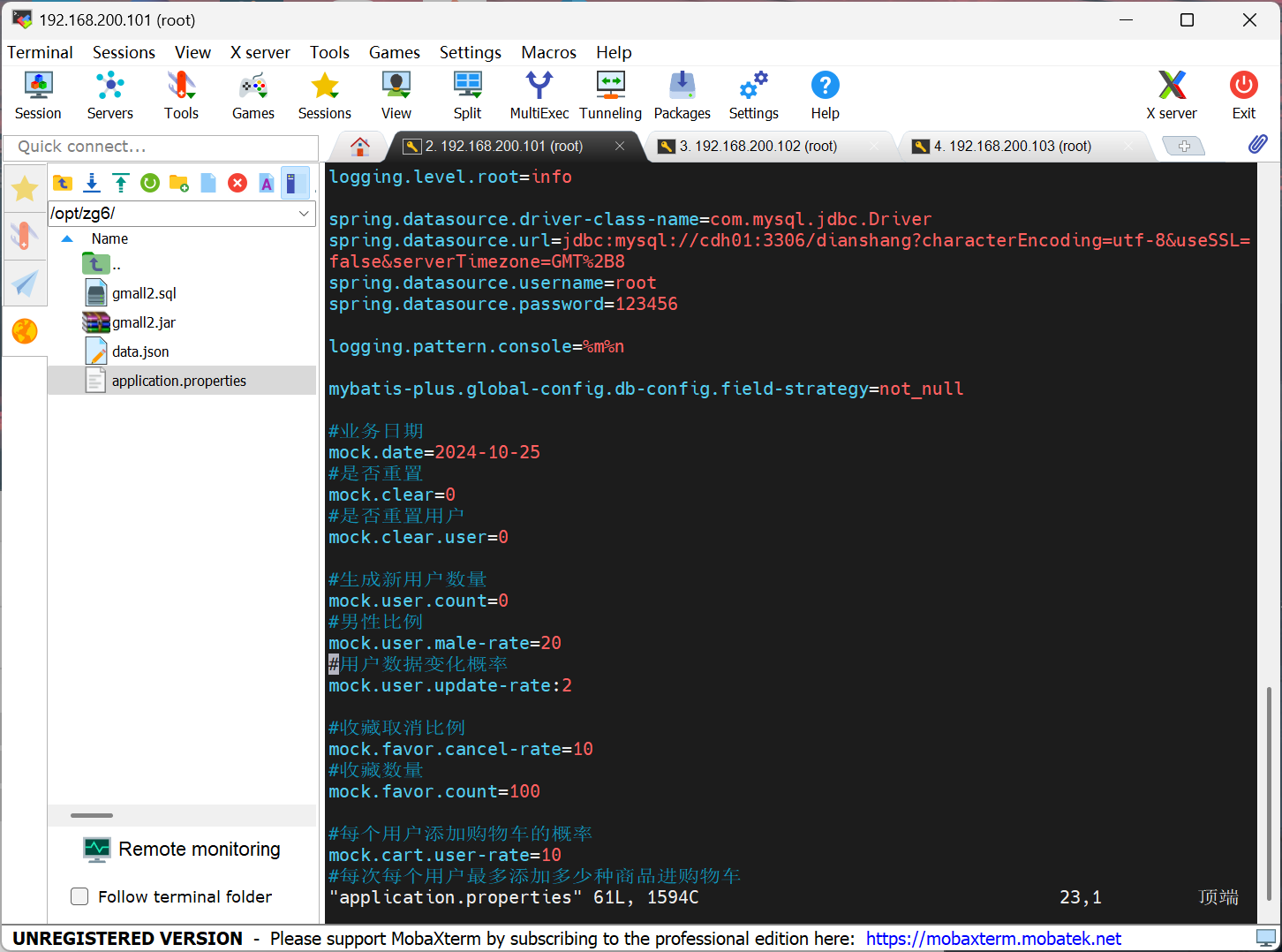
数据源：模拟数据生成器（Java JAR）

执行环境：Linux

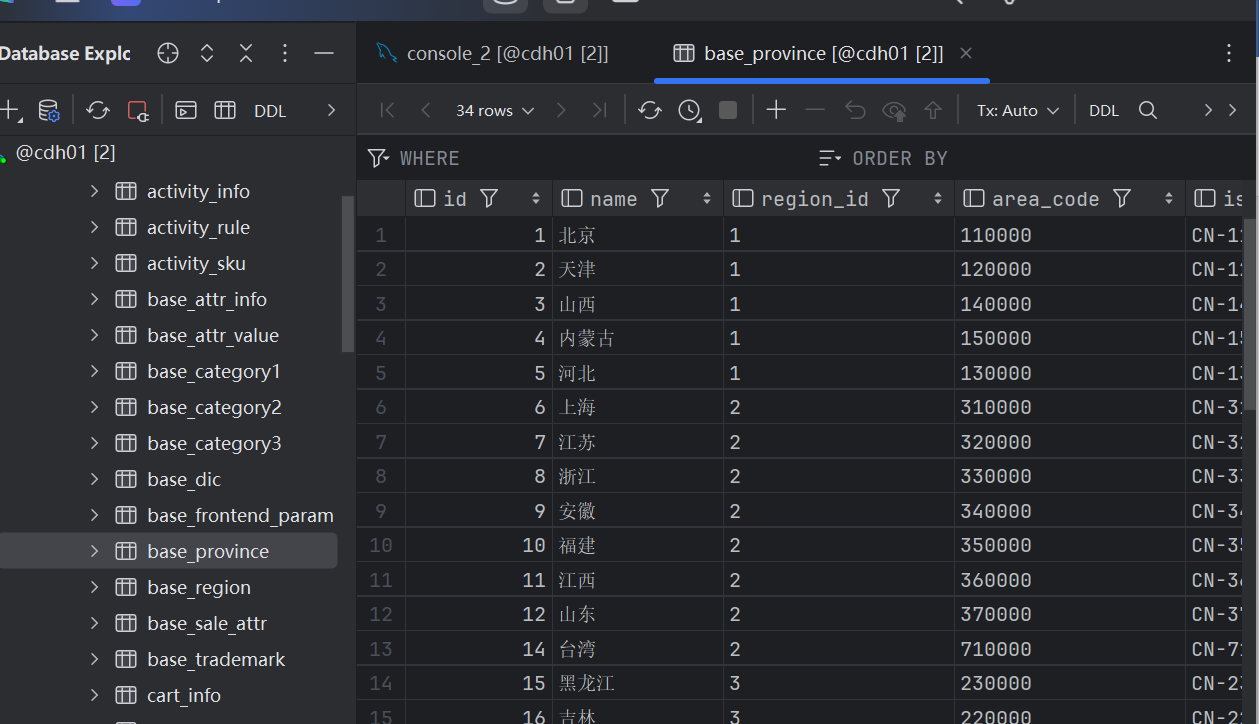
调度方式：Shell 脚本

输出格式：CSV / JSON / Hive 表

采集目标层：ODS（Operational Data Store）







MySQL 数据同步到 Hive

本项目实现了 MySQL 到 Hive 的离线数据同步机制。通过 Python 脚本自动生成多张 MySQL 表的 Seatunnel 配置文件（包括表结构映射、字段类型转换、数据同步逻辑），并通过 Shell 脚本批量调度执行 Seatunnel 同步任务，完成 MySQL 表结构与数据的自动入湖。该流程适用于数据仓库的 ODS 层数据采集，具备可扩展性和重复使用能力。

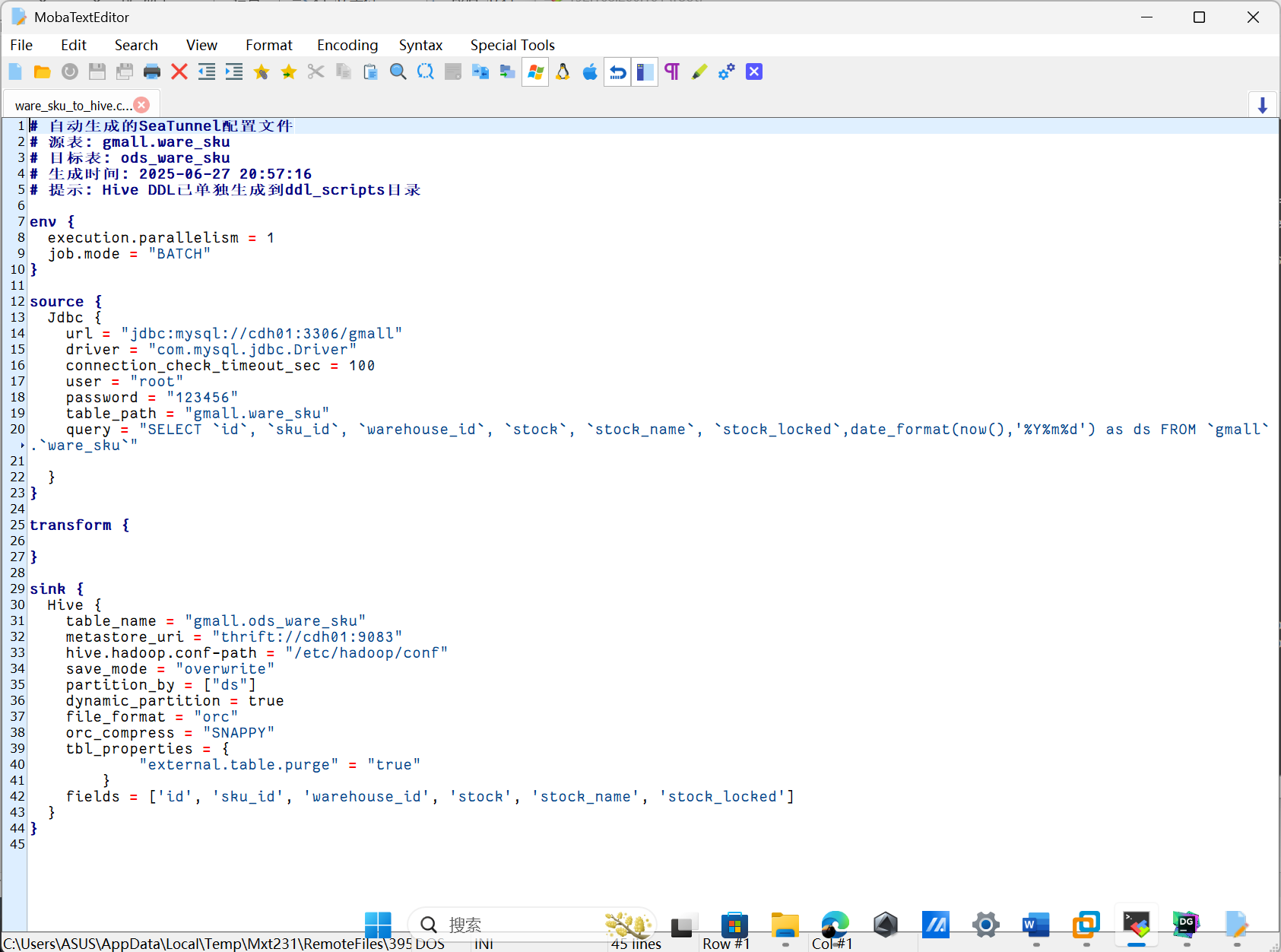
[MySQL 数据源]

