Data Model / Data Sets Description

Behavioural information about each user that can be used to infer user's interests are provided as table datasets - SearchInfo, AdsInfo, trainsearchstream, phonerequestsstream, visitsstream.

SearchInfo - details of searches performed by user includes search query, search parameters, userID, categoryID, LocationID, IP address of user- IPID.

One search per row.

Can be joined to category, location, UserInfo

Level in Location (once you join SearchInfo and Location on SearchLocationID = LocationID) represents geo level where a search (and impression) took place. Values could be 3 (city), 2 (region), or 1 (whole country).

Level in Category (once you join SearchInfo and Category on SearchCategoryID = CategoryID) represents category level where a search (and impression) took place. Values could be 3 (subcategory), 2 (category), or 1 (no specific category or subcategory were selected by visitor).

Join with UserInfo on UserID

-*Simulation of SearchInfo data*

-Bronze layer staging table

-Silver - Joining with tables and enriching the data

AdsInfo - Data of all the ads ie title/description, parameters, LocationID, categoryID, price, iscontext. Iscontext indicates whether it is contextual ad. Though this is master data, this has 37 million records.

Can be joined to category, location

Level in Location (once you join AdsInfo and Location using LocationID) represents geo level belongs to a non-contextual ad. It has to be 3 (city). Note that for context ad LocationID is NULL.

Level in Category (once you join AdsInfo and Category using CategoryID) represents category level belongs to any ad. Values could be either 3 (subcategory, for all non-contextual and some contextual ads), or 2 (category, for some contextual ads).

-*Simulation of AdsInfo data*

-Bronze layer staging table

-Silver - Joining with tables and enriching the data

> Location - Add Geo Level for non-contextual Ad – 3 (city). Location is null for contextual Ad. Else location id has value
>
> Category – 3 (subcategory, for all non-contextual and some contextual ads) or 2 (category, for some contextual ads).

Category - Master table of all categories. Level = 1/2/3

Location - Master table of all locations. Level = 1/2/3

UserInfo - Master table of all users

trainsearchstream - Each record describes one "impression" (an ad that is shown to a particular user based on a search). Fields are as follows:

SearchID - identifier for a visitors's search event.

AdID - identifier of the ad (see also ad description in AdsInfo.tsv).

Position - position of the ad in search result page (1 - is first ad on a page starting from the top). Only ads on position 1, 2, 6, 7, and 8 are logged.

ObjectType - type of the ad shown to user. The options are: 1 - regular free ads added by users; 2 - highlighted regular (owners have to pay fixed price to highlight them and stick to the top for some period of time); 3 - contextual ads (owners have to pay per visitor's click).

HistCTR - some naive history-based estimation of click-through rate for contextual ads, calculated when the ad is showed. For non-contextual ads this field equals NULL.

IsClick - 1 if there was a click on this ad. Otherwise 0. For non-contextual ads this field equals NULL.

Can be joined with SearchInfo on Search Id, AdsInfo on AdID

VisitsStream.tsv, PhoneRequestsStream.tsv

These are samples of users' visits to non-contextual ad landing pages and the corresponding phone request (if one occurred). Each ad's landing page shows the hidden seller's phone number. To be able to contact the seller, the user needs to click the request phone button, which indicates a high level of interest and is in "PhoneRequestsStream"

Both can be joined with UserID on UserInfo, AdID on AdsInfo

PhoneRequestsStream has phonerequestdate instead of viewdate

Events after impression

Can be joined to impressions for conversion analysis

-Bronze layer staging table

-Silver - Joining with tables and enriching the data

Merge the tables using UserID and AdID

testsearchstream is not used as trainsearchStream has almost same data along with isclick.

trainsearchStream_staging, trainsearchstream_silver, csv_extract_marker, staging_extract_marker are the tables I created for bronze & silver.

SampleSubmission and **SampleSubmission**_HistCTR has some prediction data for click through rate of contextual ads and not used.

Gold layer Analysis

A. Ad Performance Analytics

- **CTR (Click Through Rate) by ad type**
  - CTR = clicks / impressions - Helps identify which ads (highlighted, contextual, free) perform better.
- **Top ads by clicks / conversions (phone requests / visits)**
- **Revenue contribution by ad type**
  - For contextual ads: Revenue = clicks × CPC
  - For highlighted ads: Revenue = flat highlight fee

B. User Behavior Insights

- **User interest profiling**
  - Aggregate categories, locations, and ad types searched by a user.
  - Example: “User123 is mostly searching for *real estate in Mumbai*.”
- **Engagement metrics**
  - Avg. impressions per search, avg. CTR per user, repeat visits.
- **Conversion funnel**
  - Search ⟶ Impression ⟶ Click ⟶ Visit ⟶ Phone Request.

C. Search & Market Trends

- **Trending categories & locations**
  - Which categories are seeing more searches over time (e.g., “Cars up 15% MoM”).
- **Search demand vs ad supply**
  - Searches in “Category X” vs ads available in the same category.
- **Seasonality**
  - Ads CTR or search frequency by time of day, day of week, or month.

