Below is a **professionally formatted**, **clean**, **code-free**, developer-friendly version of the **Complete ML Data Dictionary**.
It includes:

✔ Beautiful, organized layout
✔ Complete tables with all columns
✔ All derived ML datasets (A–F)
✔ Explanations optimized for ML + MLOps teams
✔ Cross-Dataset ML Usage Map
✔ Storage & pipeline guidance
✔ No SQL, no Python (per your request)

You can paste this directly into your repo under:

📁 `docs/data_dictionary.md`
If you want a PDF export, just tell me.

📘 ML Data Dictionary — Complete & Reformatted

Version: v2.0 — Clean, code-free, fully structured
Prepared For: Machine Learning, Data Engineering, MLOps Teams
Project: Fashion E-Commerce ML System (Recommender, Forecasting, NLP, Segmentation)

TABLE OF CONTENTS

---

1. Base Database Tables (Canonical Raw Data)

These tables represent the **source-of-truth** for the ML pipelines.

🔷 Customers

Description:
Contains demographic and account-level attributes of each customer.

Column	Description
customer_id (PK)	Unique user identifier
age	Age of customer
postal_code	Postal/zip code
club_member_status	Member status (e.g., active, inactive, tier)
fashion_news_frequency	Subscription frequency
active	Whether account is currently active
first_name	First name *(PII – exclude from ML)*
last_name	Last name *(PII – exclude from ML)*
email	Email *(PII – hash/anonymize)*
signup_date	Account creation date

Purpose for ML:
Segmentation, personalization, RFM modeling, churn recency calculations.

🔷 Articles

Description:
Full metadata about products, including textual & categorical attributes.

Column	Description
article_id (PK)	Unique product identifier
product_code	Vendor/product code

Column	Description
prod_name	Product short name
product_type_name	Product type label
product_group_name	Product group classification
graphical_appearance_name	Pattern/appearance
colour_group_name	Standardized color
department_no	Department numeric identifier
department_name	Department label
index_name	Index category
index_group_name	Index grouping
section_name	Section label
garment_group_name	Garment group
detail_desc	Free text product description
price	Price of article
stock	Current stock quantity
category_id (FK)	Category reference
created_at	Date article was added
last_updated	Date article metadata was modified

Purpose for ML:
Content-based recommender, embeddings, similarity search, forecasting.

🔷 Categories

Description:
Hierarchical category system for articles.

Column	Description
category_id (PK)	Category ID
name	Category name
parent_category_id	Parent category (self-reference)

Purpose for ML:
Hierarchy embeddings, taxonomy-based recommendations.

🔷 Transactions

Description:
Line-level purchases for all customers.

Column	Description
transaction_id (PK)	Unique transaction
t_dat	Transaction timestamp
customer_id (FK)	Buyer
article_id (FK)	Item purchased
price	Price paid
sales_channel_id	Channel (web, store, app, etc.)

Purpose for ML:
Collaborative filtering, time series forecasting, RFM, LTV.

🔷 Orders

Column	Description
order_id (PK)	Unique order
customer_id (FK)	Buyer
order_date	Order timestamp
total_amount	Order total
payment_status	Paid / Pending / Refunded
shipping_address	Shipping details *(PII – exclude from ML)*

🔷 Order_items

Column	Description

Column	Description
order_item_id (PK)	Order item line
order_id (FK)	Order reference
article_id (FK)	Product
quantity	Units
unit_price	Price per unit
line_total	unit_price × quantity

🔷 Reviews

Column	Description
review_id (PK)	Unique review
customer_id (FK)	Reviewer
article_id (FK)	Reviewed item
rating	Rating (1–5)
review_text	Free text content
created_at	Review date

Purpose for ML:
NLP sentiment, topic modeling, embedding-based ranking.

🔷 Events

Contains user behavior (implicit signals).

Column	Description
event_id (PK)	Unique event
session_id	Session identifier
customer_id (FK)	User *(nullable for guests)*
article_id (FK)	Involved article
event_type	view / click / cart / wishlist / buy / etc.

Column	Description
campaign_id	Marketing origin
created_at	Event timestamp

Purpose for ML:
Session modeling, funnels, re-ranking, real-time recommendations.

🔷 Cart

Column	Description
Cart_id (PK)	Unique cart entry
Customer_id (FK)	Owner
Article_id (FK)	Item
Quantity	Quantity added
Added_at	Timestamp

🔷 Wishlist

Column	Description
Wishlist_id (PK)	Unique wishlist entry
Customer_id (FK)	Owner
Article_id (FK)	Wished item
added_at	Timestamp

---

2. Derived ML Datasets (A–F)

These are the datasets your downstream ML models ACTUALLY use.

🟩 Dataset A — User-Item Interaction Dataset (CF)

Purpose: Collaborative filtering, implicit feedback modeling.

