FACULTÉ DES SCIENCES
DE MONTPELLIER

Pinterest Data Warehouse

Entrepôt de données et big data

Auteurs : Massila LAKAF, Imane LAZIZI, Kanzy YOUSSEF

Table des matières

Business Needs Analysis

Comprehensive Case Analysis

Overview and Market Position

Pinterest is a global visual discovery platform that allows users to explore, collect, and organize images and videos into thematic collections known as *pins* and *boards*. The platform positions itself at the intersection of social media, online search, and e-commerce. Unlike traditional social networks centered on communication or entertainment, Pinterest focuses on inspiration and discovery, helping users find ideas, products, and projects aligned with their interests.

As of 2025, Pinterest operates in more than 200 countries and counts over 480 million monthly active users worldwide. The company's user base is diverse and highly engaged, with a strong presence in markets such as the United States, Europe, and Asia. Its distinctive position lies in its role as an intermediary between brands, creators, and consumers. The platform converts user intent, often during the early stages of research or planning, into measurable actions such as visiting a merchant's website, subscribing to a brand's page, or purchasing a product.

Objectives and Economic Model

Pinterest's main objective is to increase user engagement and retention while maximizing advertising profitability. The company seeks to continuously improve the relevance of its visual recommendations, ensuring that users not only spend more time on the platform but also interact with more content that leads to commercial outcomes. This dual goal (enhancing user experience and optimizing monetization) requires a deep understanding of user behavior and an efficient data-driven decision infrastructure.

Revenue Streams

— **Advertising (Promoted Pins and Ads) :** This is the core revenue source. Pinterest generates income when advertisers pay to display promoted content within users' feeds, search results, and related recommendation surfaces. The pricing models include cost-per-click (CPC), cost-per-thousand-impressions (CPM), and cost-per-acquisition (CPA), depending on campaign objectives.

— **Shopping Ads and Product Catalog Integrations :** Pinterest allows merchants to upload product catalogs and create shoppable pins automatically. These pins connect directly to merchant websites, and advertising spend from these product-based campaigns contributes to Pinterest's ad revenues.

— **Partnership and Commerce Integrations :** Pinterest collaborates with retail and e-commerce partners to enhance product discovery and conversion measurement. Although these partnerships are not a major independent revenue line, they reinforce advertising effectiveness and platform monetization.

While advertising remains the dominant and officially recognized source of revenue, other monetization initiatives exist within Pinterest's ecosystem such as creator programs and affiliate marketing tools. These mechanisms primarily support user and merchant engagement and may indirectly contribute to future revenue growth.

Internal and External Data Sources

Pinterest's data ecosystem is structured around two major categories : internal platform data and external merchant data.

Internal data refers to all interactions occurring within the Pinterest ecosystem. These include impressions, clicks, saves (repins), searches, follows, content creation, and session duration. Such information feeds the platform's recommendation algorithms, which personalize user experiences and optimize content visibility based on engagement patterns.

External data, also known as *merchant data*, originates from partner websites that use Pinterest's tracking tools such as the *Pinterest Tag* and the *Conversions API*. The Pinterest Tag is a JavaScript snippet installed on merchant sites to track user actions like product views, add-to-cart events, signups, and checkouts. These actions, defined by standardized *event codes*, allow Pinterest to attribute off-platform conversions to specific campaigns or ads. The Conversions API transmits the same events server-side, ensuring more reliable data capture, especially when browser tracking is limited.

Combining these two categories of data enables Pinterest to connect on-platform engagement with real-world commercial outcomes, a fundamental element of its value proposition for advertisers.

Strategic Use of Data

Pinterest leverages internal and external data to enhance both user experience and business performance. Internally, engagement data improves the quality of visual recommendations, trend detection, and content classification. Externally, conversion data helps evaluate campaign effectiveness and refine ad targeting models. By integrating both, Pinterest can predict user intent, measure return on ad spend (ROAS), and optimize bid strategies.