↓

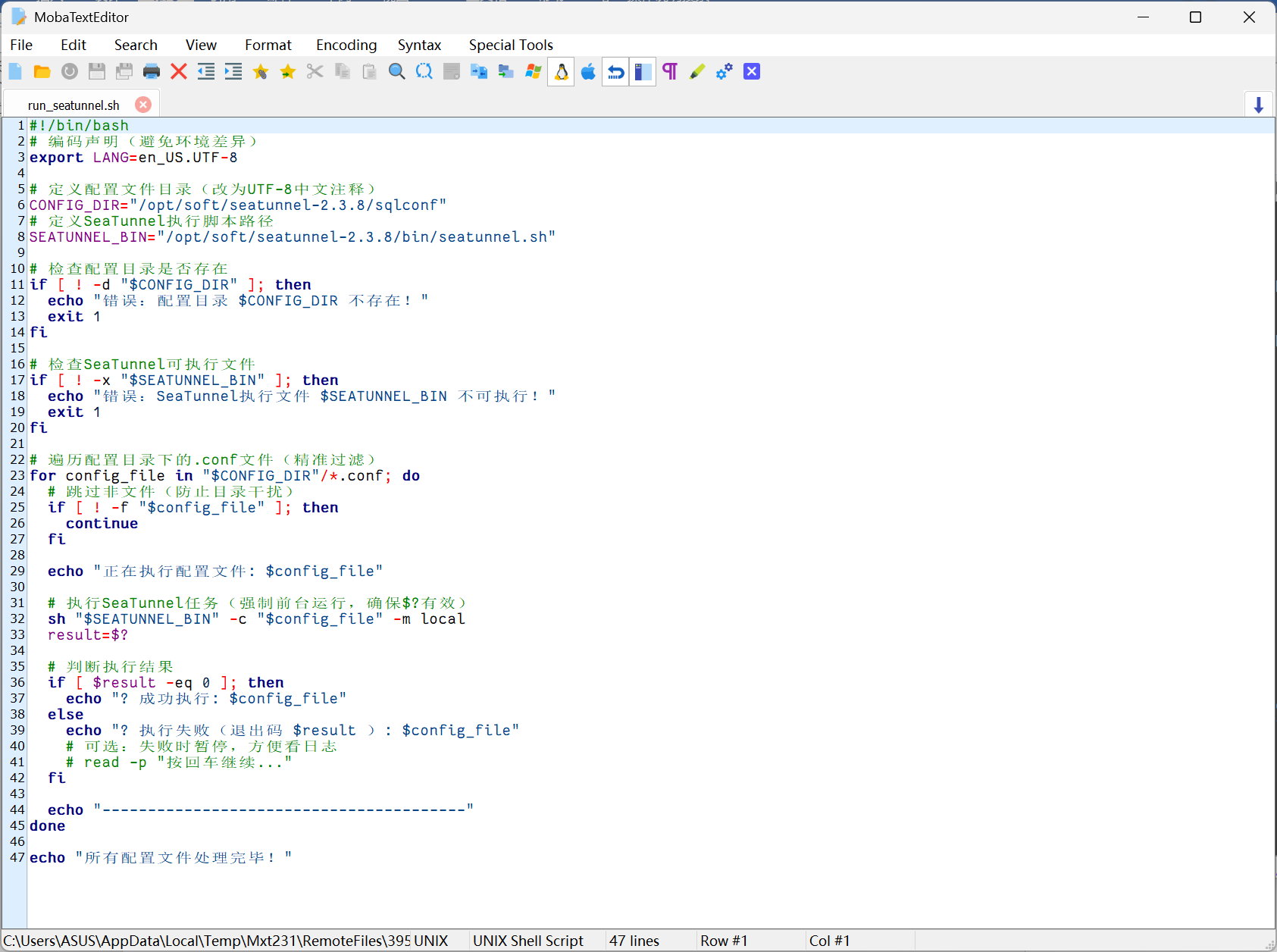
[Seatunnel 同步任务（Python生成配置 + Shell调度）]

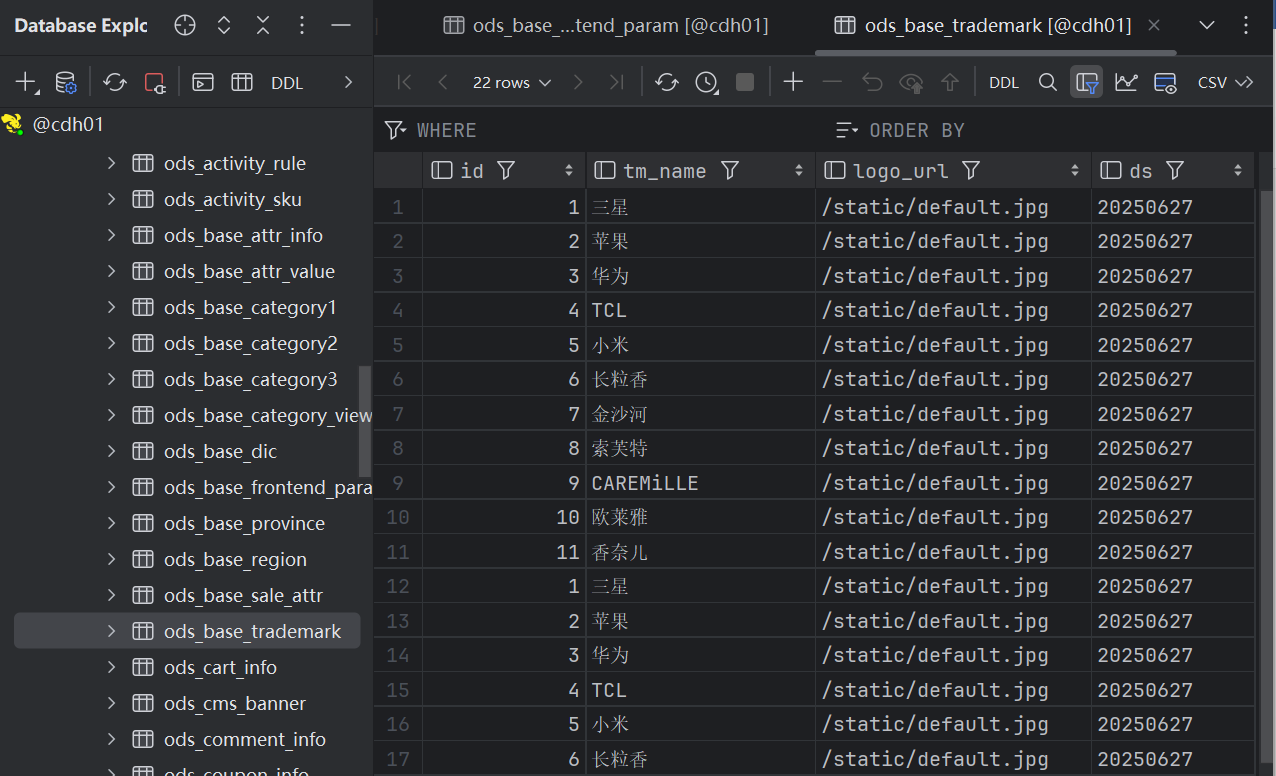
↓

[Hive（ODS 层）]









数据加工

总线矩阵设计

总线矩阵（Bus Matrix）是数仓设计中，用于规划“业务主题”与“维度/事实表”关系的二维表格，它决定了你的数仓主题粒度和维度建模方向。

总线矩阵是一个二维表格：

* \*\*行（横向）\*\*表示业务过程（即“事实”或“主题”），如订单、支付、发货、退款、用户行为等。
* \*\*列（纵向）\*\*表示维度，如时间、用户、商品、门店、渠道、地区等。



|  |  |  |  |
| --- | --- | --- | --- |
| **维度分类** | **维度表名** | **描述** | **是否使用** |
| 用户维 | dim\_user\_full | 用户ID、注册时间、性别、年龄、等级等 | ✔️ |
| 商品维 | dim\_sku\_full | 商品ID、名称、类目、品牌、规格等 | ✔️ |
| 时间维 | dim\_date | 日期ID、年、月、季度、星期、节假日等 | ✔️ |
| 地区维 | dim\_province\_full | 地区ID、省、市、区、行政等级等 | ✔️ |
| 活动维 | dim\_activity\_full | 活动ID、类型、开始/结束时间、预算等 | ✔️ |
| 渠道维 | dim\_channel\_full | 渠道类型、来源终端（App、小程序、PC等） | ✔️ |
| 优惠券维 | dim\_coupon\_full | 优惠券ID、面额、使用条件、类型 | ✔️ |
| 日期维 | dim\_date | 下单时间、支付时间、发货时间 | ✔️ |

|  |  |  |
| --- | --- | --- |
| 事实表名 | 粒度 | 描述 |
| dwd\_trade\_order\_detail | 一行一条明细 | DWD层，订单明细级别的交易记录 |
| dws\_trade\_user\_order\_1d | 用户-天 | 用户每日下单情况（下单数、金额） |
| dws\_trade\_sku\_order\_1d | 商品-天 | 商品每日销售汇总（销量、销售额） |
| dws\_trade\_order\_1d | 订单-天 | 订单整体视角每日统计 |
| ads\_trade\_order\_stats | 汇总统计层 | 每日订单概况，供报表使用 |

DWD 层原子指标梳理表

|  |  |  |
| --- | --- | --- |
| **表名** | **原子指标字段** | **指标说明** |
| dwd\_trade\_cart\_add\_inc | sku\_num | 加购商品件数 |
| dwd\_trade\_order\_detail\_inc | sku\_num  split\_original\_amount  split\_activity\_amount  split\_coupon\_amount  split\_total\_amount | 下单商品件数原始价格活动优惠金额优惠券优惠金额最终支付价格 |
| dwd\_trade\_pay\_detail\_suc\_inc | sku\_numsplit\_original\_amount  split\_activity\_amount  split\_coupon\_amount  split\_payment\_amount | 支付商品件数原始价格活动优惠金额优惠券优惠金额实际支付金额 |
| dwd\_trade\_cart\_full | sku\_num | 当前购物车中商品件数 |
| dwd\_trade\_trade\_flow\_acc | order\_original\_amount  order\_activity\_amount  order\_coupon\_amount  order\_total\_amount  payment\_amount | 下单原始金额下单活动优惠下单优惠券金额下单应付金额实际支付金额 |
| dwd\_tool\_coupon\_used\_inc | —— | 主要是行为记录，无数量类指标 |
| dwd\_interaction\_favor\_add\_inc | —— | 收藏行为记录，不含指标 |
| dwd\_traffic\_page\_view\_inc | during\_time | 页面浏览持续时间（毫秒） |
| dwd\_user\_register\_inc | —— | 用户注册行为记录 |
| dwd\_user\_login\_inc | —— | 用户登录行为记录 |