Columns

Column	Description
customer_id	User
article_id	Product
purchase_count	Total purchases of this article by the user
total_spent	Sum of all prices for this user–item pair
first_purchase_date	First time user bought this item
last_purchase_date	Most recent purchase
recency_days	Days since last purchase
time_weighted_score	Recency-adjusted weighting
avg_price_paid	Average unit price
last_sales_channel	Most recent sales channel for this pair

🟩 Dataset B — Article Content Feature Dataset

Purpose: Content-based recommender, similarity search, hybrid models.

Structured Columns

Column	Description
article_id	Identifier
product_code	Product code
prod_name	Name
product_type_name	Product type
product_group_name	Product group
graphical_appearance_name	Appearance/pattern
colour_group_name	Color
department_no	Numeric department

Column	Description
department_name	Department name
index_name	Index label
index_group_name	Index grouping
section_name	Section
garment_group_name	Garment type
detail_desc	Raw description
created_at	Creation timestamp
last_updated	Last update
price	Monetary price
stock	Available quantity
category_id	Category
category_name	Category name
parent_category_id	Parent category
category_depth	Hierarchy depth (derived)
lifecycle_days	Days since creation
normalized_price	Scaled price
normalized_stock	Scaled stock

Embedding/Text Columns

Column	Description
tfidf_vector_ref	Reference to TF-IDF vector
bert_embedding_ref	Reference to BERT embedding

🟩 Dataset C — Customer Feature Dataset

Purpose: Segmentation, personalization, LTV, hybrid ranking.

Columns

Column	Description
customer_id	Unique user
age	Demographic
signup_date	Signup timestamp
signup_age_days	Account lifetime
club_member_status	Membership
fashion_news_frequency	Engagement via newsletter
active	Active/inactive status
total_transactions	Count
total_amount_spent	Sum of spend
avg_transaction_value	Average spend per transaction
first_purchase_date	First transaction
last_purchase_date	Most recent purchase
recency_days	Recency metric
frequency_6m	6-month purchase frequency
frequency_12m	12-month frequency
monetary_6m	6-month spend
monetary_12m	12-month spend
R_score	Recency percentile bucket
F_score	Frequency percentile bucket
M_score	Monetary percentile bucket
RFM_score	Combined score
total_orders	Orders placed
avg_basket_size	Items per order
total_items_bought	Total units
wishlist_size	Wishlist item count
cart_additions	Total cart events
cart_abandon_rate	Carts without purchase
most_used_sales_channel	Dominant channel

Column	Description
top_category	Most purchased category
category_distribution_ref	Vector representing category preferences
session_events	Total events
views	Product views
clicks	Click-throughs
carts	Cart events
buys	Purchases
conversion_rate	buys/views
click_through_rate	clicks/views
add_to_cart_rate	carts/clicks
view_to_buy_rate	buys/views
dominant_event_type	Most frequent behavior
lifetime_value_estimate	Optional LTV feature

🟩 Dataset D — Time-Series Dataset (Daily & Weekly)

Purpose: Demand forecasting, inventory planning, category-level trends.

Columns

Column	Description
article_id	Product
date	Day
daily_sales	Unit sales
daily_revenue	Revenue
avg_price	Average price
is_weekend	Weekend flag
weekday	Day of week
month	Month

Column	Description
year	Year
rolling_7	7-day average
rolling_30	30-day average
category_id	Product category
promo_flag	Promotion indicator
holiday_flag	Holiday indicator

🟩 Dataset E — Reviews NLP Dataset

Purpose: Sentiment analysis, review embeddings, re-ranking.

Columns

Column	Description
review_id	Review
customer_id	Reviewer
article_id	Product
category_id	Product category
rating	Rating 1–5
review_text	Raw review
clean_text	Preprocessed review
created_at	Timestamp
review_age_days	Age of review
sentiment_score	Continuous sentiment (0–1)
sentiment_label	Negative/Neutral/Positive
tfidf_vector_ref	TF-IDF vector reference
bert_embedding_ref	Embedding
toxicity_score	Toxicity detection score
topic_id	Topic cluster

🟩 Dataset F — Behavioral Data (Events, Sessions, Funnels, Customer Intent)

Dataset F1 — Enriched Event-Level Dataset

Column	Description
event_id	Identifier
session_id	Session
customer_id	User
article_id	Product
category_id	Category
event_type	View/click/cart/buy/etc.
campaign_id	Campaign
created_at	Timestamp
event_hour	Hour of day
event_day	Day of week
session_start	First event time
time_since_session_start	Seconds since session began
is_conversion	Whether event is a purchase

Dataset F2 — Session Funnel Dataset

Column	Description
session_id	Session
customer_id	User
views	Count of view events
clicks	Count of click events
carts	Count of cart events
buys	Count of purchase events
wishlist	Wishlist interactions

