# 淘宝用户行为数仓系统的设计与实现

|  |  |
| --- | --- |
| **小组成员** | **分工** |
| 刘灵菲  37620222204762 | 完成项目前期的背景调查、需求分析和概念设计 |
| 陆浩楠  34520232201088 | 完成项目逻辑设计、物理设计 |
| 苏一涵（组长）36720232204041 | 参与开发全程，组织调配整体项目任务，完成系统的搭建、数据处理以及具体实现，ppt制作 |
| 陈昊睿  37520232205079 | 完成多维分析与可视化部分 |

## 项目背景

随着数字经济的快速发展，电商行业已成为零售领域的核心支柱，用户行为数据成为平台优化运营、提升转化效率的核心资产。淘宝作为国内领先的电商平台，积累了海量用户浏览、收藏、加购、购买等行为数据，这些数据蕴含着用户消费偏好、商品热度分布、转化路径规律等关键业务信息，对精准营销、商品运营、用户留存具有重要指导意义。

本项目以阿里天池公开的《淘宝用户行为数据集》为基础数据源，该数据集包含超过1000万条真实的用户隐去敏感信息后的行为记录，覆盖了淘宝平台上一段时间内的用户与商品交互全貌。我们希望通过构建本数据仓库及后续分析，能够系统性地揭示用户从访问到决策的全链路行为规律，量化核心业务指标，从而为精准营销、活动策划、库存优化、产品体验提升等关键运营决策提供坚实的数据支撑。

## 需求分析

电商平台每天产生海量用户行为日志，数据价值巨大，但原始的行为日志数据庞杂、分散且无法直接用于分析，业务部门面临多重挑战。原始日志以文本形式分散存储，缺乏统一标准，无法支持高效的多维查询与分析，形成了“数据孤岛”。运营、产品和市场团队无法快速获取统一的、可信的业务指标，决策多依赖经验而非数据。每当有新的分析需求，数据工程师或分析师都需要从原始日志开始重新编写复杂的处理脚本，分析效率低下。

为解决上述问题，本项目的核心目标构建企业级数据仓库，实现从“原始日志”到“商业洞察”的转化。为了更好地存储、处理和分析淘宝用户行为数据，并最大化其业务与教学价值，我们明确了以下核心需求：

1. 数据提取与集成需求

能够完整、准确地导入给定的静态历史数据集至数据仓库系统。设计的数据模型应具备良好的扩展性，以支持未来可能增加的实时流数据或新的行为类型。在数据加载过程中，需内置对原始数据的清洗（如过滤无效字段、纠正时区）和转换逻辑，使其适应数据仓库的规范。

2. 数据模型设计需求

采用经典的维度建模方法，设计清晰的事实表与维度表结构，以支持高效、灵活的多维分析。核心模型需包括：用户行为事实表，并围绕其构建时间、用户、商品、商品类目、行为类型等多个维度表。模型设计需考虑业务分析的自然粒度，并支持向上汇总至日、小时、用户、商品类目等多个维度组合。

3. 数据质量保障需求

在ETL处理过程中，实施数据质量检查，例如确保关键业务键非空，过滤行为类型不在预设范围内的脏数据。通过数据验证查询来监测数据加载的完整性。建立一致性的数据处理口径，确保从ODS原始层到ADS应用层的指标计算逻辑统一、可信。

4. 查询性能与可扩展性需求

利用Hive分区技术（按日期dt分区）对大数据量表进行物理划分，显著提升按时间范围查询的效率。采用列式存储格式并配置压缩，以减少I/O开销，提升查询速度。数据模型与ETL流程设计应模块化，确保当分析需求变更或数据量增长时，系统易于调整和扩展。

5. 分析功能与报表输出需求

数据仓库的最终输出（ADS层）必须直接支持四大核心业务分析场景：流量分析、转化漏斗分析、用户消费偏好分析和用户价值分析。预先计算并固化关键业务指标表，例如：每日用户行为转化漏斗表、热门商品/类目TOP10排行榜、日活跃用户及复购率统计表。输出的数据表结构应清晰、直观，便于直接连接可视化工具进行图形化报表展示。

6. 项目文档与可复用性需求

提供完整、详细的技术与业务文档，清晰阐述数据分层（ODS, DWD, DWS, ADS）定义、ETL逻辑、指标计算口径，确保项目可被理解、维护与重现。代码与脚本应具备良好的注释和结构，作为数据仓库教学的典型案例，供学习者参考与实践。

7. 安全与隐私考量

本项目使用的为已公开的脱敏数据集，不涉及真实的用户隐私信息。但在企业实际环境中，需强调数据权限管控，确保敏感数据仅对授权用户开放。

以上需求分析旨在确保本项目构建的数据仓库不仅能高效、可靠地承载数据，更能紧密贴合业务分析目标，产出具有直接指导价值的洞察结果，同时作为一项规范的技术资产，具备教育意义和参考价值。

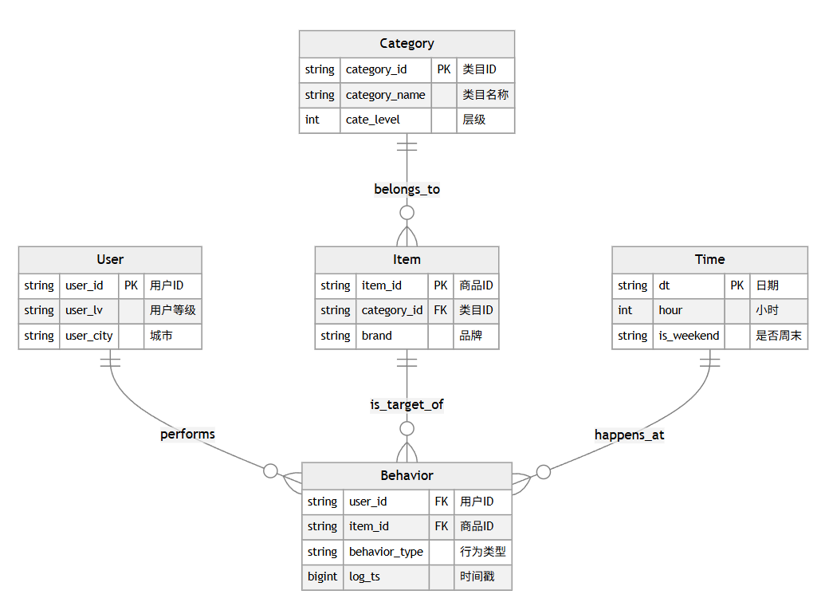
## 概念设计

核心实体：用户，商品，商品类目，时间以及行为。

实体之间的关系如下：

1. 用户 - 行为: 一对多 (1:N)。一个用户可以产生多条行为记录。
2. 商品 - 行为: 一对多 (1:N)。一个商品可以被多次操作。
3. 商品 - 类目: 多对一 (N:1)。一个商品属于一个类目（简化模型）。
4. 时间 - 行为: 一对多 (1:N)。一个时间点（如某天某小时）发生多个行为。

实体关系ER图如下所示：



## 4.逻辑设计

4.1 设计理念与核心模型

本项目逻辑设计遵循维度建模思想，以 “用户行为” 为核心事实，围绕 “人、货、场” 业务逻辑构建星型模型，确保数据支持高效多维分析。核心设计理念包括：

聚焦业务核心：以用户浏览、加购、收藏、购买等行为为事实核心，覆盖电商核心转化链路；

维度全面支撑：构建时间、用户、商品、商品类目、行为类型五大维度，满足多视角分析需求；

粒度精准可控：事实表粒度细化至 “单条用户行为”，支持从明细到汇总的全层级分析；

指标统一口径：在逻辑层定义核心指标计算规则，确保后续数据处理与分析的一致性。

4.2 星型模型结构

星型模型以事实表为中心，维度表围绕事实表展开关联，具体结构如下：

4.2.1 核心事实表

|  |  |  |  |
| --- | --- | --- | --- |
| 事实表名称 | 核心作用 | 关联维度 | 关键度量值 |
| user\_behavior\_fact（用户行为明细事实表） | 存储清洗后的单条用户行为记录，是整个数仓的核心事实载体 | 时间、用户、商品、商品类目、行为类型 | 行为次数（基础度量，后续汇总衍生其他指标） |