DWS衍生指标明细表



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **衍生指标名** | **所在表名** | **粒度** | **计算逻辑** | **来源字段（明细表或上游汇总表）** |
| order\_count\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | count(\*) | dwd\_trade\_order\_detail\_inc |
| order\_count\_1d | dws\_trade\_user\_order\_1d | 用户 | count(distinct order\_id) | dwd\_trade\_order\_detail\_inc |
| order\_count\_1d | dws\_trade\_province\_order\_1d | 省份 | count(distinct order\_id) | dwd\_trade\_order\_detail\_inc |
| order\_num\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | sum(sku\_num) | dwd\_trade\_order\_detail\_inc |
| order\_num\_1d | dws\_trade\_user\_order\_1d | 用户 | sum(sku\_num) | dwd\_trade\_order\_detail\_inc |
| order\_num\_1d | dws\_trade\_province\_order\_1d | 省份 | sum(sku\_num) | dwd\_trade\_order\_detail\_inc |
| order\_original\_amount\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | sum(split\_original\_amount) | dwd\_trade\_order\_detail\_inc |
| order\_original\_amount\_1d | dws\_trade\_user\_order\_1d | 用户 | sum(split\_original\_amount) | dwd\_trade\_order\_detail\_inc |
| order\_original\_amount\_1d | dws\_trade\_province\_order\_1d | 省份 | sum(split\_original\_amount) | dwd\_trade\_order\_detail\_inc |
| activity\_reduce\_amount\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | sum(nvl(split\_activity\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| activity\_reduce\_amount\_1d | dws\_trade\_user\_order\_1d | 用户 | sum(nvl(split\_activity\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| activity\_reduce\_amount\_1d | dws\_trade\_province\_order\_1d | 省份 | sum(nvl(split\_activity\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| coupon\_reduce\_amount\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | sum(nvl(split\_coupon\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| coupon\_reduce\_amount\_1d | dws\_trade\_user\_order\_1d | 用户 | sum(nvl(split\_coupon\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| coupon\_reduce\_amount\_1d | dws\_trade\_province\_order\_1d | 省份 | sum(nvl(split\_coupon\_amount, 0)) | dwd\_trade\_order\_detail\_inc |
| order\_total\_amount\_1d | dws\_trade\_user\_sku\_order\_1d | 用户 + SKU | sum(split\_total\_amount) | dwd\_trade\_order\_detail\_inc |
| order\_total\_amount\_1d | dws\_trade\_user\_order\_1d | 用户 | sum(split\_total\_amount) | dwd\_trade\_order\_detail\_inc |
| order\_total\_amount\_1d | dws\_trade\_province\_order\_1d | 省份 | sum(split\_total\_amount) | dwd\_trade\_order\_detail\_inc |
| cart\_add\_count\_1d | dws\_trade\_user\_cart\_add\_1d | 用户 | count(\*) | dwd\_trade\_cart\_add\_inc |
| cart\_add\_num\_1d | dws\_trade\_user\_cart\_add\_1d | 用户 | sum(sku\_num) | dwd\_trade\_cart\_add\_inc |
| payment\_count\_1d | dws\_trade\_user\_payment\_1d | 用户 | count(distinct order\_id) | dwd\_trade\_pay\_detail\_suc\_inc |
| payment\_num\_1d | dws\_trade\_user\_payment\_1d | 用户 | sum(sku\_num) | dwd\_trade\_pay\_detail\_suc\_inc |
| payment\_amount\_1d | dws\_trade\_user\_payment\_1d | 用户 | sum(split\_payment\_amount) | dwd\_trade\_pay\_detail\_suc\_inc |
| favor\_add\_count\_1d | dws\_interaction\_sku\_favor\_add\_1d | SKU | count(\*) | dwd\_interaction\_favor\_add\_inc |
| during\_time\_1d | dws\_traffic\_session\_page\_view\_1d | 会话 | sum(during\_time) | dwd\_traffic\_page\_view\_inc |
| page\_count\_1d | dws\_traffic\_session\_page\_view\_1d | 会话 | count(\*) | dwd\_traffic\_page\_view\_inc |
| view\_count\_1d | dws\_traffic\_page\_visitor\_page\_view\_1d | 访客 + 页面 | count(\*) | dwd\_traffic\_page\_view\_inc |
| order\_count\_7d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 过去7天 order\_count\_1d 累加 | dws\_trade\_user\_sku\_order\_1d |
| order\_num\_7d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 过去7天 order\_num\_1d 累加 | dws\_trade\_user\_sku\_order\_1d |
| order\_total\_amount\_7d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 过去7天 order\_total\_amount\_1d | dws\_trade\_user\_sku\_order\_1d |
| order\_count\_30d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 最近30天 order\_count\_1d 累加 | dws\_trade\_user\_sku\_order\_1d |
| order\_num\_30d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 最近30天 order\_num\_1d 累加 | dws\_trade\_user\_sku\_order\_1d |
| order\_total\_amount\_30d | dws\_trade\_user\_sku\_order\_nd | 用户 + SKU | 最近30天 order\_total\_amount\_1d | dws\_trade\_user\_sku\_order\_1d |
| login\_count\_td | dws\_user\_user\_login\_td | 用户 | 历史+当日登录次数累计 | dwd\_user\_login\_inc + 历史累加 |
| order\_count\_td | dws\_trade\_user\_order\_td | 用户 | 历史+当日订单数累计 | dws\_trade\_user\_order\_1d + 累加 |
| order\_date\_first | dws\_trade\_user\_order\_td | 用户 | 首次下单日期（最小值） | dws\_trade\_user\_order\_1d |
| order\_date\_last | dws\_trade\_user\_order\_td | 用户 | 最近下单日期（最大值） | dws\_trade\_user\_order\_1d |
| used\_count\_1d | dws\_tool\_user\_coupon\_coupon\_used\_1d | 用户 + 优惠券 | count(\*) | dwd\_tool\_coupon\_used\_inc |

ADS派生指标明细表



|  |  |  |  |
| --- | --- | --- | --- |
| **指标名称** | **所属表** | **来源字段/表** | **派生逻辑说明** |
| uv\_count | ads\_traffic\_stats\_by\_channel | count(distinct mid\_id) from dws\_traffic\_session\_page\_view\_1d | 基于会话页面视图中不同设备ID计数 |
| avg\_duration\_sec | ads\_traffic\_stats\_by\_channel | avg(during\_time\_1d)/1000 from dws\_traffic\_session\_page\_view\_1d | 将毫秒转为秒 |
| avg\_page\_count | ads\_traffic\_stats\_by\_channel | avg(page\_count\_1d) from dws\_traffic\_session\_page\_view\_1d | 会话浏览页数平均值 |
| sv\_count | ads\_traffic\_stats\_by\_channel | count(\*) from dws\_traffic\_session\_page\_view\_1d | 会话总数 |
| bounce\_rate | ads\_traffic\_stats\_by\_channel | sum(if(page\_count\_1d=1,1,0))/count(\*) | 只浏览一个页面的会话占比 |
| user\_churn\_count | ads\_user\_change | login\_date\_last = date\_add(dt, -7) from dws\_user\_user\_login\_td | 最近登录时间等于一周前，当前未活跃 |
| user\_back\_count | ads\_user\_change | 差值 >= 8 天（回流判断） | 当前有登录，前一天没登录，且上次登录距今大于7天 |
| new\_user\_count | ads\_user\_stats | login\_date\_first >= date\_add(dt, -recent\_days+1) | 计算在最近N日注册的新用户 |
| active\_user\_count | ads\_user\_stats | login\_date\_last >= date\_add(dt, -recent\_days+1) | 计算最近N日有登录行为的用户 |
| order\_repeat\_rate | ads\_repeat\_purchase\_by\_tm | sum(order\_count>=2)/sum(order\_count>=1) | 每个品牌中下单次数大于等于2的用户占比 |
| order\_user\_count | ads\_order\_stats\_by\_tm / by\_cate / by\_province | count(distinct user\_id) | 同一时间段内对品牌、品类或省份有下单行为的用户数 |
| cart\_num | ads\_sku\_cart\_num\_top3\_by\_cate | sum(sku\_num) from dwd\_trade\_cart\_full | SKU 加购数量汇总 |
| favor\_count | ads\_sku\_favor\_count\_top3\_by\_tm | favor\_add\_count\_1d from dws\_interaction\_sku\_favor\_add\_1d | SKU 单日收藏数 |
| order\_to\_pay\_interval\_avg | ads\_order\_to\_pay\_interval\_avg | avg(to\_unix\_timestamp(payment\_time) - to\_unix\_timestamp(order\_time)) | 计算下单到支付的平均时间（秒） |
| order\_total\_amount | ads\_order\_by\_province | 来自 dws 的 order\_total\_amount\_1d/7d/30d | 分区不同时间窗口汇总金额 |
| used\_count | ads\_coupon\_stats | sum(used\_count\_1d) from dws\_tool\_user\_coupon\_coupon\_used\_1d | 优惠券使用次数 |
| used\_user\_count | ads\_coupon\_stats | count(\*) from dws\_tool\_user\_coupon\_coupon\_used\_1d | 优惠券使用用户数 |
| path\_count | ads\_page\_path | count(\*) over source -> target 路径 | 用户页面跳转路径分析 |
| order\_continuously\_user\_count | ads\_order\_continuously\_user\_count | 连续三天 dt 差值为 2 | 连续三天下单的用户数统计 |
| home\_count | ads\_user\_action | sum(if(page\_id='home',1,0)) from dws\_traffic\_page\_visitor\_page\_view\_1d | 浏览首页人数 |
| good\_detail\_count | ads\_user\_action | sum(if(page\_id='good\_detail',1,0)) from dws\_traffic\_page\_visitor\_page\_view\_1d | 浏览商品详情页人数 |

数仓分层建模



| 分层目的 | 具体好处 |
| --- | --- |
| 解耦 | 各层关注点不同，方便维护与扩展 |
| 标准化 | 数据口径一致，统一数据来源 |
| 复用性强 | 下游多个报表、模型可共用上游清洗结果 |
| 数据质量可控 | 分层有助于定位问题、追溯源头 |
| 权限清晰 | 不同角色可访问不同层级的数据 |
|  |  |

DIM 层

DIM（Dimension）层 是数仓模型中用于描述业务实体属性的维度表集合，主要作用是提供数据分析时的“维度上下文”和“查询粒度”。

它的主要作用包括：

1. 维度扩展：描述事实数据的业务背景（如：用户、商品、活动、时间、区域等）；
2. 降维查询：配合事实表进行数据聚合分析、分组统计、透视报表等；
3. 维持一致性：统一全域维度定义，防止口径不一致（维表建模是总线矩阵的核心）；
4. 缓慢变化处理：处理 SCD（Slowly Changing Dimension），支持历史回溯分析。