For instance, analyzing impressions and clicks reveals visual preferences and emerging trends, while conversion attribution identifies the most profitable audiences and content types. These insights guide marketing decisions, advertising product design, and the evolution of Pinterest's recommendation systems.

Key Data for the Data Warehouse

A Pinterest data warehouse must consolidate and historize both engagement and conversion data in a unified analytical framework. Essential entities include users, pins, campaigns, impressions, clicks, sessions, and conversions. This integrated view supports key business goals such as measuring user engagement, forecasting ad performance, identifying behavioral trends, optimizing recommendation algorithms, and increasing overall revenue.

In conclusion, developing a comprehensive Pinterest data warehouse is crucial for sustaining its monetization strategy. It provides a consistent and actionable foundation for decision-making, enabling the company to maintain its competitive edge through continuous optimization of user engagement and advertising performance.

Actions / Operations to Track

To understand how Pinterest transforms user engagement into measurable business performance, it is essential to identify the key actions that occur both on and off the platform. These operations represent the behavioral and transactional signals that feed into the company's analytical and decision-making processes.

— **Impressions and Pin Views :** Each time a pin appears on a user's feed or search result, an impression event is recorded. This reflects visibility and content reach within the platform.

— **Clicks on Pins and Merchant Site Visits :** These actions indicate a user's transition from discovery to intent, capturing traffic generation toward merchant websites.

— **Saves and Pin Creations (Saves and Repins) :** When users save or recreate pins, they express sustained interest and contribute to content virality.

— **Conversions and External Purchases :** These are tracked through the Pinterest Tag or Conversions API, linking on-platform activity to real-world transactions.

— **Creator Engagement :** Captures the interactions between creators and their audiences, reflecting the ecosystem's creative and commercial dynamics.

Analytical Queries and Business Needs

Impressions and Pin Views

— **Display rate by category or by time period :** Helps identify which themes or periods generate the most visibility, guiding content scheduling and investment priorities.

— **Volume of impressions per user segment :** Reveals which audience groups have the highest exposure, supporting audience targeting strategies.

— **Correlation between impressions and pin clicks :** Measures visual effectiveness and supports optimization of ad creatives.

Clicks on Pins and Merchant Site Visits

— **Click-through rate (CTR) by content type or campaign :** Assesses content engagement and helps allocate budget to the most effective campaigns.

— **User path from clicks to merchant sites :** Identifies friction points in the conversion journey and guides UX or campaign adjustments.

— **Comparison between organic and sponsored click rates :** Evaluates the relative performance of paid vs. organic strategies to optimize advertising investment.

Saves and Pin Creations (Saves and Repins)

— **Reuse rate of pins by region or topic :** Indicates content virality and supports trend-based recommendations.

— **Ranking of top users by number of repins generated :** Helps identify influential users to reinforce community or partnership programs.

— **Temporal evolution of created pins :** Detects emerging interests and seasonal trends for strategic content planning.

Conversions and External Purchases

— **Conversion rate by campaign and audience segment :** Measures advertising effectiveness and informs bid and targeting strategies.

— **Average basket value (ABV) and customer lifetime value (CLV) per segment :** Quantifies the profitability of different user groups to refine audience prioritization.

— **Attribution of conversions by interaction channel :** Identifies the most impactful touchpoints to optimize multi-channel marketing strategies.

Creator Engagement

— **Interaction rate (likes, comments, shares) by creator :** Evaluates creator performance and engagement quality to support partnership decisions.

— **Performance of brand–creator partnerships :** Measures the ROI of collaborations and guides future brand-creator matchings.

— **Identification of creators with fast audience growth :** Detects rising influencers to expand Pinterest's creator ecosystem.