4.2.2 维度表设计

各维度表聚焦单一业务视角，包含基础属性与扩展字段，支持灵活筛选与聚合分析：

|  |  |  |  |
| --- | --- | --- | --- |
| 维度表名称 | 维度核心 | 关键字段 | 支持分析场景 |
| dim\_time（时间维度表） | 时间粒度拆分 | 日期、小时、年、季、月、周、是否周末、是否节假日 | 分时 / 分日 / 分周流量分析、高峰期识别 |
| dim\_user（用户维度表） | 用户基础属性 | 用户 ID、用户等级、注册日期、所在城市 / 省份 | 用户分层、地域分布、高价值用户识别 |
| dim\_item（商品维度表） | 商品基础属性 | 商品 ID、类目 ID、商品名称、品牌、价格 | 商品热度、品牌分析、价格带转化 |
| dim\_category（商品类目维度表） | 类目层级关系 | 类目 ID、类目名称、父类目 ID、类目层级 | 类目热度排名、层级转化分析 |
| dim\_behavior\_type（行为类型维度表） | 行为定义与权重 | 行为类型编码、行为名称、行为权重 | 行为价值量化、综合活跃度计算 |

4.3 表间关联关系

事实表与维度表通过 “外键 - 主键” 关联，所有关联字段均为字符串类型，确保关联稳定性：

user\_behavior\_fact.user\_id → dim\_user.user\_id（用户维度关联）

user\_behavior\_fact.item\_id → dim\_item.item\_id（商品维度关联）

user\_behavior\_fact.category\_id → dim\_category.category\_id（类目维度关联）

user\_behavior\_fact.dt\_hh → dim\_time.dt\_hh（时间维度关联）

user\_behavior\_fact.behavior\_type → dim\_behavior\_type.behavior\_type（行为类型关联）

维度表之间独立设计，避免维度表直接关联，降低模型复杂度，提升查询效率（如商品与类目通过事实表间接关联，而非直接建立表间关系）。

4.4 数据粒度与聚合逻辑

基础粒度：事实表粒度为 “单条用户行为”（1 条记录对应 1 次用户-商品交互）；

聚合粒度：通过后续DWS层汇总，支持多维度组合聚合（如 “日期+类目”“日期+用户”“小时+商品” 等），衍生出浏览量、加购量、购买量等核心指标。

## 5.物理设计

5.1 设计原则

物理设计基于Hive数据仓库特性，兼顾存储效率、查询性能、可扩展性三大核心目标，具体原则：

分区策略：按日期（dt）分区存储大表，降低查询时的数据扫描范围；

存储格式：采用ORC列式存储，配合SNAPPY压缩，减少I/O开销与存储占用；

数据类型：核心字段（ID类）采用STRING类型避免溢出，数值字段（计数、价格）按需定义精度（如DECIMAL (10,2) 存储价格）；

索引优化：对高频筛选字段（如user\_id、item\_id）设置ORC bloom filter索引，提升过滤效率；

路径规范：按 “数据库-表类型-表名” 统一规划HDFS存储路径，便于管理与维护。

5.2 分层物理表实现

基于数仓标准四层架构（ODS→DWD→DWS→ADS），逐层设计物理表结构，每一层均有明确的存储与处理目标：

5.2.1 ODS 层（原始数据层）

核心目标：原样存储原始数据，不做清洗，保留数据完整性；

表结构设计：完全贴合源数据格式，字段与源数据一致，仅规范字段名称与注释；

关键表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名 | 存储内容 | 存储格式 | 分区策略 | 关键配置 |
| ods.taobao\_user\_behavior\_ods | 淘宝用户行为原始日志（用户 ID、商品 ID、类目 ID、行为类型、时间戳） | TEXTFILE | 无分区（原始数据按文件存储） | 字段分隔符为逗号，跳过表头 |
| ods.taobao\_category\_ods | 淘宝商品类目原始数据（类目 ID、父类目 ID、类目名称等） | TEXTFILE | 无分区 | 字段分隔符为逗号，跳过表头 |

5.2.2 DIM 层（维度表层）

核心目标：存储结构化维度数据，支持事实表关联与筛选；

表结构设计：包含主键、基础属性字段、扩展字段，确保维度完整性；

关键表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名 | 主键 | 存储格式 | 关键配置 | 存储路径 |
| dim.dim\_time | dt\_hh（日期+小时） | ORC（SNAPPY压缩） | 包含年/季/月/日/小时等多粒度字段 | /user/hive/warehouse/taobao\_dw.db/dim/dim\_time |
| dim.dim\_user | user\_id | ORC（SNAPPY压缩） | 扩展用户等级、地域等字段 | /user/hive/warehouse/taobao\_dw.db/dim/dim\_user |
| dim.dim\_item | item\_id | ORC（SNAPPY压缩） | 包含商品品牌、价格等扩展字段 | /user/hive/warehouse/taobao\_dw.db/dim/dim\_item |
| dim.dim\_category | category\_id | ORC（SNAPPY压缩） | 支持类目层级（父类目ID、类目层级） | /user/hive/warehouse/taobao\_dw.db/dim/dim\_category |
| dim.dim\_behavior\_type | behavior\_type | ORC（SNAPPY压缩） | 包含行为权重字段，支持价值量化 | /user/hive/warehouse/taobao\_dw.db/dim/dim\_behavior\_type |

5.2.3 DWD 层（明细数据层）

核心目标：清洗原始数据，关联维度外键，生成标准化明细事实数据；

表结构设计：基于 ODS 层数据清洗后，添加维度外键，按日期分区；

关键表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表名 | 核心字段 | 存储格式 | 分区策略 | 清洗规则 | 存储路径 |
| dwd.taobao\_user\_behavior\_fact | user\_id、  item\_id、  category\_id、  behavior\_typ  e、dt\_hh、  log\_ts | ORC（SNAPPY压缩） | 按日期分区  （dt，格式：yyyy-MM-dd） | 1. 过滤无效行为类型； 2. 剔除用户ID/商品ID为空的脏数据；   3. 时间戳转换为东八区日期+小时 | /user/hive/warehouse/dwd.db/taobao\_user\_behavior\_fact |

5.2.4 DWS 层（数据汇总层）

核心目标：按指定维度聚合明细数据，生成中间汇总指标，提升 ADS 层查询效率；

表结构设计：按 “聚合维度 + 度量值” 设计，保留核心指标，按日期分区；

关键表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表名 | 聚合维度 | 核心度量值 | 存储格式 | 分区策略 | 存储路径 |
| dws.taobao\_behavior\_cate\_agg | 日期（dt）+类目ID（category\_id） | 浏览量（pv\_cnt）、  加购量（cart\_cnt）、  购买量（buy\_cnt）、  收藏量（fav\_cnt） | ORC（SNAPPY压缩） | 按日期（dt）分区 | /user/hive/warehouse/dws.db/taobao\_behavior\_cate\_agg |
| dws.taobao\_behavior\_user\_agg | 日期（dt）+用户ID（user\_id） | 总行为数（total\_behavior\_cnt）、  购买次数（buy\_cnt）、  主要活跃小时（active\_hour） | ORC（SNAPPY压缩） | 按日期（dt）分区 | /user/hive/warehouse/dws.db/taobao\_behavior\_user\_agg |

5.2.5 ADS 层（数据应用层）

核心目标：基于 DWS 层汇总数据，计算最终业务指标，直接支撑可视化与运营决策；

表结构设计：按业务场景设计，包含最终 KPI 指标，字段直观易懂；

关键表（对应核心业务场景）：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表名 | 业务场景 | 核心字段 | 存储格式 | 存储路径 |
| ads.taobao\_conversion\_funnel | 行为转化漏斗分析 | 日期、  浏览用户数、  加购用户数、  收藏用户数、  购买用户数、  各环节转化率 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_conversion\_funnel |
| ads.taobao\_hot\_item\_top10 | 热门商品 TOP10分析 | 日期、  商品排名、  商品ID、  浏览量、  购买量、  购买转化率 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_hot\_item\_top10 |
| ads.taobao\_user\_dau | 日活跃用户分析 | 日期、  日活跃用户数（DAU）、  7日累计活跃用户数（MAU\_7d）、  日活跃率 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_user\_dau |
| ads.taobao\_traffic\_hourly\_stats | 分时流量趋势分析 | 日期、  小时、  浏览次数、  加购次数、  收藏次数、  购买次数 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_traffic\_hourly\_stats |
| ads.taobao\_category\_top10 | 热门类目 TOP10 分析 | 日期、  类目排名、  类目ID、  浏览量、  购买量、  购买转化率 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_category\_top10 |
| ads.taobao\_repurchase\_stats | 用户复购率分析 | 日期、  购买用户数、  复购用户数、  复购率 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_repurchase\_stats |
| ads.taobao\_peak\_hour\_distribution | 活跃高峰时段分析 | 日期、  活跃小时、  达峰用户数、  用户占比、  平均行为次数 | ORC（SNAPPY压缩） | /user/hive/warehouse/taobao\_dw.db/ads/taobao\_peak\_hour\_distribution |