DIM 层的来源

DIM 层数据通常来自于 ODS 层（业务主表）或第三方系统，经过清洗、建模、去重、拉链处理等步骤构建而成。

来源示例：

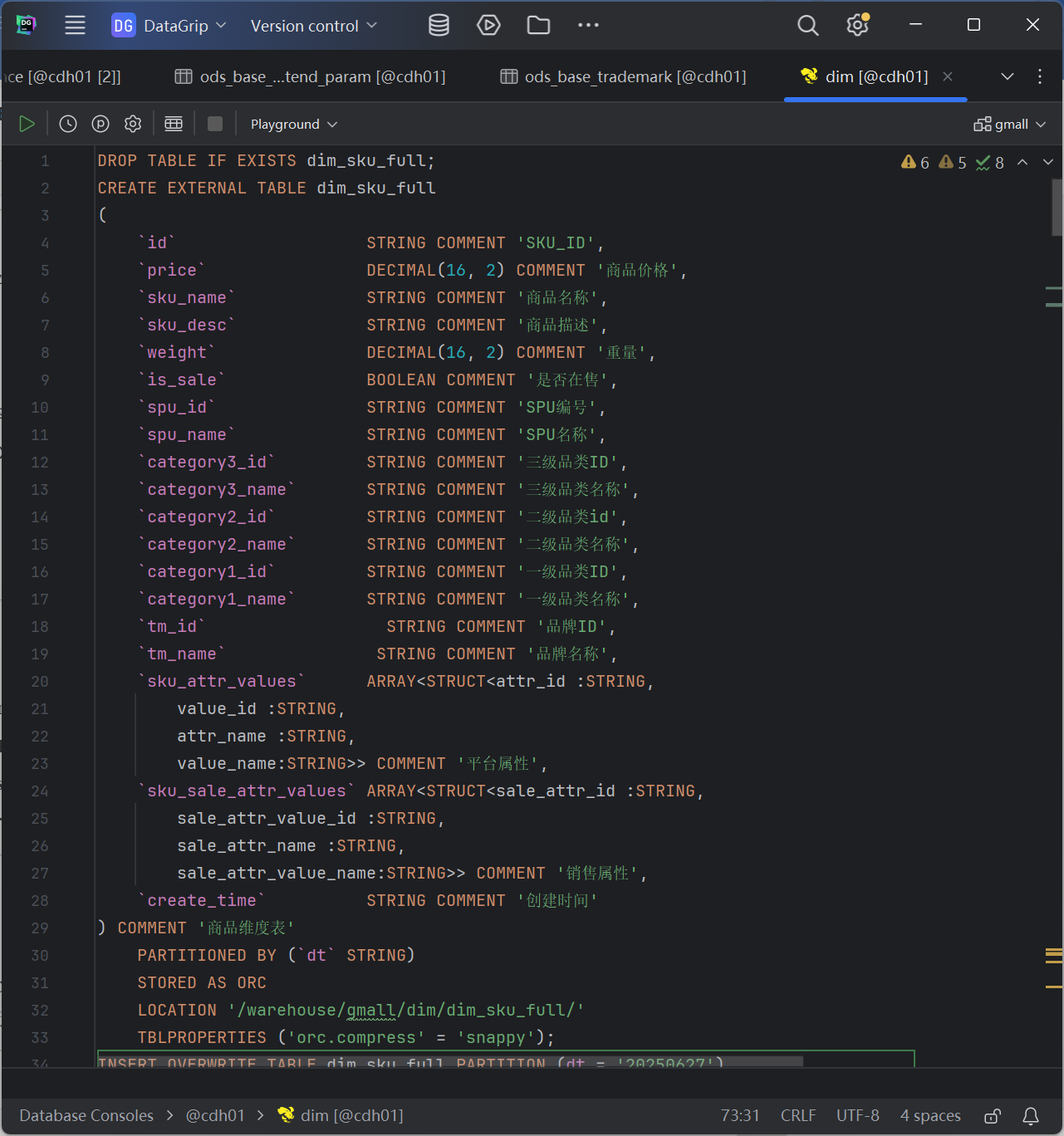
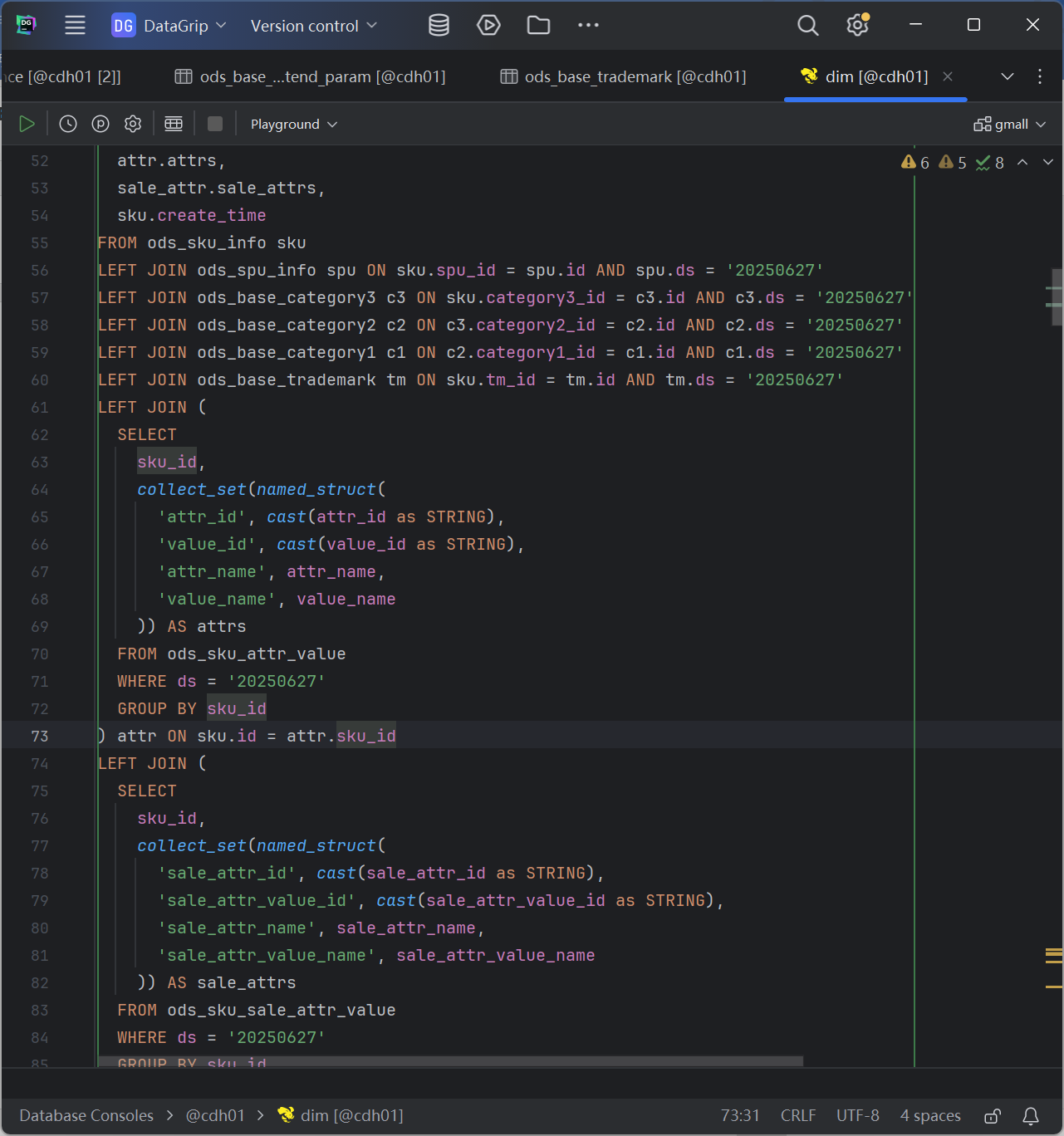
| DIM 表名 | 来源说明 |
| --- | --- |
| dim\_user\_full | ODS 中的用户主数据表（注册表、账户表等） |
| dim\_sku\_full | 商品主数据（商品基础表 + 品类表 + 品表） |
| dim\_coupon\_full | 优惠券基础信息表 |
| dim\_date | 系统生成的日期维表 |
| dim\_province\_full | 行政区域划分 |

DIM 层分类（按建模方式）

| 类型 | 说明 | 示例 |
| --- | --- | --- |
| 静态维度表 | 数据不变，如日期、区域 | dim\_date, dim\_province |
| 常规维度表 | 取主表 + 字典表清洗得到 | dim\_user\_full, dim\_sku\_full |
| 拉链维度表 | 处理历史变化，保留版本 | dim\_user\_zip，dim\_sku\_zip |
| 枚举/字典维度 | 状态、类型、来源字典等 | dim\_status, dim\_source\_type |

维度表设计要点

* 要有业务主键（natural key）；
* 尽量结构宽、字段多，支持各种查询；
* 保留有效时间字段：start\_date, end\_date, is\_current（SCD）；
* 命名规范：dim\_xxx\_full（当前快照），dim\_xxx\_zip（历史拉链）