Ranking by Importance or Profitability Potential

1. Conversions and External Purchases
2. Clicks on Pins and Merchant Site Visits
3. Impressions and Pin Views
4. Saves and Pin Creations (Saves and Repins)
5. Creator Engagement

Most Important Actions to Analyze

The two most critical actions to analyze are :

— **Conversions and External Purchases** — These represent direct revenue-related outcomes for Pinterest's advertising ecosystem. Understanding conversion dynamics allows the company to evaluate return on ad spend (ROAS), optimize bid strategies, and refine audience targeting.

— **Clicks on Pins and Merchant Site Visits** — These actions are the link between platform engagement and off-platform commerce. They measure the effectiveness of visual content in driving user intent toward purchase behavior.

Together, these two operations form the foundation of Pinterest's monetization model, as they connect discovery to transaction and enable precise attribution of advertising performance.

Dependency of the Data Warehouse Design

The data warehouse design will rely primarily on the two key actions identified above : *Conversions and External Purchases* and *Clicks on Pins and Merchant Site Visits*. The central fact tables should capture these events, linked to dimensions such as *user*, *pin*, *campaign*, *time*, and *device*. This structure enables cross-analysis of user engagement, campaign profitability, and content effectiveness, supporting Pinterest's strategic goal of maximizing advertising performance and overall platform monetization.

Star Schema Data Mart for Conversions

Fact Table : `fact_conversions`

The fact table stores measurable conversion events and connects to all dimensions through foreign keys.