5.3 关键技术配置

动态分区配置：开启非严格模式（SET hive.exec.dynamic.partition.mode=nonstrict），支持动态生成日期分区；

并行度优化：设置 Hive/Spark 执行参数（如 spark.executor.instances=60），提升大数据量处理效率；

数据质量保障：通过 TBLPROPERTIES 设置跳过表头、ORC 索引等配置，确保数据加载准确性；

时区处理：原始时间戳统一转换为东八区时间（+28800 秒），避免时区偏差影响分析结果。

## 6.系统实现

### 6.1开发工具和部署平台

平台：

Hadoop (HDFS + YARN)集群：分布式文件系统，存储天池数据集等的 CSV 文件

共使用3台虚拟机作为服务端，用windows本机做为客户端

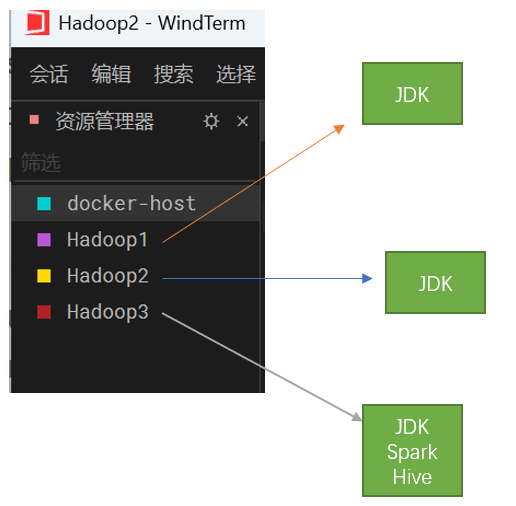
集群配置如下



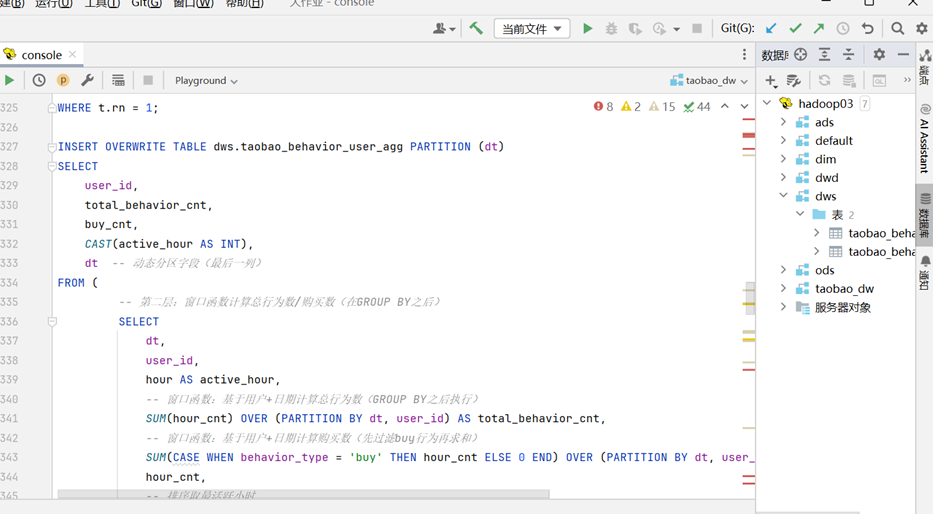
数仓工具是**Hive**：作为数据仓库的元数据存储和查询引擎,通过HQL将结构化查询转化为MapReduce任务，实现高层次的SQL-like查询。

搭建了**Hive-on-Spark环境**，将 HQL 查询转换为 Spark 作业执行

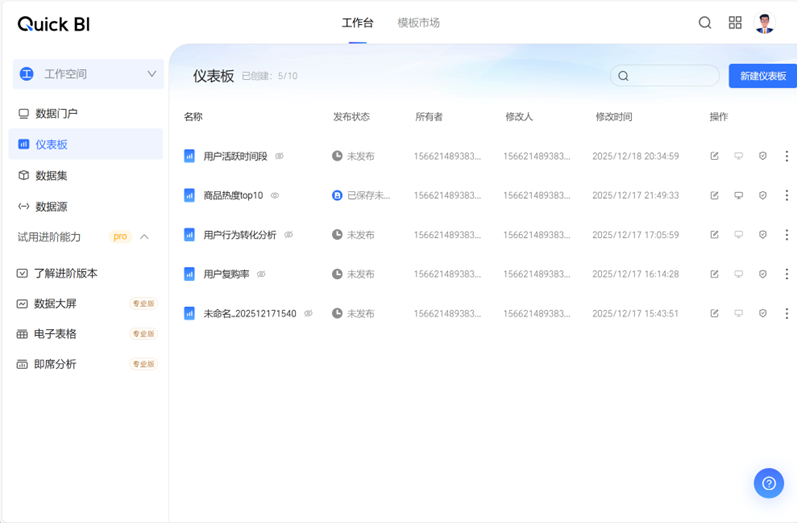
各个机器的开发环境配置如下，通过**WindTerm**统一管理所有虚拟机



部署完成后通过**IDEA**连接hive数据库进行开发



可视化工具：**QuickBI**



绘制事实表和维度表工具：starUML

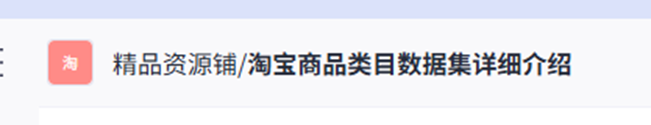


### 6.2 数据来源

核心数据集：[淘宝用户购物行为数据集\_数据集-阿里云天池](https://tianchi.aliyun.com/dataset/649)

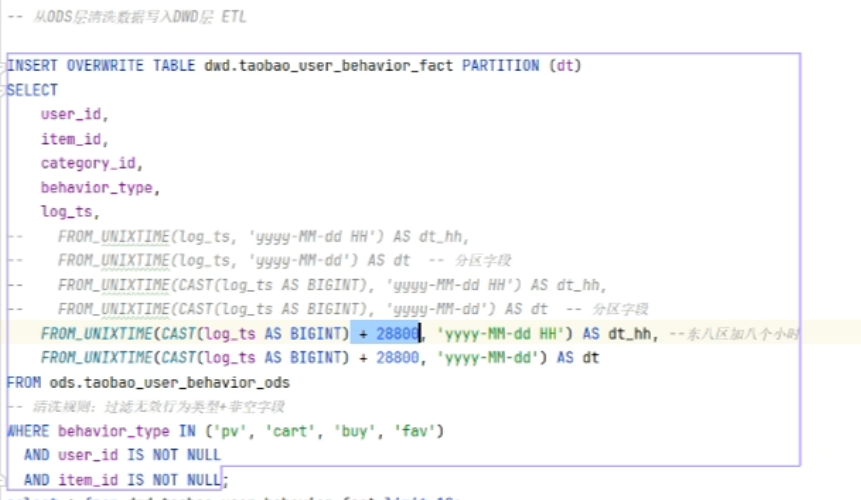


商品数据集：[淘宝商品类目数据集](https://gitcode.com/Premium-Resources/9d6e4/?utm_source=article_gitcode_universal&index=top&type=card&&uuid_tt_dd=10_6202980440-1735729714820-253335&isLogin=9&from_id=147912117&from_link=33613e86681689e176468b58acbfab1c)