DWD 是明细数据层，位于 ODS 和 DWS 之间，承担着“承上启下”的关键作用。它的主要职责包括：

1. 数据清洗和标准化

* 去除重复、脏数据
* 统一字段格式（如时间格式、金额单位等）
* 字段类型转换、命名规范统一

例如：将 create\_time 字段统一格式为 yyyy-MM-dd HH:mm:ss

2. 维度补充与退化

* 将事实表与维度表进行 Join 补全维度信息
* 生成“宽表”，便于后续聚合

如：订单事实表关联用户维度表，补充用户性别、年龄段、省份等

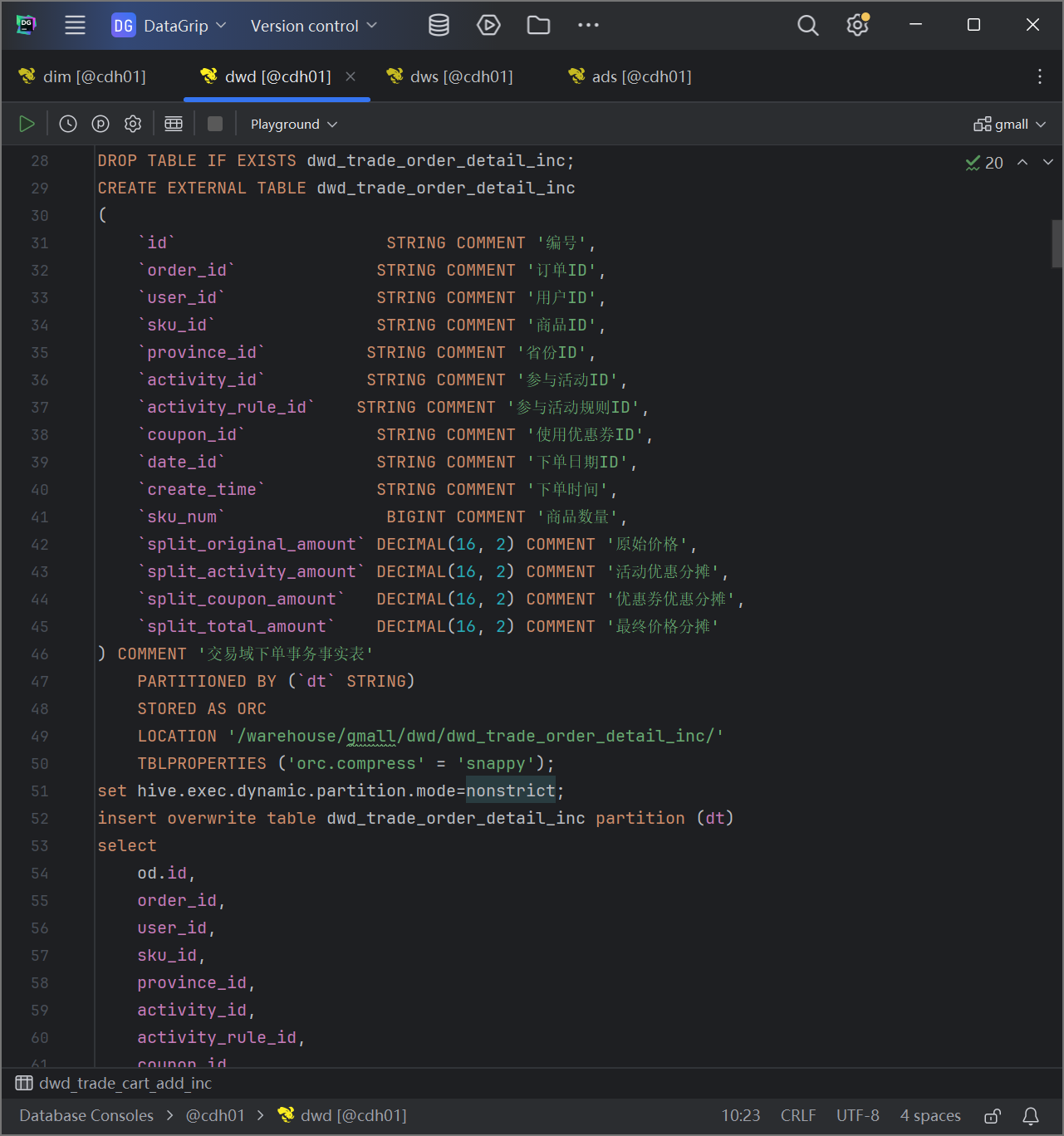
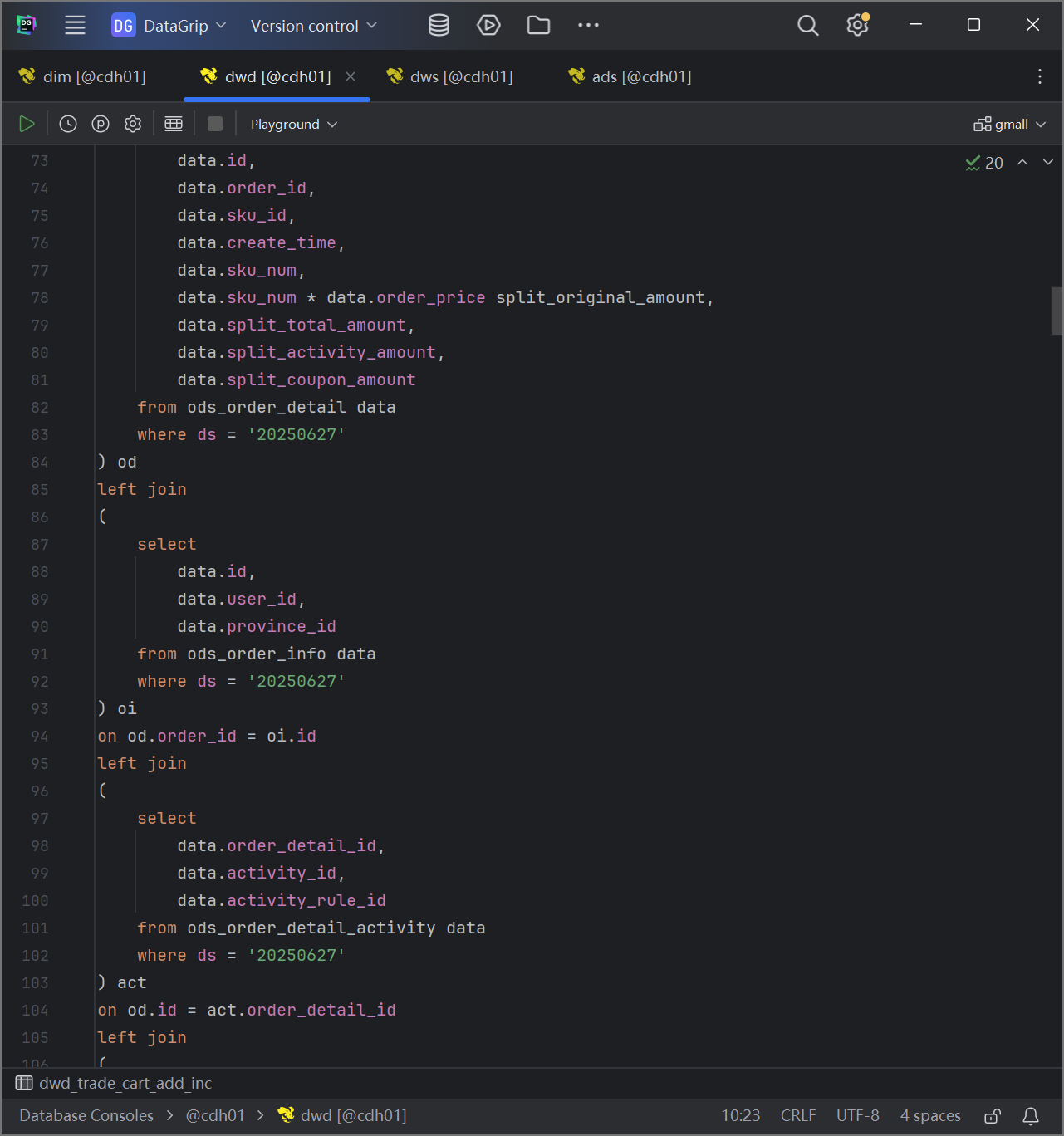
、

3. 统一粒度

* 把原始数据按业务实体和粒度进行统一，如“订单明细粒度”“用户行为粒度”等
* 不同来源但相同业务含义的数据在此合并

4. 构建事实宽表（可复用）

* 构建多个可复用的事实表，如订单表、加购表、支付表、收藏表等
* 支撑后续 DWS 层的聚合分析



DWD 层的数据来源（dwd\_trade\_order\_detail\_inc）

来源表：

* 订单主表：ods\_order\_info\_inc
* 订单明细表：ods\_order\_detail\_inc
* 商品维度：dim\_sku\_full
* 时间维度：dim\_date
* 用户维度：dim\_user\_zip

| 操作类型 | 示例说明 |
| --- | --- |
| 清洗 | 筛选 dt 分区，去空值 |
| 关联 | 订单主表 join 明细表 join SKU维度 |
| 字段转换 | 时间戳 → yyyy-MM-dd HH:mm:ss |
| 构建宽表 | 汇总订单金额、数量、用户、商品等字段 |

DWS 层是 汇总数据层，也称为 中间层 / 主题层，它将 DWD 明细数据进行归类聚合、主题建模，支撑 ADS 层的指标复用和可视化呈现。

主要作用如下：

1. 基于主题的数据整合和汇总

* DWS 层按照业务域（交易、用户、商品、流量等）进行组织
* 将 DWD 层的明细表聚合为用户级、SKU级、品牌级等不同粒度的宽表

例如：将每天的订单明细聚合成“每日用户下单统计表（dws\_trade\_user\_order\_1d）”