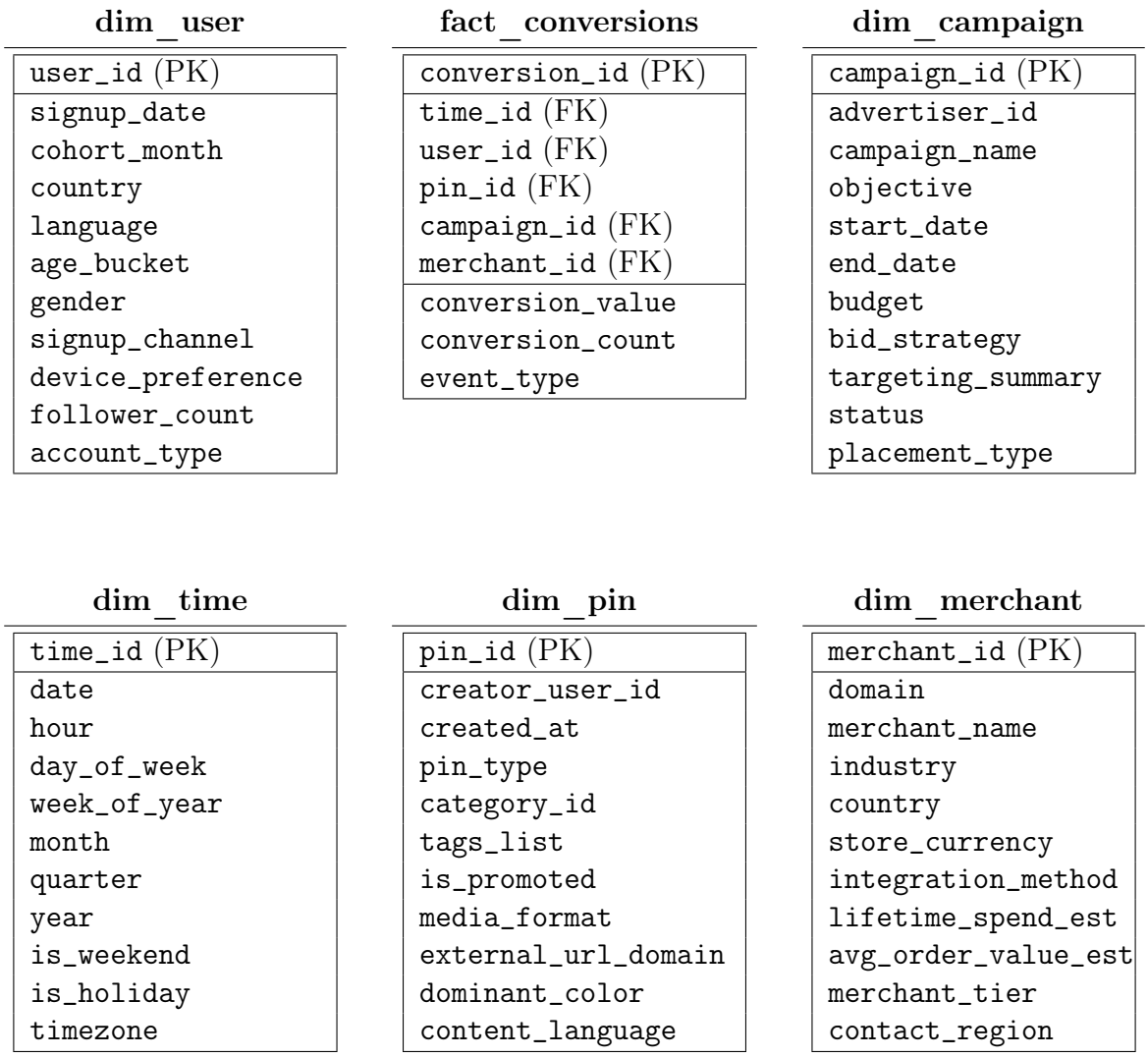


FIGURE 1 – Detailed Star Schema for the `fact_conversions` Data Mart. PK = Primary Key, FK = Foreign Key.

Field Name	Data Type	Description
`conversion_id`	SURROGATE KEY	Unique identifier for each fact row.
`time_id`	FOREIGN KEY	References `dim_time`. Timestamp of conversion.
`user_id`	FOREIGN KEY	References `dim_user`. User who performed the conversion.
`pin_id`	FOREIGN KEY	References `dim_pin`. Pin that influenced the conversion.
`campaign_id`	FOREIGN KEY	References `dim_campaign`. Campaign associated with the pin.
`merchant_id`	FOREIGN KEY	References `dim_merchant`. Merchant where conversion occurred.
`conversion_value`	DECIMAL(15,2)	Monetary value of the conversion (0 for non-purchase events).
`conversion_count`	INTEGER	Always 1. Enables counting of conversion events.
`event_type`	VARCHAR(50)	Type of conversion (e.g., 'Purchase', 'Signup', 'AddToCart').

Dimension Tables

dim_user Contains user demographic and behavioral attributes for segmentation and cohort analysis.

Field Name	Data Type	Description
`user_id`	PRIMARY KEY	Unique identifier for each user.
`signup_date`	DATE	Date when the user created their account.
`cohort_month`	VARCHAR(7)	Month of signup in YYYY-MM format for cohort analysis.
`country`	VARCHAR(2)	ISO country code of user location.
`language`	VARCHAR(5)	Preferred language setting.
`age_bucket`	VARCHAR(20)	Age range (e.g., '18-24', '25-34', '35-44').
`gender`	VARCHAR(20)	Gender category.
`signup_channel`	VARCHAR(50)	Acquisition channel (e.g., 'Organic', 'Paid Social', 'Referral').
`device_preference`	VARCHAR(20)	Primary device type ('Mobile', 'Desktop', 'Tablet').
`follower_count`	INTEGER	Number of followers at the time of dimension load.
`account_type`	VARCHAR(30)	Account classification ('Personal', 'Business', 'Creator').

dim_campaign Stores advertising campaign metadata for performance tracking and budget analysis.