### 6.3 ETL

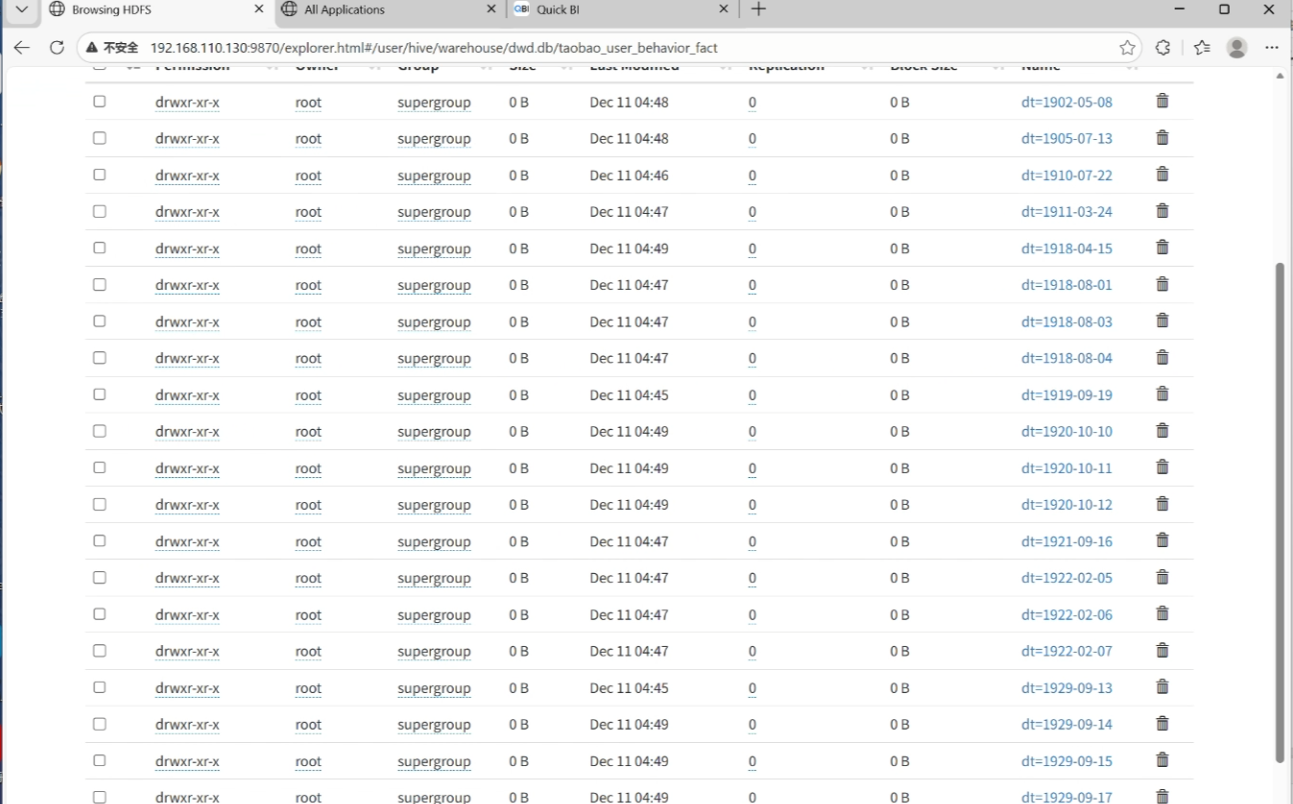
ods层到dwd层的加载逻辑

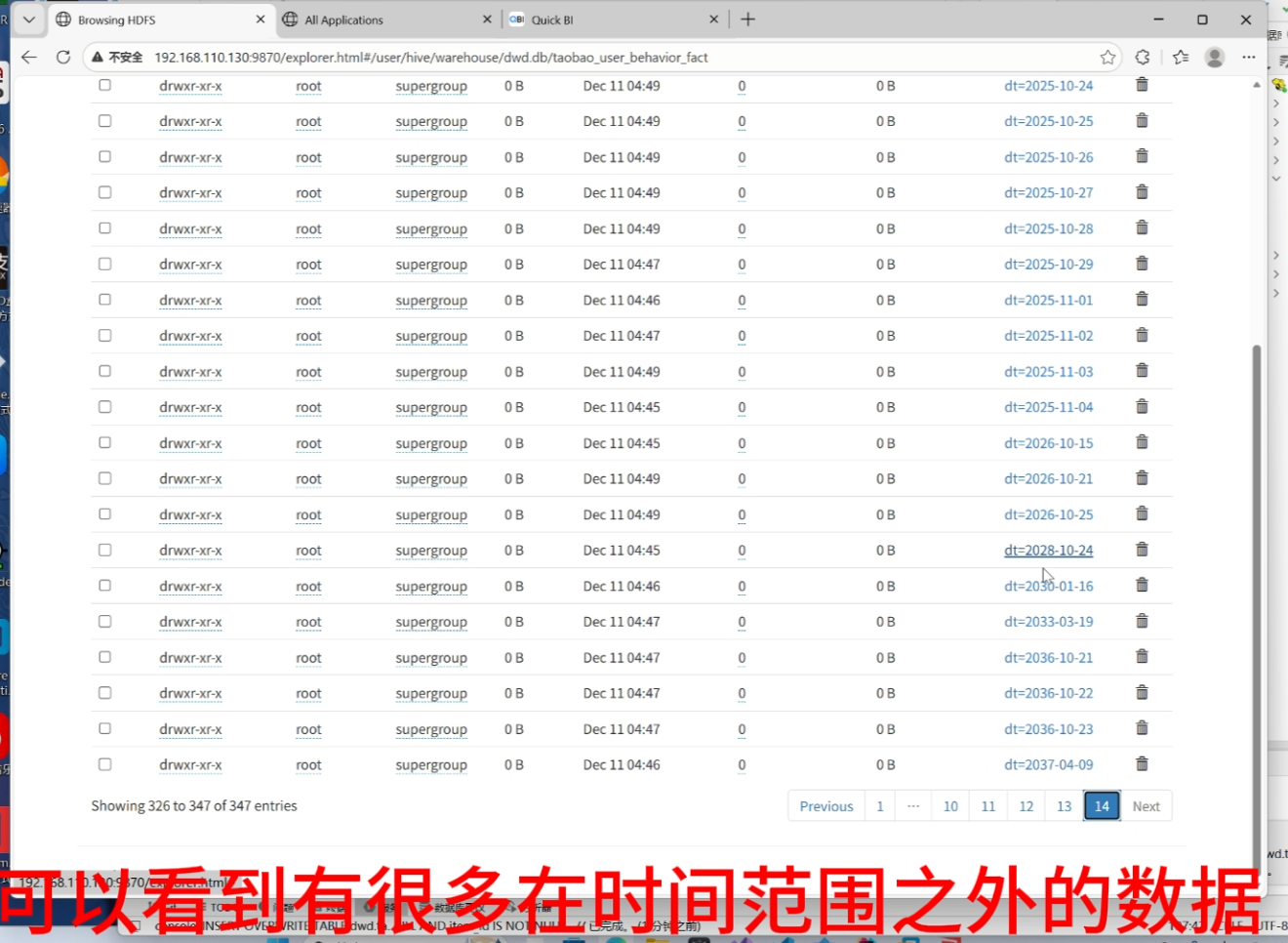


一个是时间，从时间戳转化为utc时间，然后因为北京时间是utc+8，所以时间戳要加28800

然后是过滤掉behavior等字段是空值的数据。

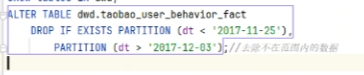
加载之后情况如下



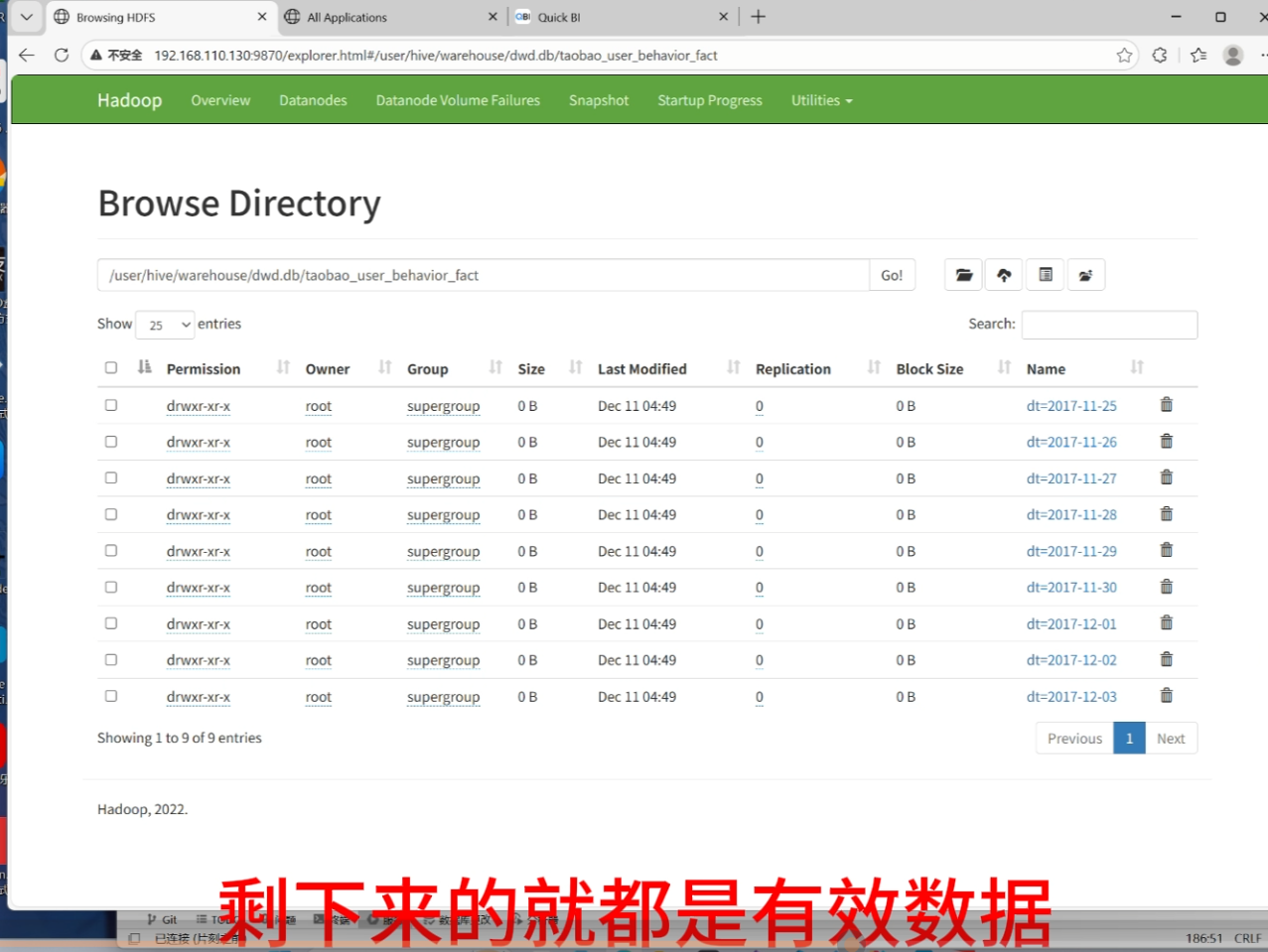


有很多在时间范围之外的数据

这边再给他清洗掉

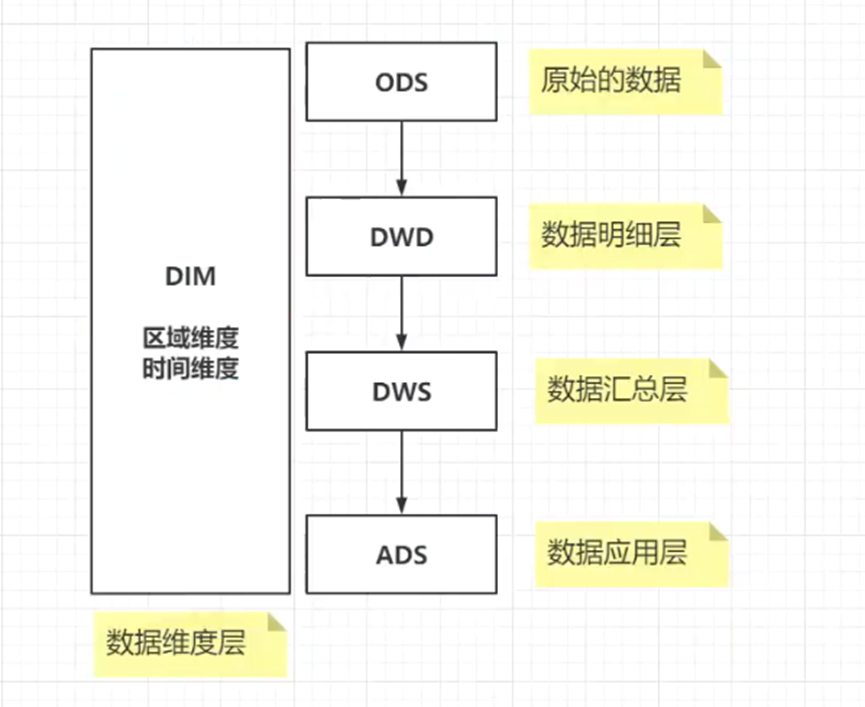


剩下来的就都是有效数据了

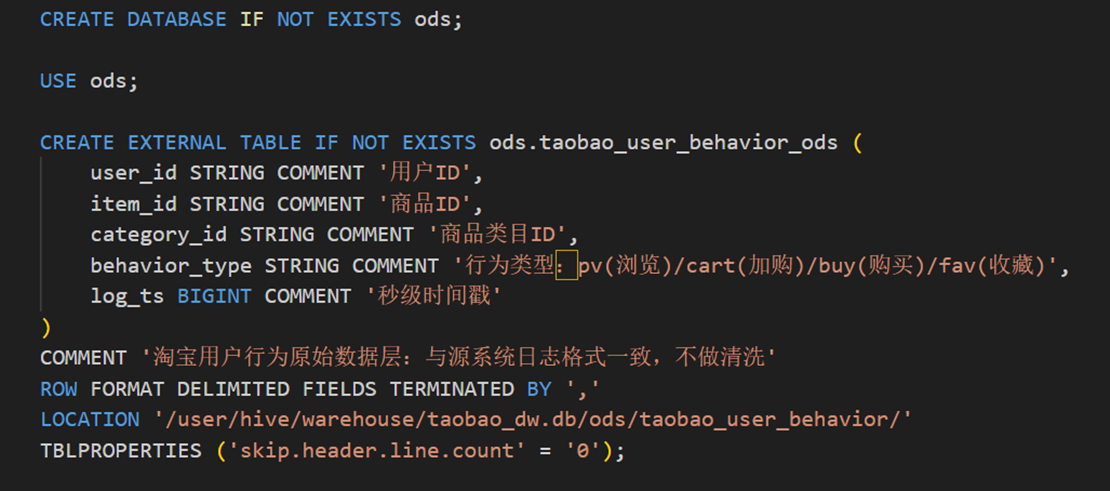


### 6.4 数据分层

按如下分层



ods层：ods.taobao\_user\_behavior\_ods



dim层：dim.dim\_time



dim.dim\_user



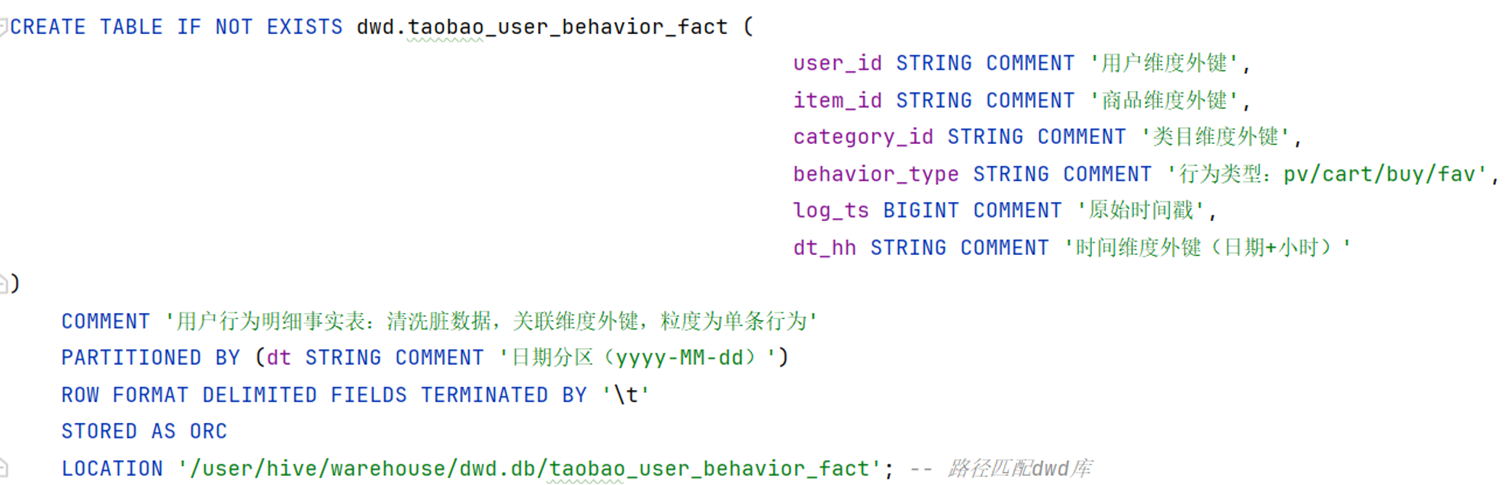
dim.dim\_item



dim.dim\_category



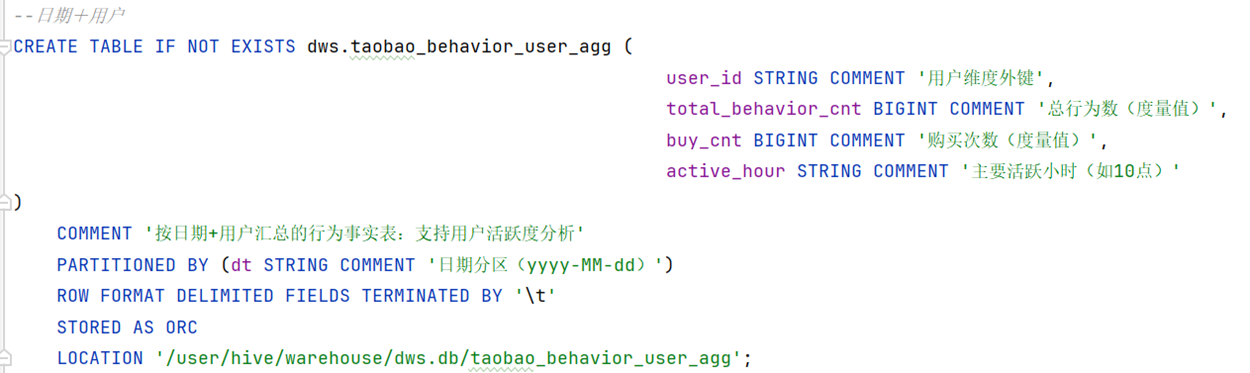
dwd层：dwd.taobao\_user\_behavior\_fact



dws层：dws.taobao\_behavior\_cate\_agg



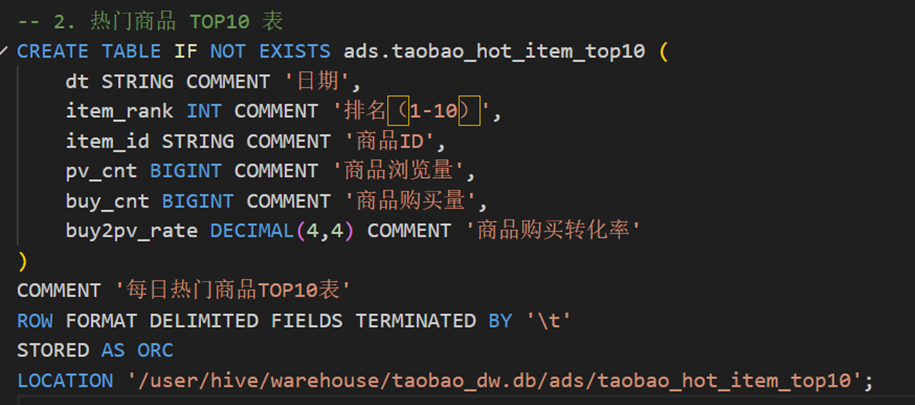
dws.taobao\_behavior\_user\_agg