2. 支持多种粒度的数据分析

DWS 层一般以以下时间粒度组织：

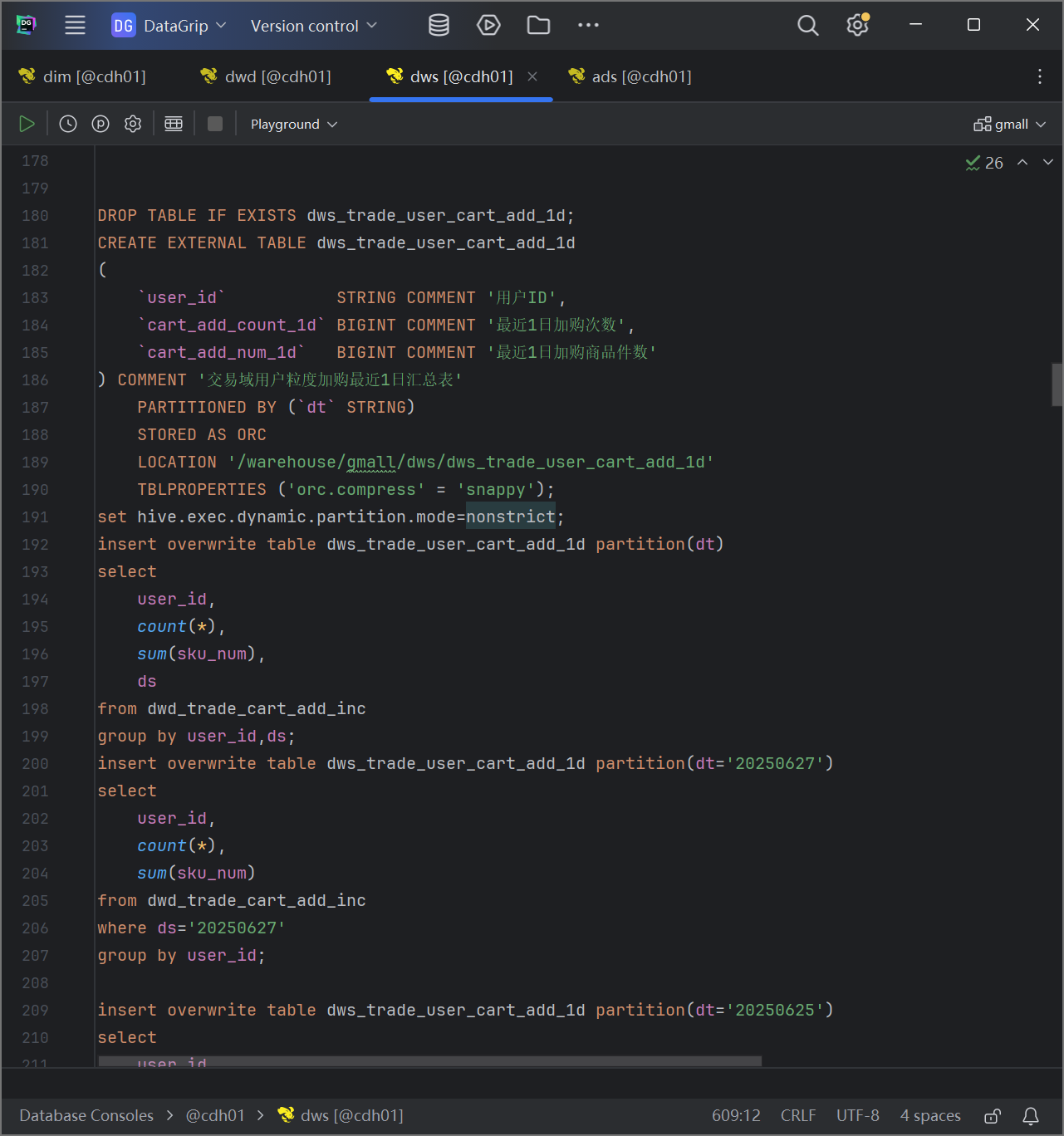
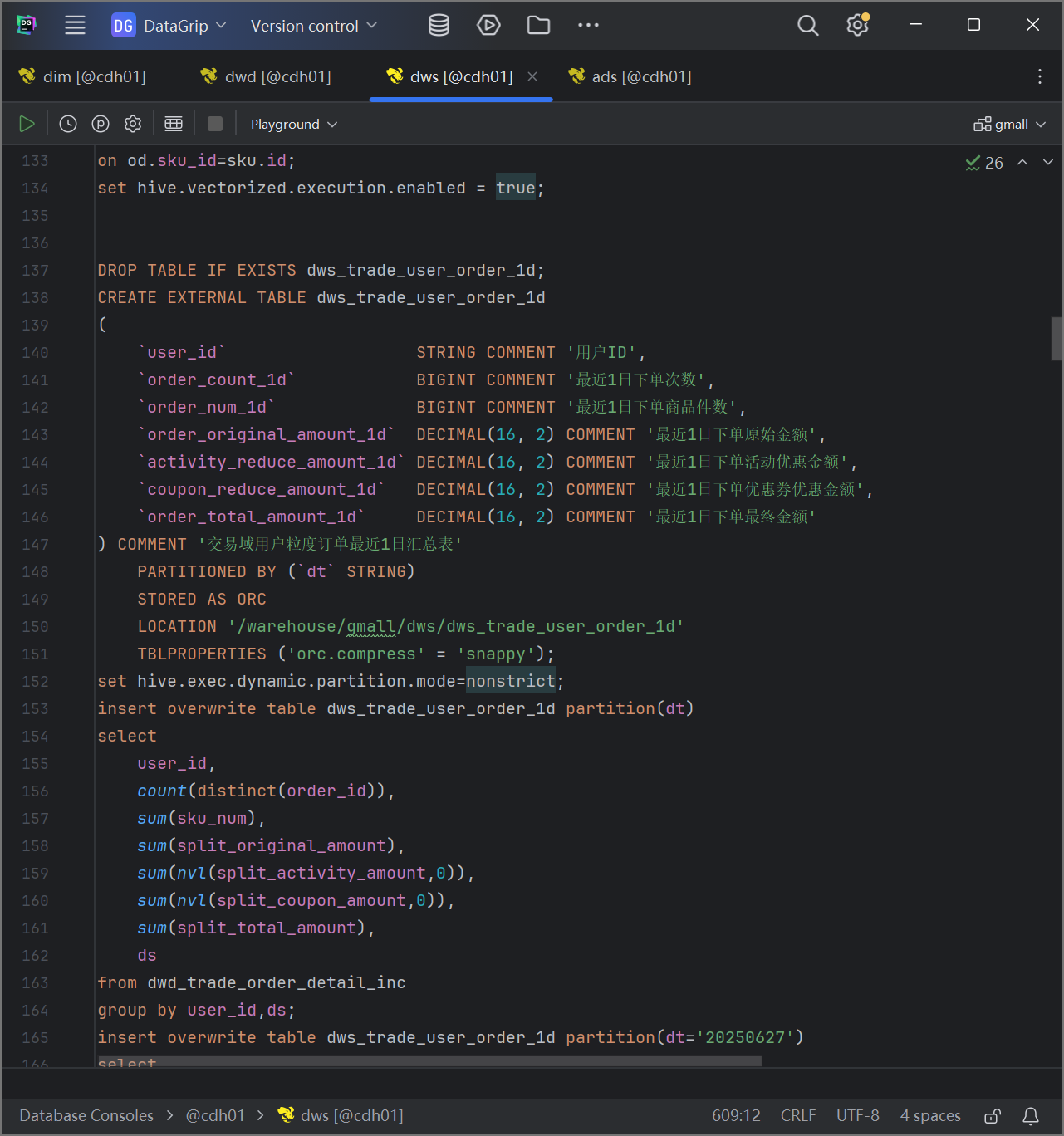
| 粒度 | 说明 |
| --- | --- |
| 1d | 按天汇总 |
| nd | 按最近 N 天滑窗汇总，如 7 天、30 天 |
| td | 截至某日的累积汇总（to-date） |

3. 打通主题之间的数据链路

* 如：用户下单行为 + 用户浏览行为 → 支撑漏斗分析
* 品牌购买行为 + 商品收藏行为 → 支撑复购率/热度分析

4. 提升复用性，简化 ADS 层开发

DWS 层构建的宽表被多个指标/图表复用，减少重复开发，保障一致性。



| 来源层 | 内容说明 |
| --- | --- |
| DWD 层明细表 | 各类经过清洗标准化的事实表 |
| DIM 层维度表 | 品类、品牌、商品、时间等维度，用于 enrich |
| ODS 层少量辅助数据 | 如最新的用户注册信息、部分外部表 |

以分析粒度组织

* 用户级：dws\_trade\_user\_order\_1d、dws\_interaction\_user\_favor\_add\_1d
* SKU级：dws\_trade\_sku\_order\_1d
* 品牌级：dws\_trade\_tm\_order\_nd
* 渠道级：dws\_traffic\_channel\_uv\_count\_nd

| 表名示例 | 粒度 | 说明 |
| --- | --- | --- |
| dws\_trade\_user\_order\_1d | 用户粒度，1天 | 每日每用户下单汇总 |
| dws\_trade\_user\_order\_nd | 用户粒度，n天 | 最近 n 天下单统计 |
| dws\_trade\_sku\_order\_1d | SKU 粒度，1天 | 每日每 SKU 的销量 |
| dws\_traffic\_session\_page\_view\_1d | 会话粒度，1天 | 页面浏览路径分析 |

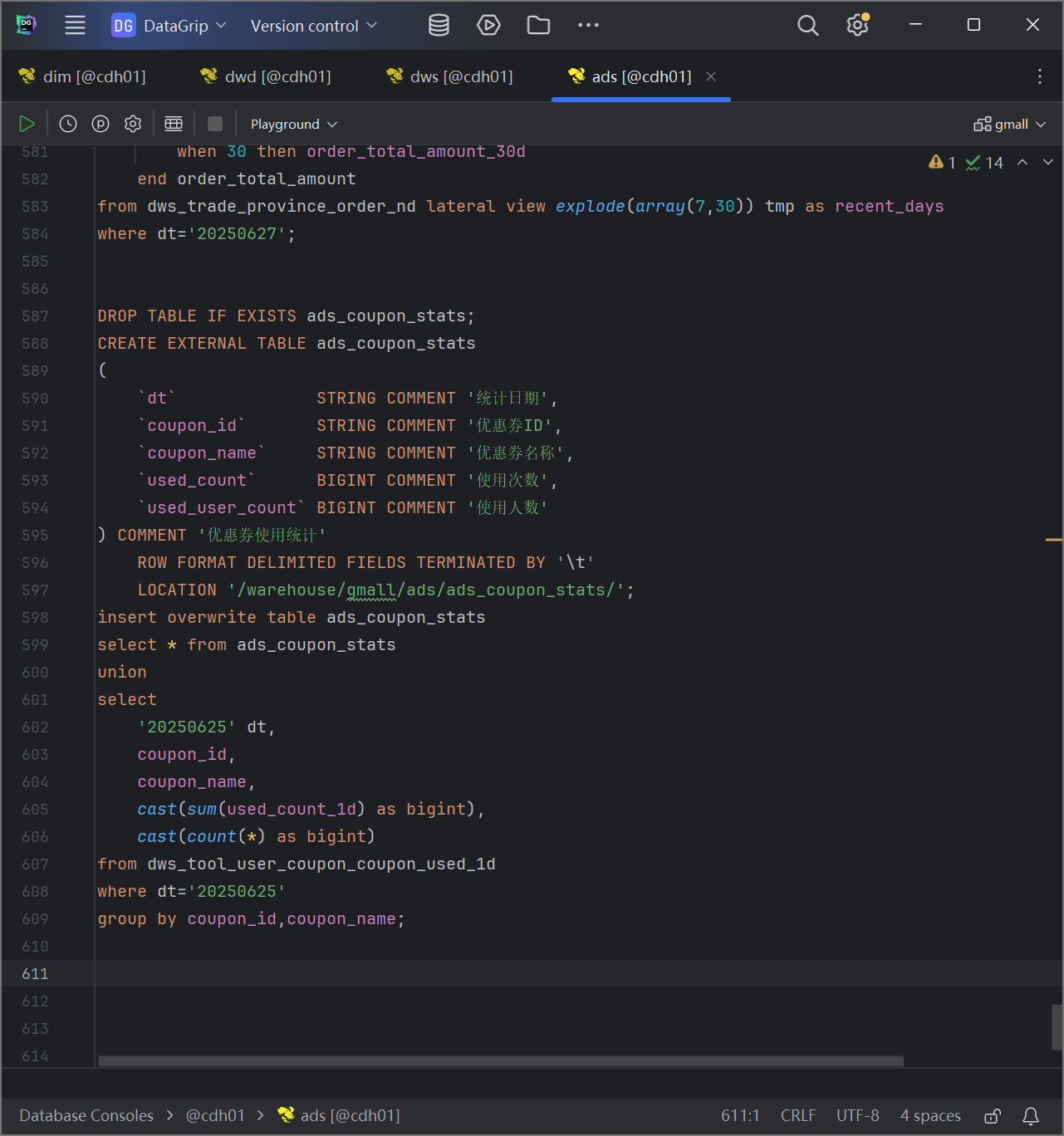
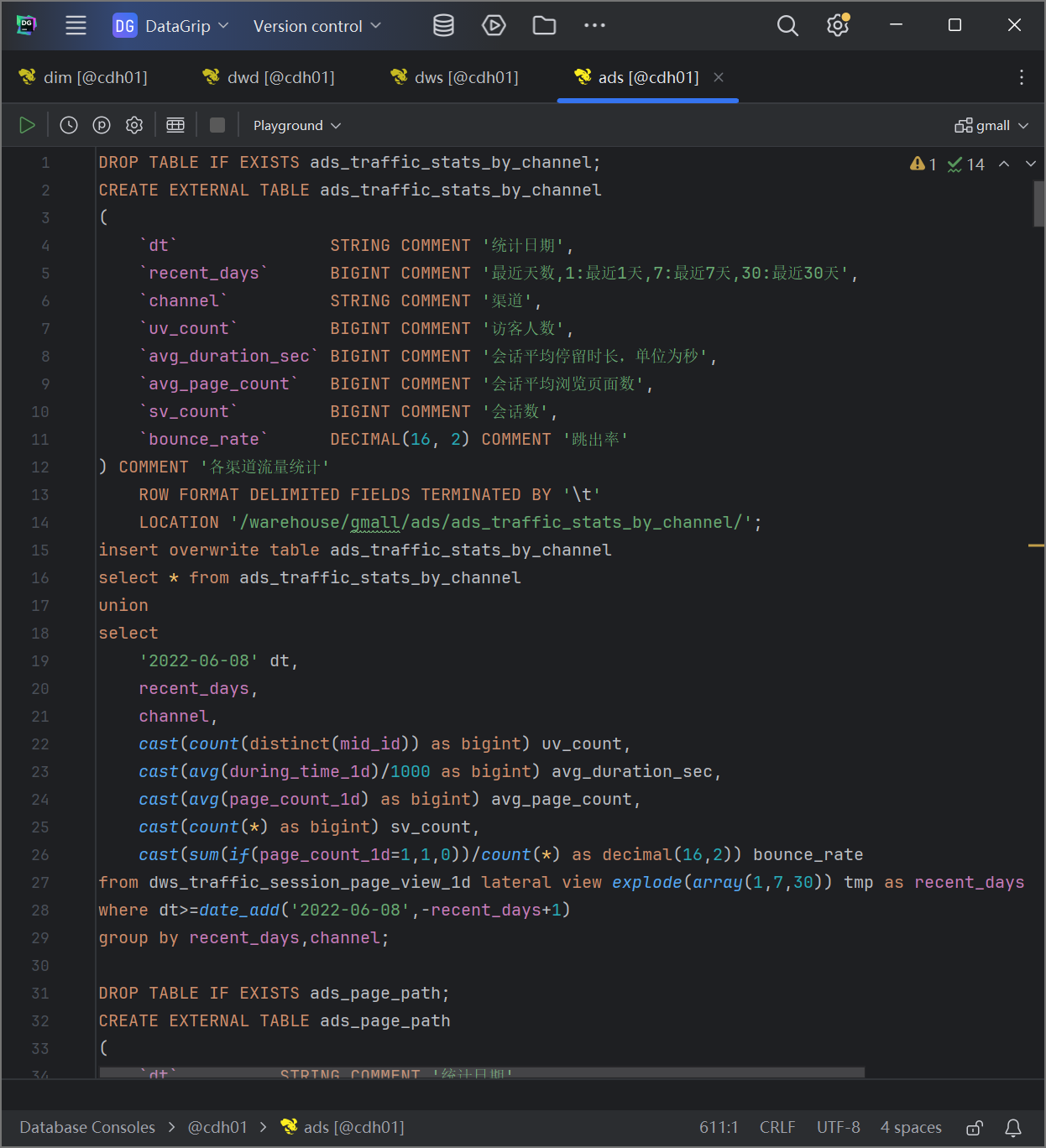
ADS 层是面向业务的指标层，直接为业务部门（产品、运营、市场等）提供数据支持，是数据仓库价值的最终体现。