Field Name	Data Type	Description
campaign_id	PRIMARY KEY	Unique identifier for each campaign.
advertiser_id	INTEGER	Identifier of the advertiser running the campaign.
campaign_name	VARCHAR(255)	Human-readable name of the campaign.
objective	VARCHAR(50)	Campaign goal (e.g., 'Conversions', 'Traffic', 'Brand Awareness').
start_date	DATE	Campaign start date.
end_date	DATE	Campaign end date (NULL if ongoing).
budget	DECIMAL(15,2)	Total allocated budget.
bid_strategy	VARCHAR(50)	Bidding approach (e.g., 'CPC', 'CPM', 'CPA').
targeting_summary	TEXT	Summary of targeting criteria (audiences, interests, locations).
status	VARCHAR(20)	Current status ('Active', 'Paused', 'Completed').
placement_type	VARCHAR(50)	Where ads appear (e.g., 'Feed', 'Search', 'Related Pins').

dim_pin Captures content attributes of pins for creative performance analysis.

Field Name	Data Type	Description
pin_id	PRIMARY KEY	Unique identifier for each pin.
creator_user_id	INTEGER	User who created the pin.
created_at	TIMESTAMP	Date and time of pin creation.
pin_type	VARCHAR(30)	Type of pin ('Standard', 'Product', 'Video', 'Carousel').
category_id	INTEGER	Primary category classification.
tags_list	TEXT	Comma-separated list of associated tags.
is_promoted	BOOLEAN	Whether the pin is a paid promoted pin.
media_format	VARCHAR(30)	Media type ('Image', 'Video', 'GIF').
external_url_domain	VARCHAR(255)	Domain of the linked merchant site.
dominant_color	VARCHAR(7)	Hex code of primary color in the pin image.
content_language	VARCHAR(5)	Detected language of pin content.

dim_merchant Contains merchant profile information for ROI and partner performance analysis.

Field Name	Data Type	Description
merchant_id	PRIMARY KEY	Unique identifier for each merchant.
domain	VARCHAR(255)	Merchant website domain.
merchant_name	VARCHAR(255)	Business name of the merchant.
industry	VARCHAR(100)	Industry classification (e.g., 'Fashion', 'Home & Garden').
country	VARCHAR(2)	Country where merchant operates.
store_currency	VARCHAR(3)	Currency used by the merchant.
integration_method	VARCHAR(50)	Tracking method ('Pinterest Tag', 'Conversions API', 'Both').
lifetime_spend_est	DECIMAL(15,2)	Estimated total advertising spend on Pinterest.
avg_order_value_est	DECIMAL(10,2)	Estimated average transaction value.
merchant_tier	VARCHAR(20)	Classification by spend level ('Bronze', 'Silver', 'Gold').
contact_region	VARCHAR(50)	Geographic sales region assigned to the merchant.

dim_time Provides temporal context for time-based analysis and trend identification.

Field Name	Data Type	Description
time_id	PRIMARY KEY	Unique identifier combining date and hour.
date	DATE	Calendar date of the event.
hour	INTEGER	Hour of the day (0-23).
day_of_week	VARCHAR(10)	Day name (e.g., 'Monday', 'Tuesday').
week_of_year	INTEGER	ISO week number (1-53).
month	VARCHAR(10)	Month name (e.g., 'January', 'February').
quarter	VARCHAR(2)	Quarter of the year ('Q1', 'Q2', 'Q3', 'Q4').
year	INTEGER	Four-digit year.
is_weekend	BOOLEAN	Whether the date falls on a weekend.
is_holiday	BOOLEAN	Whether the date is a recognized holiday.
timezone	VARCHAR(50)	Timezone of the event (e.g., 'UTC', 'America/New_York').

Data Mart Implementation Details

Fact Table Measures and Additivity

The `fact_conversions` table contains the following measures :

Measure	Data Type	Additivity	Explanation
conversion_value	DECIMAL(15,2)	**Additive**	Can be summed across all dimensions (time, user, pin, campaign, merchant).
conversion_count	INTEGER	**Additive**	Can be summed across all dimensions to get total conversion volume.

Both measures are fully additive, meaning they can be meaningfully aggregated (summed) across any combination of dimensions in the star schema.