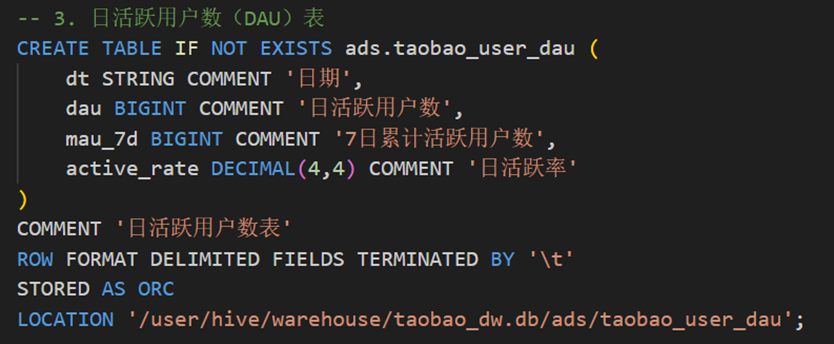
ads层：ads.taobao\_conversion\_funnel



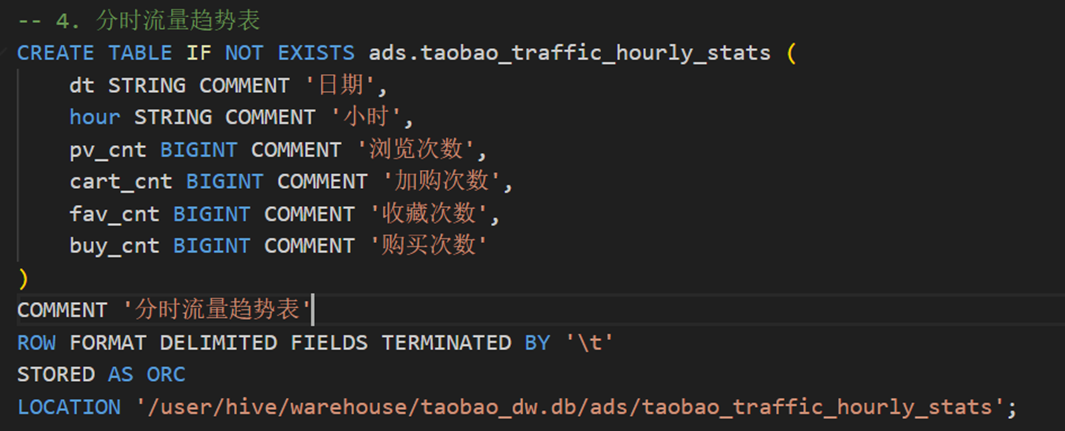
ads.taobao\_hot\_item\_top10



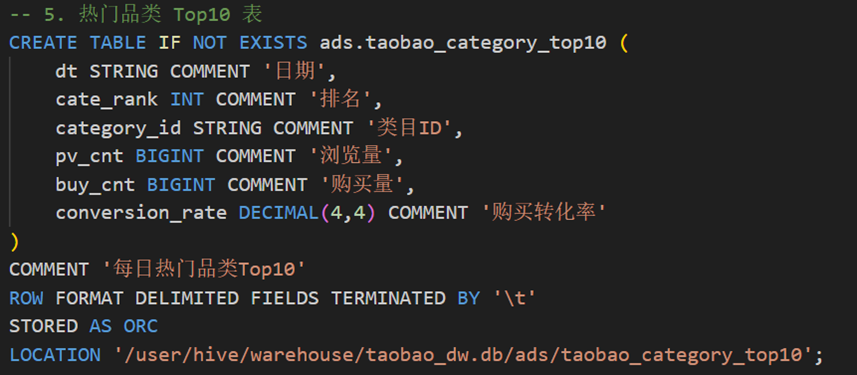
ads.taobao\_user\_dau



ads.taobao\_traffic\_hourly\_stats



ads.taobao\_category\_top10



ads.taobao\_repurchase\_stats



### 6.5 KPI

1. 流量与活跃度指标 (Traffic & Engagement)

设计目的：监控平台流量规模与用户活跃规律，指导营销活动投放时间。

| **指标名称** | **定义/计算逻辑** | **业务价值** |
| --- | --- | --- |
| **日活跃用户数** | 每日有任意行为（PV/Cart/Fav/Buy）的去重用户数。 | 衡量平台核心流量规模，判断大促活动的引流效果。 |
| **分时活跃度** | 统计全天 24 小时每小时的 PV/UV 量。 | 发现用户活跃高峰期，指导广告推送和服务器扩容。 |
| **日活跃用户数** | 最近 30 天内有活跃行为的去重用户数。 | 衡量平台的长期用户粘性。 |
| **活跃率** | DAU / MAU（或总注册用户）。 | 评估用户的活跃程度，活跃率越高说明用户粘性越强。 |

2. 转化与漏斗指标 (Conversion & Funnel)

设计目的：分析用户从浏览到购买的全链路转化效率，定位流失环节。