D. Ad Quality & Pricing Insights

- **Impact of price on CTR**
  - Average CTR grouped by ad price ranges.
- **Effect of ad type**
  - Compare CTR across free, highlighted, contextual.
- **High CTR ads with low conversions**
- **Ads with strong engagement but poor conversion (indicating possible fraud or misleading content).**

E. Location & Category Drilldowns

- **Geo-level performance**
  - **CTR, conversion rate, and impressions at city / region / country level.**
- **Category hierarchy analysis**
  - Electronics > Mobile Phones > Smartphones ⟶ CTR, conversions, demand trend.

F. Fraud & Anomaly Detection (ML features)

- **Unusual activity detection**
  - Users with abnormally high CTR / clicks in short time.
- **Ad anomaly**
  - Ads with high impressions but **0 clicks** (possible poor quality).
- **User session features**
  - Avg. time between search and click
  - # of different categories searched.

G. ML Feature Store

For training ML models (e.g., CTR prediction, recommendation engine):

- **User-level features**
  - Avg CTR, most searched categories, location preference.
- **Ad-level features**
  - Price, ad type, past CTR, category, location.
- **Search-level features**
  - Query text embedding, time of search, category hierarchy.
- **Interaction-level features**
  - Position on page, historic CTR at that slot.

Analysis on Whether table or View is required in gold layer for each business use case scenario

Grouping	Sub Group	Recommended Gold storage	Reason
A. Ad Performance Analytics			
	CTR (Click Through Rate) by ad type / category / region *Helps identify which ads (highlighted, contextual, free) perform better.*	Table Gold_ctr_adperf	Heavy aggregation, joins with phonerequestsstream. Expensive to compute each time
	Top ads by clicks / conversions (phone requests / visits)	Table gold_top_ads	“
	Revenue contribution by ad type *For contextual ads: Revenue = clicks × CPC* *For highlighted ads: Revenue = flat highlight fee*	Table gold_adtype_revenue	“

B. User Behavior Insights			
	User interest profiling - Aggregate categories, locations, and ad types searched by a user. *Example: "User123 is mostly searching for real estate in Mumbai."*	View Gold_user_profiling	Join with UserInfo only which has less rows
	Engagement metrics - Avg. impressions per search, avg. CTR per user, repeat visits.	Table gold_engagement_metrics	Heavy aggregations/ expensive joins with VisitsStream Also needed indexing on userid on VisitsStream
	Conversion funnel: Search → Impression → Click → Visit → Phone Request	Table gold_conversion_funnel	Heavy joins. VisitsStream and PhoneRequestsStream
C. Search & Market Trends			
	Trending categories & locations *Which categories are seeing more searches over time (e.g., "Cars up 15% MoM").*	Table Gold_search_trends	Heavy aggregation with searchInfo Expensive to compute each time
	Search demand vs ad supply *Searches in "Category X" vs ads available in the same category.*	Table gold_search_vs_supply	Expensive subquery joins
	Seasonality - Ads CTR or search frequency by time of day, day of week, or month.	View gold_seasonality	Less expensive query
D. Ad Quality & Pricing Insights			
	Impact of price on CTR *Average CTR grouped by ad price ranges.*	View Gold_avgctr_by_price	Single table query
	Effect of ad type *Compare CTR across free, highlighted, contextual*	View gold_ctr_by_adtype	Single table query (returning summarized single row)
	High CTR ads with low conversions	Table gold_highctr_lowconv	Heavy joins with PhoneRequestsstream and VisitsStream
	Ads with strong engagement but poor conversion	Table gold_suspicious_ads	"

	(indicating possible fraud or misleading content)		
E. Location & Category Drilldowns			
	Geo-level performance - CTR, conversion rate, and impressions at city / region / country level.	Table Gold_geo_perf	Heavy aggregation with AdsInfo
	Category hierarchy analysis - Electronics > Mobile Phones > Smartphones → CTR, conversions, demand trend	Table Gold_cat_heirarchy	Costly joins / aggregation with Phonerequestsstream, searchInfo
F. Fraud & Anomaly Detection (ML features)			
	Unusual activity detection - Users with abnormally high CTR / clicks in short time	View Table is historical tracking is needed, going with view Gold_unusual_activity	Can be done on the fly in single table
	Ad anomaly - Ads with high impressions but 0 clicks (possible poor quality)	View Gold_ad_anomaly	“
	User session features - Avg. time between search and click, # of different categories searched	View Gold_user_session_anomaly Gold_User_categories_searched	“
G. ML Feature Store			
	User-level features - Avg CTR, most searched categories, location preference	Table gold_user_mlfeatures	Expensive Subqueries
	Ad-level features - Price, ad type, past CTR, category, location	View gold_ad_mlfeatures	Single table summary
	Search-level features - Query text embedding, time of search, category hierarchy	Table gold_search_mlfeatures	Expensive join with SearchInfo
	Interaction-level features - Position on page, historic CTR at that slot	View gold_interaction_mlfeatures	From single table
	CTR Prediction	View Gold_ctr_prediction	From single table