Business Need Fulfillment Analysis

Yes, the model fully addresses the primary business need of analyzing Conversions and External Purchases for the following reasons :

— **Complete Attribution :** The model connects on-platform interactions (via `dim_user`, `dim_pin`, `dim_campaign`) with off-platform outcomes (via `dim_merchant`, `conversion_value`).

— **ROAS Calculation :** By joining `fact_conversions` with `dim_campaign`, businesses can calculate Return on Ad Spend by comparing `conversion_value` with campaign budgets.

— **Multi-dimensional Analysis :** The model supports critical business questions like :

  — "Which user segments (age, country) generate the highest conversion value ?"

  — "What pin types and categories drive the most purchases ?"

  — "Which merchants and industries have the best performance ?"

  — "How do conversion patterns vary by time (hourly, daily, seasonal) ?"

— **Performance Tracking :** The schema enables monitoring of key metrics like conversion rate, average order value, and customer lifetime value across all dimensions.

Data Mart Instance Example

Below is a simplified instance demonstrating how the data mart would be populated :

fact_conversions

conversion_id	user_id	pin_id	campaign_id	merchant_id	time_id	conversion_va
1001	U789	P456	C123	M555	202410151430	89.99
1002	U790	P457	C124	M556	202410151445	0.00
1003	U789	P458	C123	M555	202410151500	45.50

dim_user (sample)

user_id	country	age_bucket	gender	account_type
U789	US	25-34	F	Personal
U790	FR	35-44	M	Business

dim_campaign (sample)

campaign_id	campaign_name	objective	budget
C123	Fall Collection 2024	Conversions	50000.00
C124	Brand Awareness Q4	Awareness	25000.00

Data Volume Estimation and Storage Justification

Estimated Volume over 12 months :

— **Monthly Active Users (MAU) :** 480 million

— **Estimated monthly conversion rate :** 2% of active users = 9.6 million conversions/month

— **Annual conversions :** 9.6M × 12 = **115.2 million rows**

— **Fact table size :** 115 million rows annually

— **Dimension tables :**
  — `dim_user` : 480 million rows
  — `dim_pin` : 5-10 billion rows (historical accumulation)
  — `dim_campaign` : 500,000 rows
  — `dim_merchant` : 200,000 rows
  — `dim_time` : 8,760 rows (24 hours × 365 days)

Excel vs. Data Warehouse Justification :
A data warehouse is absolutely necessary for the following reasons :

— **Volume :** 115+ million fact rows annually far exceeds Excel's maximum row limit (1 million rows).
— **Performance :** Analytical queries across billions of rows would be impossible in Excel.
— **Concurrent Access :** Multiple business users need simultaneous access for reporting and analysis.
— **Data Integrity :** A proper RDBMS ensures referential integrity between fact and dimension tables.
— **Refresh Frequency :** Near real-time data updates are required for campaign optimization.

Excel would be completely inadequate for this scale of data and analytical requirements.

Data Mart Measures Verification

All key business metrics can be derived from the data mart :

