SHENKAR — שנקר
ENGINEERING. DESIGN. ART . הנדסה. עיצוב. אמנות
The Pernick Faculty of Engineering . הפקולטה להנדסה ע"ש פרניק
המחלקה להנדסת תוכנה
Software Engineering Dept.

Nail Technician

Software Engineering B.Sc. Big Data Project

Software Design Document

Authors:

Yoav Zucker

Oranit Yogev

Nave Maymon

Supervisor: Miri Nudelman

18/05/2025

Data Sources

1. streaming data source - Real-Time Customer Ratings

Attribute	Details
Source	Customer reviews and ratings
Description	After each appointment, the customer is invited to rate the service. The review is submitted in real time via a mobile app or web form
Real-Time Processing	The data is immediately streamed through platforms such as Kafka or Azure Event Hub and ingested into the Bronze layer
Event-Driven Architecture	Each review submission triggers an event that activates the ingestion process into the system
Business Use	Enables immediate tracking of customer satisfaction, early detection of service issues, and enhancement of the overall customer experience

These are live, real-time data points that provide **a** direct indicator of service quality, making this a critical source for rapid monitoring and response

2. Late Arrival Source – Instagram Post Engagement

Attribute	Details
Source	Engagement data from Instagram posts (likes, comments)
Description	Every night, a CSV/Excel file is exported from the inventory management system, with details such as available colors, quantities, shortages, by branch
Process	The file is loaded in a daily batch into the Bronze layer, and then undergoes processing and cleaning into the Silver layer
Business Use	Allows knowing which colors are in stock, when to order products, and which colors are most popular based on usage

This is relatively slow data, but it is critical for smart inventory management and accurate ordering

3. Additional Data – Daily Inventory Update

Attribute	Details
Source	Update of the list of colors and raw materials in stock
Description	The information is collected from the Instagram API once a day, but the actions themselves (likes/comments) may occur up to 48 hours after the post was published
Handling Delay	The data is received with the original timestamp (the event's timestamp), and the system knows how to update tables and correct existing aggregations using watermarking or upserts
Business Use	To understand which colors/designs generate the most engagement, identify trends, and optimize marketing

Here we deal with data that is not chronologically ordered and with late completions – and this simulates a real-world scenario of working with third-party analytics

Data Modeling Requirements

1. Static Dimension Table

Colors Dimension Table

Represents all the colors available in the nail studio – including color code, name, status (active/inactive), category, etc

Color ID	Color Name	HEX Code	Active	Category
101	Funny Bunny	#D9CDBD	Yes	Nude
102	Soft Lavender	#B497BD	Yes	Pastel Purple
103	Pink	#FFC0CB	No	Classic Pink

Use Cases:

- Filter reports by color
- Analyze popularity by color category
- Display color names instead of IDs in dashboards

Branches Dimension Table

Represents the different physical locations (branches) of the nail studio. This table is used to link sessions, employees, and revenue data to their respective branch

Branch ID	Chain Name	Branch Name	City	Is Active	Opening Date
201	NailGlow	TLV	Tel Aviv	Yes	01/03/2020
202	NailGlow	Haifa	Haifa	Yes	15/08/2021
303	NailGlow	Eilat	Eilat	No	10/12/2019

Use Cases:

- Segmenting reports by location (e.g., revenue per branch)
- Filtering data by active/inactive branches
- Mapping customer trends based on geographic location

Employees Dimension Table

This dimension represents the staff members (for example: manicurists) working in a nail studio. The table allows to analyze individual performance, cross-reference data with sessions, enter salaries, and generate insights by employee

Experience Level	Active	Employment Date	Branch ID	Role	Full Name	Employee ID
Senior	Yes	01/01/2020	201	Manicurist	Dana Levi	501
Junior	Yes	12/05/2021	202	Pedicurist	Hila Cohen	502
Senior	No	01/03/2018	203	Manager	Shiri Kaze	503

Use Cases:

- Join session data by employee (who performed which treatment)
- Segment revenue by employee
- Analyze performance by experience level or role
- Analyze employee turnover (active/inactive employees)

Treatments Dimension Table

This dimension represents the types of treatments offered at the nail studio (manicure, pedicure, gel, special designs, etc.). This table enables analysis of popular treatment types, average treatment duration, pricing, and more

Active	Base Price	Duration (Min)	Category	Treatment Name	Treatment ID
Yes	80₪	30	Hand Cream	Classic Manicure	301
Yes	120₪	45	Hand Cream	Gel Polish	302
No	150₪	60	Foot Cream	Deluxe Pedicure	303

Use Cases:

- Filter session data by treatment type or category
- Analyze popularity and frequency of specific treatments
- Cross-reference with employee performance (e.g., who does most gel treatments)
- Track pricing trends and duration averages

- Displaying treatment names instead of IDs in dashboards

2. Slowly Changing Dimension (SCD)

Table Name : Customers Table

The solution includes a Slowly Changing Dimension (SCD) Type 2 table that tracks historical changes in customer details over time.
This enables accurate business analysis based on the customer information that was valid at the time the event occurred, rather than based on current values.