Column	Description
first_event	Start timestamp
last_event	End timestamp
converted	Boolean flag
time_to_convert	Duration until purchase
funnel_stage	Final stage reached

Dataset F3 — Customer Behavior/Intent Dataset

Column	Description
customer_id	User
total_events	All events
total_sessions	Session count
avg_session_length_seconds	Duration average
views	Total views
clicks	Total clicks
carts	Total cart additions
buys	Total purchases
wishlist_events	Wishlist actions
conversion_rate	buys/views
click_through_rate	clicks/views
add_to_cart_rate	carts/clicks
cart_abandon_rate	1 – buys/carts
view_to_buy_rate	buys/views
dominant_event_type	Most frequent behavior

---

3. Cross-Dataset ML Usage Map

✔ Collaborative Filtering (CF)

Uses:

- Dataset A (user–item interactions)
- Transactions (raw support)
- Customer segments for cold-start

✔ Content-Based Recommendations

Uses:

- Dataset B (article structured + embeddings)
- Reviews NLP embeddings (optional re-ranking)

✔ Hybrid Recommendation System

Combines:

- CF scores (Dataset A)
- Content similarity (Dataset B)
- Behavioral re-ranking (Datasets F1, F2, F3)

✔ Customer Segmentation

Uses:

- Dataset C (full enriched customer features)
- Dataset F3 (behavioral vectors)

✔ Demand Forecasting

Uses:

- Dataset D (daily & weekly time series)
- Article & category metadata (Dataset B)
- Stock levels (from articles)

✔ Trend Analysis

Uses:

- Dataset D (seasonality patterns)
- Dataset F1/F2/F3 (behavior shifts)
- Reviews (sentiment direction)

✔ Sentiment & NLP

Uses:

- Dataset E (text, sentiment, embeddings)
- Articles metadata for context

✔ Conversion & Funnel Modeling

Uses:

- Dataset F1 (events)
- Dataset F2 (session funnel)
- Dataset F3 (behavior history)

---

4. Dataset Storage & Naming Conventions

Recommended directory structure:

```
data/
  ml/
    A_user_item_interactions.parquet
    B_articles_structured.parquet
    B_articles_tfidf_vectors.npz
    B_articles_bert_embeddings.npy
    C_customer_features.parquet
    D_timeseries_daily.parquet
    D_timeseries_weekly.parquet
    E_reviews.parquet
    E_review_embeddings.npy
    F_events_enriched.parquet
    F_session_funnel.parquet
    F_customer_behavior.parquet
```

Principles:

- Use **Parquet** as the default format
- Use **NumPy/Torch** for embeddings
- Use **NPZ** for sparse vectors

- Store all artifacts under versioned folders
- Add metadata files (`schema.json` , `created_at` , `version`) for reproducibility

5. Data-Quality & Validation Guidelines

Key checks:

- Unique primary keys
- No NULL FK values (except allowed fields)
- Positive price, positive quantity
- Soft PII removal from ML exports
- Timestamps not in the future
- Category hierarchy integrity
- No duplicated (article_id, date) pairs in time series

Validation recommendation:

- Create automated checks using Great Expectations or custom Airflow tasks.

6. Feature Engineering Guidelines

Dates:

- Extract recency, seasonality, day-of-week, month
- Compute lifecycle features for articles
- Use consistent timezone

Categorical:

- Low cardinality → one-hot
- High cardinality → embeddings or target encoding

Numerical:

- Scale prices & stock using robust or quantile scaling
- Log transform heavy-tailed behavior features

Text:

- Clean, normalize, lemmatize
- Compute TF-IDF for lexical similarity
- Compute BERT embeddings for semantic similarity
- Store both sparse and dense versions

Behavioral:

- Build rolling funnels (last 24h, 7d, 30d)
- Compute per-user behavior ratios
- Apply time-decay for recent events

RFM:

- Use fixed reference date across all RFM metrics
- Store raw R, F, M and percentile-bucketed scores

---

7. Indexing, Partitioning & Refresh Cadence

Index recommendations:

- Transactions: index by customer_id, article_id, t_dat
- Events: index by session_id, created_at
- Orders: index by customer_id, order_date

Partitioning:

- Events → daily/monthly partitions
- Transactions → monthly partitions

Refresh frequencies:

- User-item interactions → daily
- Customer features → daily
- Content embeddings → monthly or when product catalog changes
- Time series → daily
- NLP embeddings → weekly or monthly

- Funnels & behavior → hourly or daily

---

8. Privacy & Operational Notes

Remove or anonymize:

- first_name
- last_name
- email
- shipping_address

Keep secure:

- customer_id mappings
- raw reviews (if containing personal references)

Operational guidance:

- Store ML artifacts with version tags
- Track schema evolution
- Maintain change logs
- Ensure reproducibility with timestamps + metadata