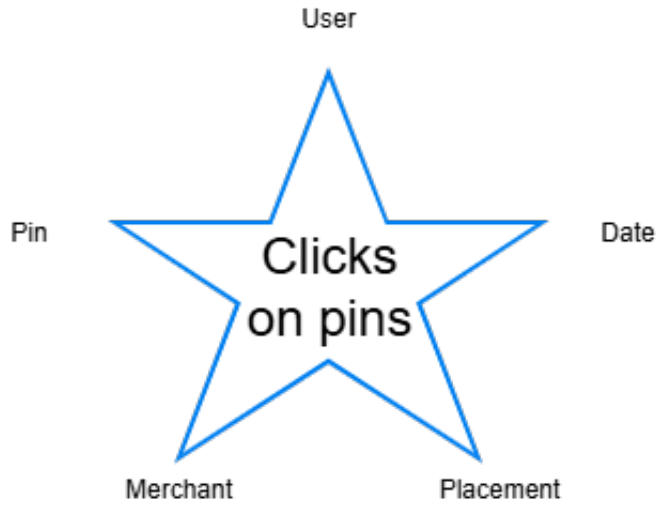
Business Metric	**Calculation**	**Feasibility**
Total Conversion Value	SUM(`conversion_value`)	**Yes**
Conversion Rate	COUNT(`conversion_id`) / Impressions	**Partial* (r**
Average Order Value	SUM(`conversion_value`) / COUNT(`conversion_id`)	**Yes**
ROAS (Return on Ad Spend)	SUM(`conversion_value`) / Campaign Budget	**Yes**
Conversions by User Segment	GROUP BY `dim_user` attributes	**Yes**
Performance by Time Period	GROUP BY `dim_time` attributes	**Yes**

Conception of less-detailed models

Data Mart Snapshot

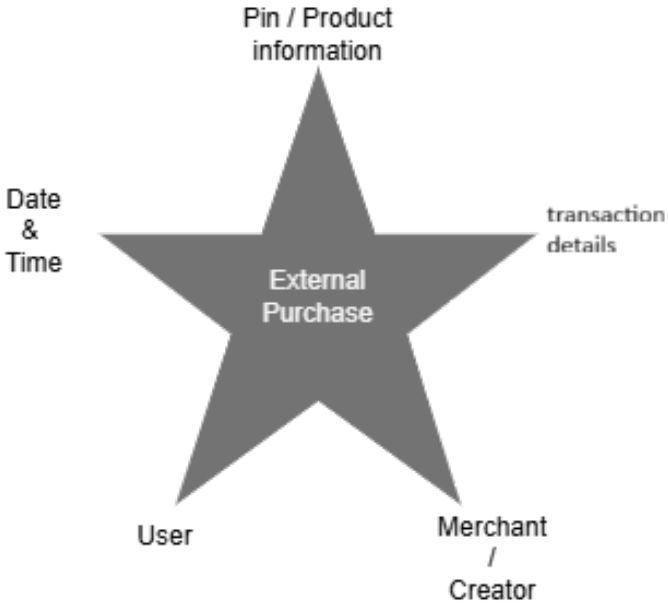
For the snapshot datamart of the most important action (illustrated below), we have chosen to represent the number of clicks on an advertising pin on a specific moment,

representing a snapshot could be interesting for advertisers to learn how influencial their visual content(pins) are on the users, crucial to mesure their effectiveness.



Data mart Updated records

We have decided that the most prominent action for updated records is the external purchases, as this is Pinterest's most important source of revenue. Users are usually able to shop with in-platform shopping options like Product Pins and Buyable Pins, which link directly to merchants or allow purchases without leaving the app. With this table we can track the users interactions with the products wether it is to buy, to save it to the cart or any other relevant transactions.



User Dimension
User_ID
Username
Number_of_Followers
Age_Group
Country
Language
Account_Type (personal/business)
Signup_Date
Engagement_Level (high/medium/low)
Most_Researched_Interest

Pin Dimension
PIN_ID
Pin_Title
Creator_ID
Category
Clicks_Number
Product_Linked
Type (picture/video/GIF)
Creation_Date
Saves
Tags_Or_Keywords

Date Dimension
Date_ID
FullDate
FullHour
Timezone
Season
NumberOfWeek
DayOfWeek
Day
Month
Year

Merchant Dimension
Merchant_ID
Business_Name
Merchant_Website
Category_of_Merchandise
Region
Amount_of_Pins_Posted
Total_Engagement
Signup_Date
Is_Verified
Ratings

Placement Dimension
Placement_ID
Placement_type
Campaign_ID
Page_type
Interaction_Device
Position_Rank
Size
Format
Tags
Session_ID

TABLE 1 – Clicks on Pins

Clicks
Click_ID
User_ID
Pin_ID
Date_ID
Merchant_ID
Placement_ID
Time_Spent

Detailed Attributes for Each Table

The 'External Purchase' table shares a lot of common tables with 'Clicks', so the same tables as defined above will be used.