| **指标名称** | **定义/计算逻辑** | **业务价值** |
| --- | --- | --- |
| **转化漏斗** | 浏览(PV) → 收藏(Fav) → 加购(Cart) → 购买(Buy) 各环节的用户数留存。 | 宏观展示用户流失情况，识别转化瓶颈（如加购后未支付）。 |
| **浏览-购买转化率** | 购买用户数 / 浏览用户数。 | 衡量整体流量变现能力。 |
| **加购-购买转化率** | 购买用户数 / 加购用户数。 | 衡量购买意向较强用户的最终成交效率，低则可能涉及价格或支付体验问题。 |
| **复购率** | 购买次数 > 1 的用户数 / 总购买用户数。 | 衡量用户忠诚度和商品质量，复购率高意味着用户生命周期价值 (LTV) 高。 |

3. 商品与类目指标 (Product & Category)

设计目的：挖掘爆款商品与潜力品类，优化推荐策略与库存管理。

| **指标名称** | **定义/计算逻辑** | **业务价值** |
| --- | --- | --- |
| **热门商品 Top10** | 按浏览量 (PV) 或 购买量 (Buy) 倒序排名的前 10 个商品。 | 识别“流量担当”和“营收担当”，用于首页推荐位选品。 |
| **热门品类 Top10** | 按浏览量或销量排名的前 10 个类目（关联类目名称）。 | 了解用户当下的消费热点（如换季时的服装、大促时的数码）。 |
| **商品转化率** | 单个商品的 购买次数 / 浏览次数。 | **诊断异常商品**：<br>1. **高曝低转**：有流量没销量（需优化详情页或价格）。<br>2. **低曝高转**：有销量没流量（需增加曝光权重）。 |

4. 维度设计 (Dimensions)