核心作用如下：

| 作用 | 说明 |
| --- | --- |
| 面向应用 | 以“某个场景”或“指标体系”为目标而建表，例如复购率、活跃人数、销售排行榜等 |
| 面向报表 | 每个 ADS 表可以直接作为报表数据源，无需额外计算 |
| 强指标表达 | 聚焦“一个主题下的一组业务指标” |
| 灵活迭代 | 随业务变化灵活调整，允许冗余字段、缓存逻辑 |

示例场景与表设计：

| 场景 | 表名 | 指标 |
| --- | --- | --- |
| 用户留存分析 | ads\_user\_change | 回流用户数、流失用户数 |
| 用户行为漏斗 | ads\_user\_action | 浏览人数、加购人数、支付人数 |
| 品牌分析 | ads\_order\_stats\_by\_tm | 下单数、下单用户数 |
| 地域销售统计 | ads\_order\_by\_province | 各省订单数、订单金额 |
| 营销转化 | ads\_coupon\_stats | 优惠券使用人数、次数 |

ADS 层的数据来源

| 来源层级 | 说明 |
| --- | --- |
| DWS 层 | 作为主要数据源，提供已经聚合好的宽表（1d/nd/td 粒度） |
| DIM 层 | 部分维度表辅助展示字段，如品牌名、品类名 |
| DWD 层（少量） | 如需更细粒度加工的个别逻辑，也可能直接引用明细表 |
| 外部接口 | 如数据同步到 Redis/Doris/ES 进行 API 服务或大屏展示 |

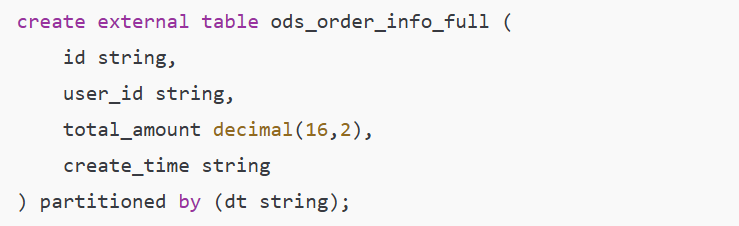
| 表名 | 统计口径 | 样例指标字段 |
| --- | --- | --- |
| ads\_user\_action | 漏斗行为 | 首页访问人数、加购人数、支付人数 |
| ads\_repeat\_purchase\_by\_tm | 品牌维度复购 | tm\_id, 复购率 |
| ads\_order\_to\_pay\_interval\_avg | 时间间隔 | 下单到支付平均时长（秒） |
| ads\_coupon\_stats | 营销统计 | 使用次数、使用人数 |
| ads\_sku\_favor\_count\_top3\_by\_tm | 品牌 TOP 收藏商品 | favor\_count, rk |

|  |  |  |  |
| --- | --- | --- | --- |
| **层级** | **全称** | **设计目标** | **特征** |
| ODS | Operational Data Store | 保留原始数据，结构与数据源一致 | 最细粒度，抽取即加载，无业务加工 |
| DWD | Data Warehouse Detail | 明细建模，按业务主题标准化存储 | 做数据清洗、去重、维度补充 |
| DWS | Data Warehouse Summary | 汇总聚合层，为ADS做准备 | 多维聚合，时间/空间/用户等维度 |
| ADS | Application Data Store | 直接服务报表和应用 | 多为宽表、轻聚合、含派生指标 |

ETL 过程说明

ODS 层（数据接入层）

* 目标：完整保留原始数据（通常为日志或业务库的快照）
* 实现方式：Seatunnel 或 CDC 工具同步
* 典型操作：追加方式写入；不做加工
* 命名规则：ods\_{source}\_{table\_name}\_inc/full



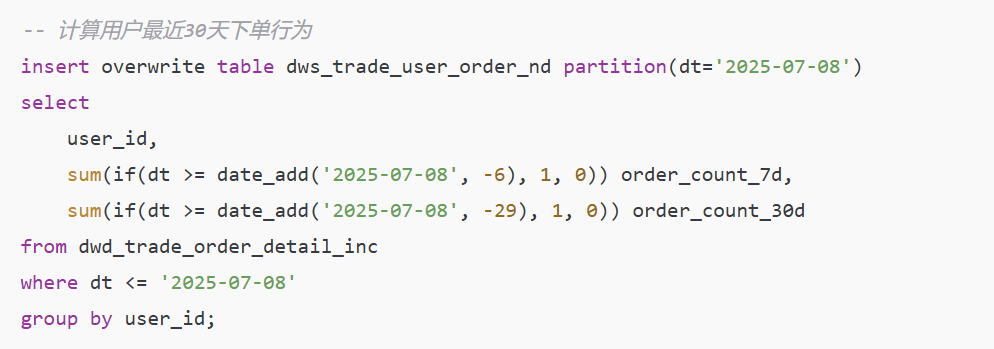
DWD 层（明细建模层）

* 目标：清洗、关联维度、统一粒度，形成原子宽表
* 典型处理：时间格式标准化、空值处理、维度拉链表关联
* 命名规则：dwd\_{domain}\_{entity}\_{inc/full}



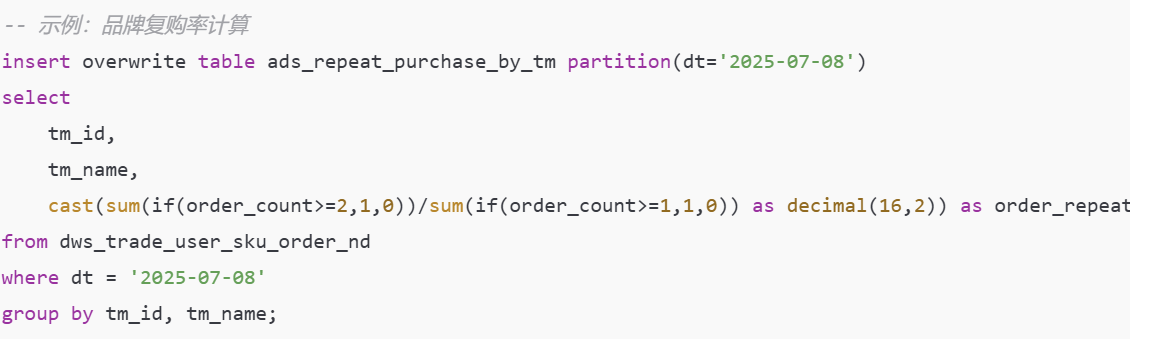
DWS 层（汇总计算层）

* 目标：按时间/空间/人群等多维进行聚合，减少ADS计算压力
* 特征：按主题域进行设计（如交易域、流量域等）
* 命名规则：dws\_{domain}\_{subject}\_{granularity}



ADS 层（应用展示层）

* 目标：为报表/API服务提供快速响应的宽表或指标数据
* 特征：多数是轻聚合 + 派生指标（如转化率、复购率等）
* 命名规则：ads\_{subject}\_stats 或 ads\_{domain}\_summary



代码样例

ODS → DWD



DWD → DWS



DWS → ADS



| 层级 | 是否建模 | 加工内容 | 目标作用 |
| --- | --- | --- | --- |
| ODS | 否 | 直接同步原始数据 | 数据备份与追溯 |
| DWD | 是 | 清洗+维度退/拉链+宽表构建 | 提供标准、统一的原子表 |
| DWS | 是 | 聚合+主题域设计 | 高复用指标、支持多场景分析 |
| ADS | 否 | 可视化报表/派生指标加工 | 报表、应用的最终指标展现 |