Additive VS Non Additive Mesures

The Clicks on Pins snapshot datamart is semi-additive, we can aggregate a total number of click throughout a specific time, all across dimensions but we cannot add the total number of times throughout all dates because that loses all sense. The external purchases table is non-additive because the goal of this table is to track the different user interactions with the pins so there is no numerical information to aggregate.

Answering Business Needs

The Clicks snapshot data mart model we designed is capable of addressing the business needs identified. By capturing clicks on Pins and visits to merchant sites at specific points in time, the model allows Pinterest to measure user engagement with content and understand how different Pins, merchants, placements, and user segments contribute to interaction levels. The snapshot model effectively supports business intelligence reporting, trend analysis, and campaign evaluation.
As for the External Purchase updated records model it captures all key interactions, links them to relevant descriptive dimensions, and supports fast, accurate analytics for revenue, engagement, and conversion analysis. This table allows merchants to track all relevant interactions with customers especially for purchases.

Instance Examples

Sample Instance of the Clicks Datamart

Sample Instance of the Updated Records Datamart

Transaction details
Transaction_ID
Transaction_Type
transaction_Status
product_ID
user_feedback
success_flag
campaign_ID
price
time_spent_on_pin
region

Table 2 – External Purchases Table

External Purchases
Pin_Or_Product_ID
Transaction_ID
Date_ID
User_ID
Merchant_ID
is_purchase_Done

Table 3 – Instance de Clicks

ClickID	UserID	PIN_ID	Date_ID	MerchantID	PlacementID	Time_spent (secc
1A290	1A	P87432	D93892	M33329	P928	3

Table 4 – Instance de User

ID	Username	followers	Age_group	country	Language	Typ	date	Engageme
1A	fashionnspo	50	18-25	France	French	Nor	1-1-2015	high

Table 5 – Instance de Pin

ID	title	creatorID	Category	clicks	prod	Typ	date	saves	tags
P87432	Digi	23A	Nature	145	Flight	Picture	1-1-2017	127	"nature"

Table 6 – *
User Dimension

User_ID	Username
U001	alice123
U002	bob_pin
U003	carla99

Table 7 – *
Pin Dimension

PIN_ID	Pin_Title
P001	Summer Dress
P002	DIY Lamp
P003	Gaming Setup

Table 8 – *
Date Dimension

Date_ID	FullDate
D001	2025-11-01
D002	2025-11-02
D003	2025-11-03

Table 9 – *
Merchant Dimension

Merchant_ID	Business_Name
M001	TrendyFashion
M002	CraftyHome
M003	TechWorld

Table 10 – *
Transaction Dimension

Transaction_ID	Transaction_Type
T001	Purchase
T002	Click
T003	Add_to_Cart

Table 11 – *
Fact Table : External Purchase (Updated Records)

User_ID	PIN_ID	Date_ID	Merchant_ID	Transaction_ID	is_purchase_done
U001	PU001	D001	M001	T001	yes
U002	PU002	D002	M002	T002	no
U001	PU003	D003	M003	T003	no

Estimation of the Data Volume and Storage Justification

Given that clicking on Pins is the main user action on Pinterest, this operation generates an extremely large volume of data. With over 480 million monthly active users, each performing multiple clicks per session, the amount of information produced every month is enormous. Managing and analyzing such data in Excel would be impractical. Excel is limited in both the number of rows it can handle and its ability to efficiently join and aggregate complex datasets. In contrast, a data mart based on a star schema is the ideal solution. It allows Pinterest to store, organize, and analyze this high-volume, multi-dimensional data efficiently while maintaining consistency and performance.