为了支持多角度分析，KPI 设计结合了以下维度：

时间维度：年、月、日、小时、是否周末、是否节假日。

商品维度：商品 ID、类目 ID、类目名称、父类目。

用户维度：用户 ID（可扩展用户等级、地域等）。

## 7.多维分析与可视化

**流量维度：**

**日活跃用户 (DAU)**



数据现象：

平稳期：11月25日至11月30日期间，日活跃用户数（DAU）保持在 70万-73万 之间，活跃率稳定在 71%-73%。

爆发期：进入12月后流量显著攀升。12月2日（周六）达到峰值 94.1万，12月3日（周日）维持在 91.7万。活跃率飙升至 95% 以上。

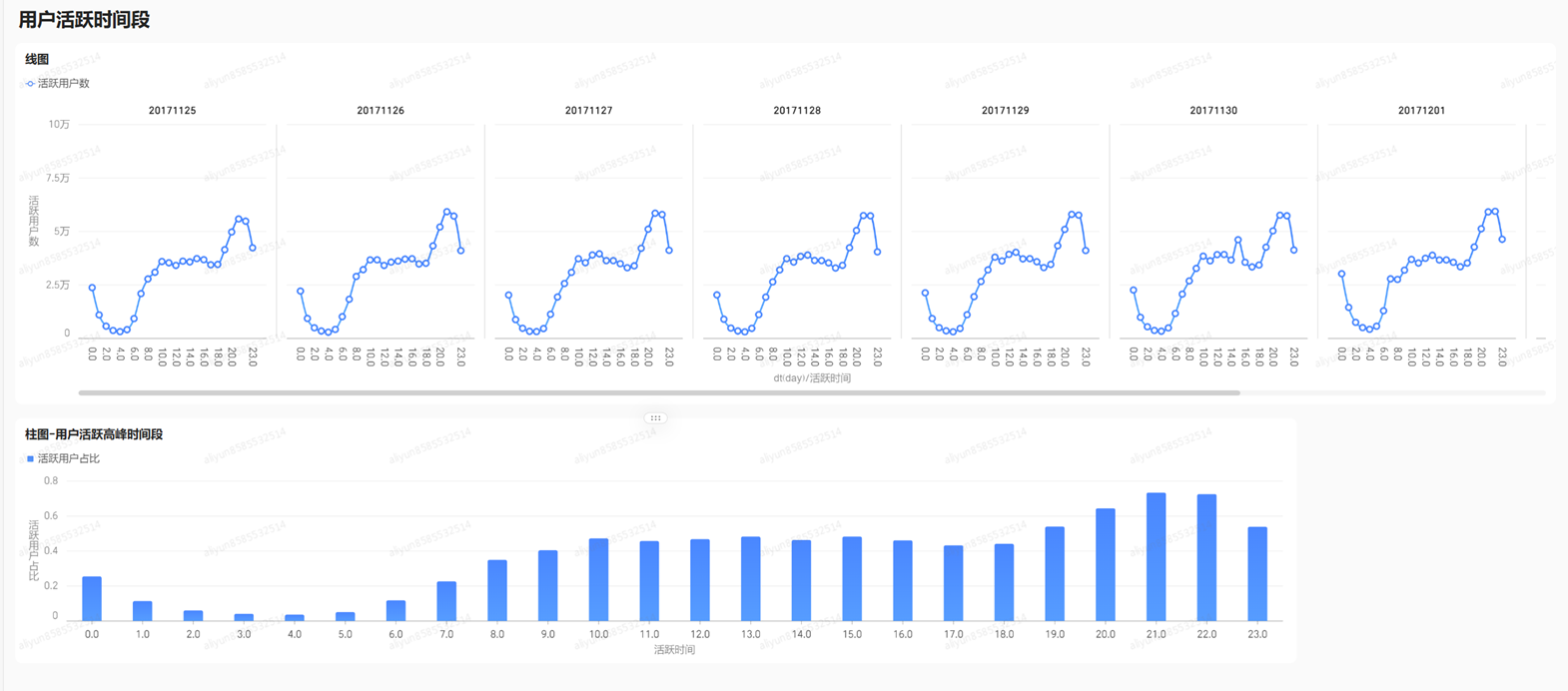
业务洞察：

周末效应显著：12月2日-3日为周末，用户有更多闲暇时间逛淘宝，导致流量激增近 30%。

大促预热：这可能也与“双12”大促前的预热活动有关，用户开始提前浏览商品。

建议：在周末或大促前夕，需重点保障服务器稳定性以应对流量洪峰；同时可在周五晚间推送营销信息，承接周末流量。

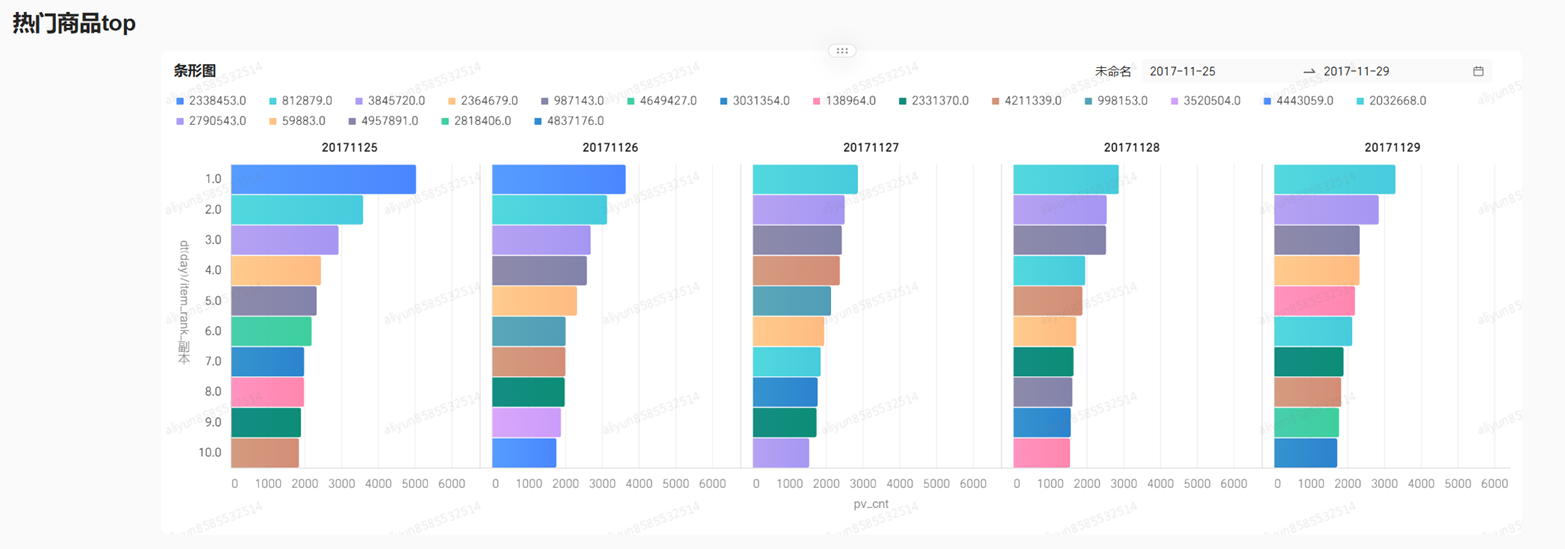
**分时流量趋势**

****

晚间 20:00-22:00 是用户活跃高峰，建议在此时间段推送营销活动；凌晨 2:00-5:00 为低谷期，适合进行系统维护。

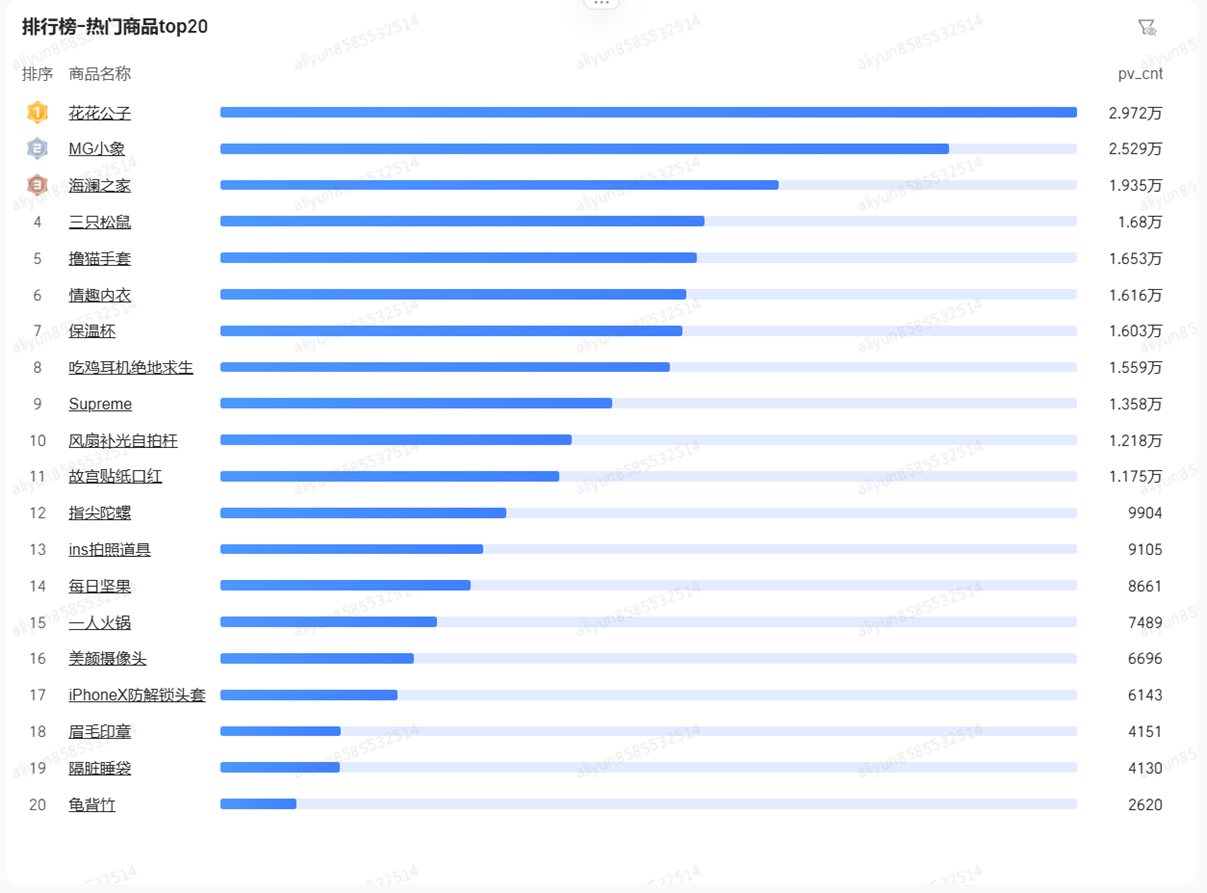
**商品维度 ：**

**热门商品 Top10**

****

可以根据商品id查看当天最热门的商品top10

**商品排名总榜top20**

****

前三都是衣服相关的，可以从这个表看出当时的热门商品以及潮流风向

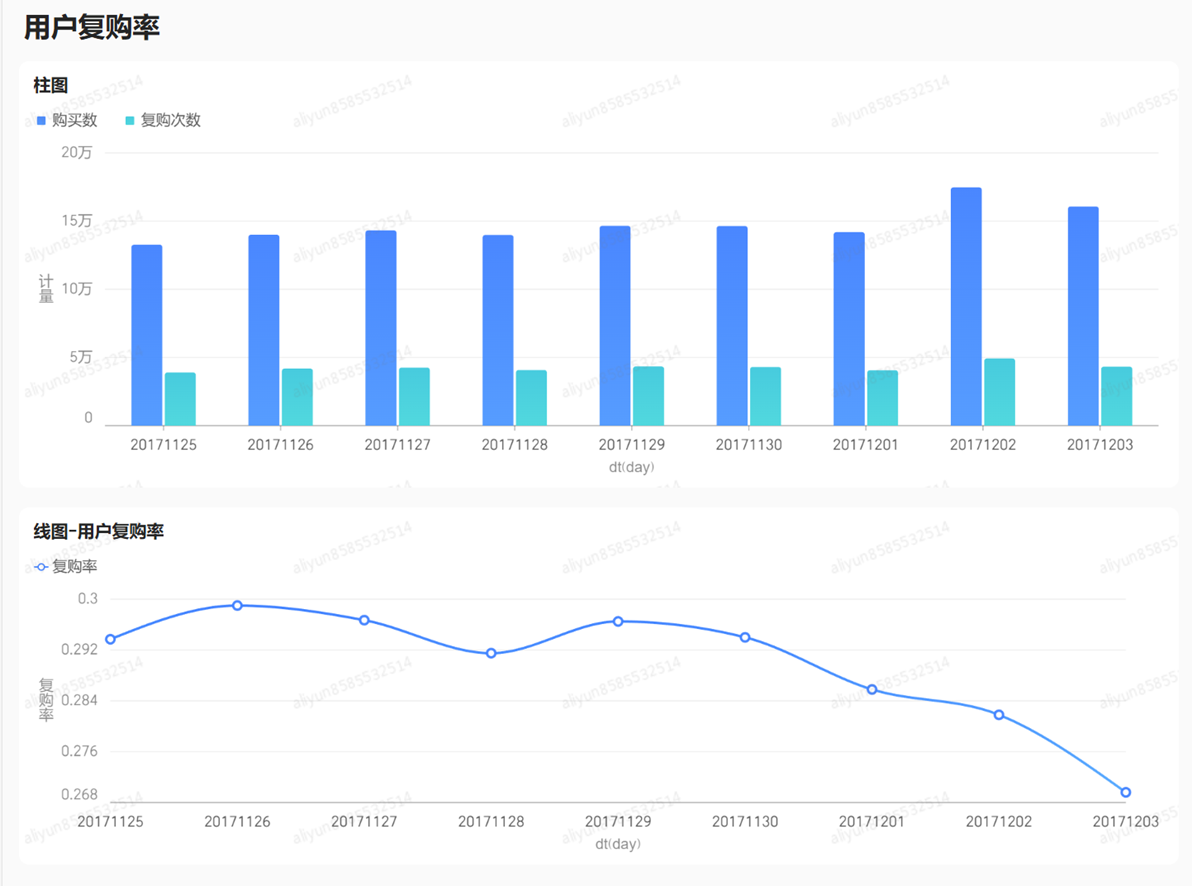