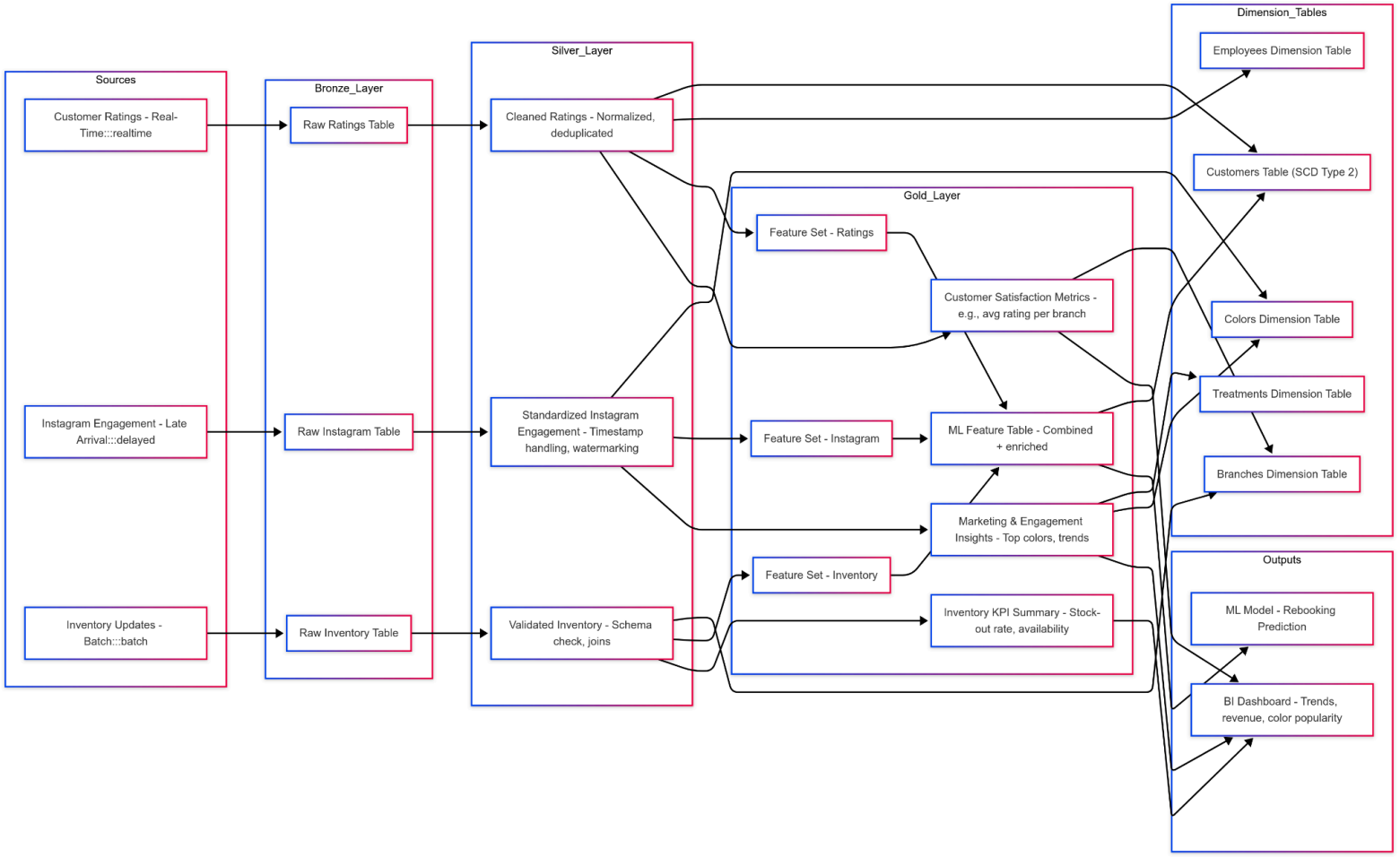
This table records changes in contact details — such as phone numbers or email addresses — by creating a new row for each change, with a clearly defined validity range.