是否使用增量的判断标准

|  |  |  |
| --- | --- | --- |
| **判断项** | **说明** | **适合使用** |
| 1. 是否存在可追踪变更字段（如 update\_time） | 表中有 create\_time / update\_time 等字段可以标记新/变更数据 | 增量（强烈建议） |
| 2. 是否是 append-only 表（追加型） | 数据一旦写入就不会更新，比如日志、交易明细 | 增量（按分区追加） |
| 3. 数据是否经常发生历史回写（补录/修改） | 比如订单表存在“取消、退单、支付状态变更”等 | 不适合纯增量，要考虑 “变更捕捉”或 全量覆盖分区 |
| 4. 数据量是否较大（每日百万级以上） | 全量成本高，ETL耗时长，资源浪费 | 增量优先 |
| 5. 是否存在主键/业务主键用于去重 | 增量场景下需要主键保证幂等性 | 支持 增量写入 |
| 6. 是否为维度表 | - 静态维表：可以全量- 慢变维表（SCD2）：必须做增量合并 | 两者都可能 |
| 7. 是否为 ADS 报表层数据 | 通常按天重算，不做精确追踪 | 覆盖式写入（伪增量） |

业务场景

|  |  |  |
| --- | --- | --- |
| **表类型** | **推荐抽取方式** | **原因说明** |
| 用户行为日志表 | 增量（按分区/时间戳） | 数据量大，更新频率高，适合按天增量 |
| 交易明细表 | 增量 + 分区覆盖 | 需处理退单、取消等历史变更，需每天处理当天数据并可能补前几天 |
| 用户维表（如 dim\_user） | 增量（SCD 类型）或全量 | 如果用 SCD2 做历史快照，则必须处理更新记录 |
| 配置类维表（如 dim\_province） | 全量每日覆盖 | 数据少，几乎不变，直接全量刷新即可 |
| 报表（ADS） | 增量 + insert overwrite 分区 | 每天计算当天的汇总指标，按天覆盖即可 |

示例 1：订单明细表



判断：

* 有 update\_time：增量
* 订单存在取消、支付状态变化：每天处理增量 + 近期分区回刷

示例 2：省份维度表 dim\_province

* 每天只几条数据，不变动
* 无时间戳字段

判断：

* 无需增量，直接 全量覆盖

数据可视化

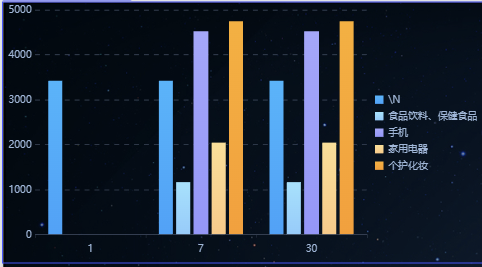
将 ADS 层（应用数据层）或 DWS 层的汇总宽表中的指标，以可视化的形式呈现出来，辅助业务分析和决策。

设计原则：

| 项目 | 说明 |
| --- | --- |
| 主题驱动 | 按业务主题设计（如交易分析、流量分析、用户行为等） |
| 指标清晰 | 每张报表展示的指标有明确业务含义和口径 |
| 来源统一 | 报表指标全部来源于 ADS（或 DWS） 层的字段 |
| 时间维度支持 | 支持按日、周、月、最近 N 日等粒度切换 |
| 支持多维分析 | 渠道、品牌、省份、品类等维度可筛选 |
|  |  |

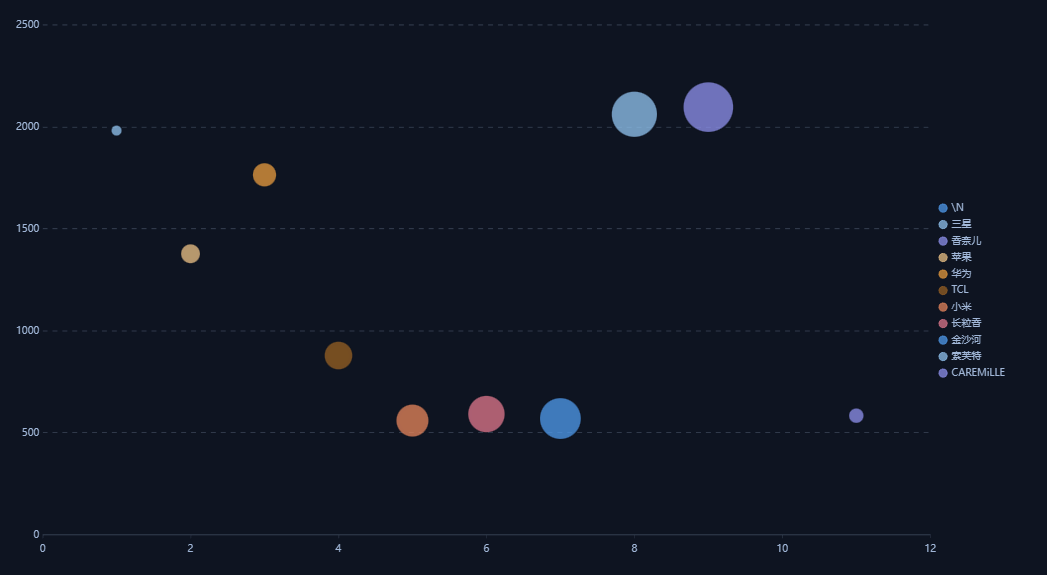
品类下单用户漏斗分析

|  |  |  |  |
| --- | --- | --- | --- |
| **指标** | **说明** | 字段 | **来源表** |
| 一级品类ID | 商品一级分类 | category1\_id | ads\_order\_stats\_by\_cate |
| 一级品类名称 | 分类名称 | category1\_name | ads\_order\_stats\_by\_cate |
| 二级品类ID | 商品二级分类 | category2\_id | ads\_order\_stats\_by\_cate |
| 下单人数（近1天） | 最近1天下单 | order\_user\_coun（recent\_days=1） | ads\_order\_stats\_by\_cate |
| 下单人数（近7天） | 最近7天下单 | order\_user\_coun（recent\_days=7） | ads\_order\_stats\_by\_cate |
| 下单人数（近30天） | 最近30天下单 | order\_user\_coun（recent\_days=30） | ads\_order\_stats\_by\_cate |



品牌 GMV 分析

|  |  |  |  |
| --- | --- | --- | --- |
| **报表指标** | **说明** | **数据来源表** | **字段** |
| 下单数 | 下单笔数 | ads\_order\_stats\_by\_tm | order\_count |
| 下单人数 | 下单用户数 | ads\_order\_stats\_by\_tm | order\_user\_count |
| 品牌名 | 品牌维度 | ads\_order\_stats\_by\_tm | tm\_name |



流量渠道统计

|  |  |  |  |
| --- | --- | --- | --- |
| **报表指标** | **说明** | **数据来源表** | **字段** |
| 渠道名 | 来源渠道，如APP、小程序 | ads\_traffic\_stats\_by\_channel | channel |
| UV | 独立访客数 | ads\_traffic\_stats\_by\_channel | uv\_count |
| 会话数 | 会话次数 | ads\_traffic\_stats\_by\_channel | sv\_count |
| 最近几天 | “最近几天”的时间范围 | ads\_traffic\_stats\_by\_channel | recent\_days |
| 跳出率 | 只看一个页面就离开的比例 | ads\_traffic\_stats\_by\_channel | bounce\_rate |



用户新增活跃趋势

|  |  |  |  |
| --- | --- | --- | --- |
| **报表指标** | **说明** | **数据来源表** | **字段** |
| 新增用户数 | 某日首次登录用户 | ads\_user\_stats | new\_user\_count |
| 活跃用户数 | 最近 N 天登录过的用户 | ads\_user\_stats | active\_user\_count |
| 最近天数 | 维度字段 | ads\_user\_stats | recent\_days |



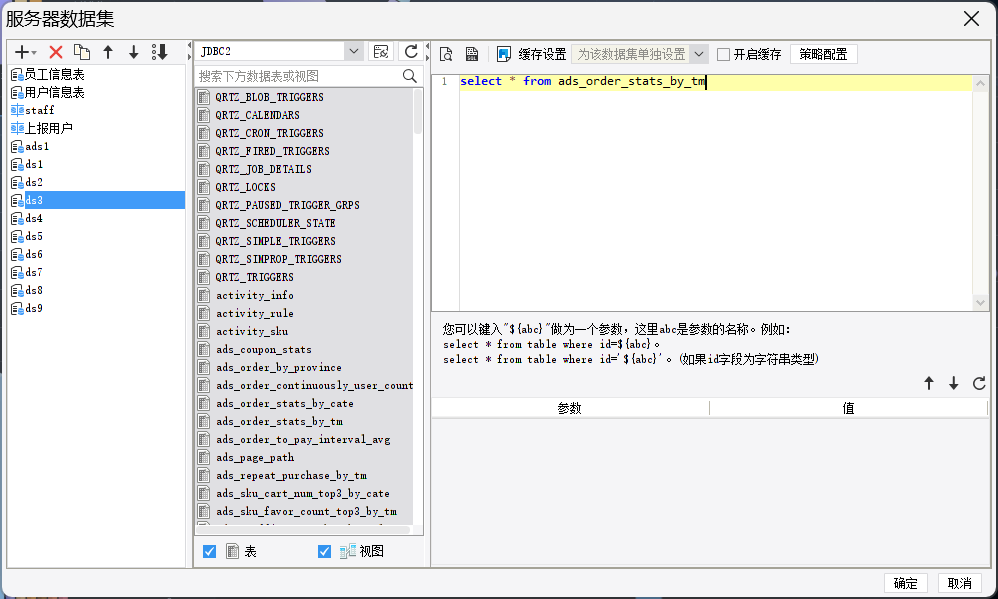
省份交易地图

|  |  |  |  |
| --- | --- | --- | --- |
| **报表指标** | **说明** | **数据来源表** | **字段** |
| 省份名称 | 地图维度 | ads\_order\_by\_province | province\_name |
| 订单金额 | 各省总交易额 | ads\_order\_by\_province | order\_total\_amount |
| 订单数 | 各省下单笔数 | ads\_order\_by\_province | order\_count |



FineReport 设计流程示例

1. 数据对接：将 Hive 或 MySQL 中的 ADS 层表连接到 FineReport 数据集（通过 JDBC 数据源）。
2. 建立数据集：用 SQL 查询汇总后的结果，如：



1. 设计报表模板：
   * 折线图 → 用户新增活跃趋势
   * 柱状图 → 品类下单用户漏斗分析
   * 地图 → 省份交易金额
   * 散点图 → 品牌 GMV 排行
   * 三维图 → 流量渠道统计
2. 参数交互：支持 日期下拉、品牌筛选、渠道维度切换 等参数联动。
3. 定时调度/发布大屏：支持每日报表推送或大屏展示。