**热门类目 Top10**

****

女装和男装热度最高，可以发现人们网购基本上是以衣物为主，其次是家电数码，搞活动可以在这些品类里搞

**用户粘性维度 ：**

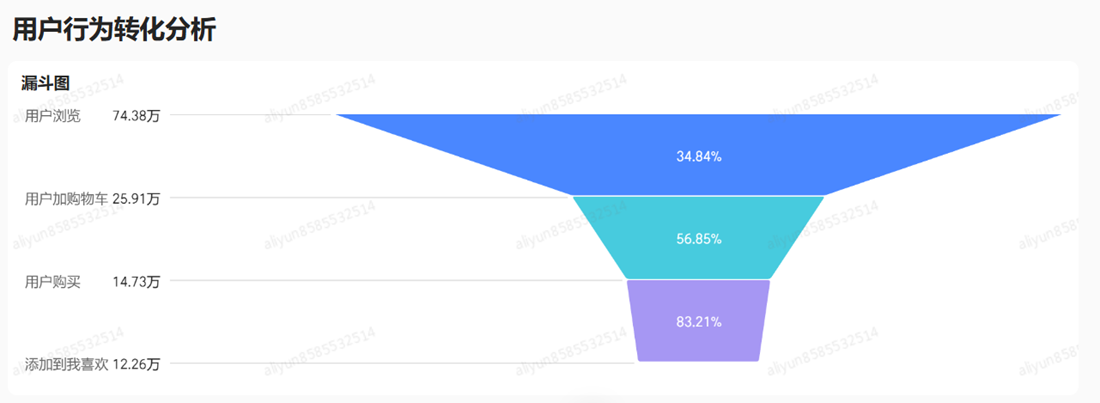
**用户复购率**

****

可以发现淘宝的用户复购率基本在25%以上，远高于通常的10%-20%，说明用户粘性很高，平台很健康。十二月二号跟三号是周末，可能在搞活动导致复购率出现波动，因为整体的购买量提升了很多

**转化维度：**

**用户行为转化表**

****

****

从浏览到购买的整体转化率为 20%，而加购物车到购买的转化率在50%以上，让用户添加到喜欢最难，想要提高购买率可能需要优化购物车结算流程或发放优惠券刺激下单。

## 结语

本次“淘宝用户行为数据仓库系统的设计与实现”项目，基于阿里天池的公开数据，从业务需求出发，系统性地完成了数据仓库的架构设计、模型构建、ETL流程开发以及多维可视化分析。采用经典的维度建模方法，构建了完整的四层数据仓库架构（ODS → DWD → DWS → ADS）。

项目重点解决了数据集成、清洗、关联、聚合及性能优化等一系列关键问题，通过分区策略、列式存储与压缩技术的应用，具备良好的查询效率与可扩展性。在分析层面，围绕流量、转化、商品与用户四大业务场景，设计并计算了十余项核心KPI指标，通过QuickBI实现了直观的可视化展示。

本项目构建的数据仓库系统，实现了对淘宝用户行为数据从原始存储到商业洞察的全链路管理。系统具备清晰的表结构设计、可复用的ETL流程、标准化的指标口径以及高性能的查询支持，能够有效支撑电商场景下的多维分析与决策需求。未来可进一步扩展实时数据处理能力、引入用户画像标签体系、对接推荐与风控等业务系统，形成更加完整的数据赋能闭环。