Customer ID	Customer Name	Phone	Email	Valid From	Valid To	Is Current
501	Shiran Cohen	050-1234567	shiran@gmail.com	01/01/2023	10/03/2023	No
501	Shiran Cohen	052-9876543	shiran_new@gmail.com	11/03/2023	NULL	Yes

Business Use Cases:

- Analyzing customer satisfaction based on the contact details valid at the time of service
- Tracking changes in customer profiles for personalized marketing
- Understanding which changes influence customer retention and repeat visits
- Maintaining reporting accuracy over time (e.g., associating a customer with a previous branch or historical data state)

Architecture Requirements



The diagram illustrates the end-to-end flow of data across the system, beginning with three primary data sources:

- Real-time customer ratings - submitted immediately after a session
- Delayed Instagram engagement data, including likes and comments on marketing posts
- Daily batch inventory updates containing stock levels and availability across branches

These data sources are processed through the three standard Medallion Architecture layers:

🟤 Bronze Layer – Raw Data Ingestion

This layer captures and stores the raw data exactly as received:

- Customer reviews enter the Raw Ratings Table

- Instagram interactions are stored in the Raw Instagram Table
- Daily inventory exports are loaded into the Raw Inventory Table

No transformation or cleaning occurs at this stage, ensuring that all original details are preserved.

⚪ Silver Layer – Cleaning, Standardization & Validation

This layer prepares data for analytics and modeling:

- Cleaned Ratings are normalized and deduplicated
- Instagram Engagement Data is standardized with proper timestamp handling and watermarking for late arrivals
- Inventory Data undergoes schema validation and joins to ensure alignment with the business model

Each cleaned table is also connected to its relevant Dimension Tables, including Customers (SCD), Employees, Branches, and Colors.

🟡 Gold Layer – Business Metrics & Feature Engineering

In this layer, advanced business insights and predictive features are generated:

- Customer Satisfaction Metrics provide KPIs such as average rating per branch
- Marketing & Engagement Insights highlight popular colors and campaign performance
- Inventory KPI Summary tracks availability and stock-out patterns
- Three Feature Sets (Ratings, Instagram, Inventory) are consolidated into a final ML Feature Table, used to power predictive models

In summary

The system provides insights through two main output channels:

- A BI dashboard that displays satisfaction trends, revenue per branch, and color/category popularity
- A machine learning model that predicts the likelihood of customers reordering, based on combined behavioral and operational data

Business Requirements

1. Business Analytics Requirements
    - Define a key business question

Key Business Question - "Which gel nail polish colors lead to the highest customer satisfaction at each branch, and which colors are most popular during specific seasons?"

This question will provide insights into:

- Seasonal marketing strategies
- Inventory management
- Tailoring services based on customer preferences

    - Design a mock dashboard to answer this question

<u>Dashboard Components:</u>

<u>1 - Bar Chart – Top Colors by Season</u>

- X-axis: Color Names (from Colors Dimension Table)
- Y-axis: Number of Sessions (derived from Customer Ratings - Cleaned Ratings)
- Filter: Season (based on session timestamps)

Built from:

- Cleaned Ratings in Silver Layer
- Joined with Colors Dimension Table and Branches Dimension Table

Logic:

- Count sessions per color per season, segmented by branches if needed
- Seasonality can be derived by extracting month → mapped to season

<u>2 - Heatmap – Avg. Rating per Color & Branch</u>

- X-axis: Color (from Colors Dimension Table)
- Y-axis: Branch (from Branches Dimension Table)
- Color intensity: Average rating (from Customer Satisfaction Metrics)

Built from:

- Customer Satisfaction Metrics (Gold Layer)
- Uses Cleaned Ratings + joins with Colors, Branches, and Customers (SCD)

Logic:

- Average rating grouped by color and branch ID
- Can further filter by date range or employee ID (via Employees Dimension Table)

3 - Line Chart – Revenue Trends by Color Popularity

- X-axis: Month
- Y-axis: Revenue
- Lines: Top 5 most-used colors

Built from:

- Marketing & Engagement Insights (Gold Layer)
- Uses data from:
    - Standardized Instagram Engagement
    - Validated Inventory (to know what was available)
    - Cleaned Ratings (to know usage and link to revenue)

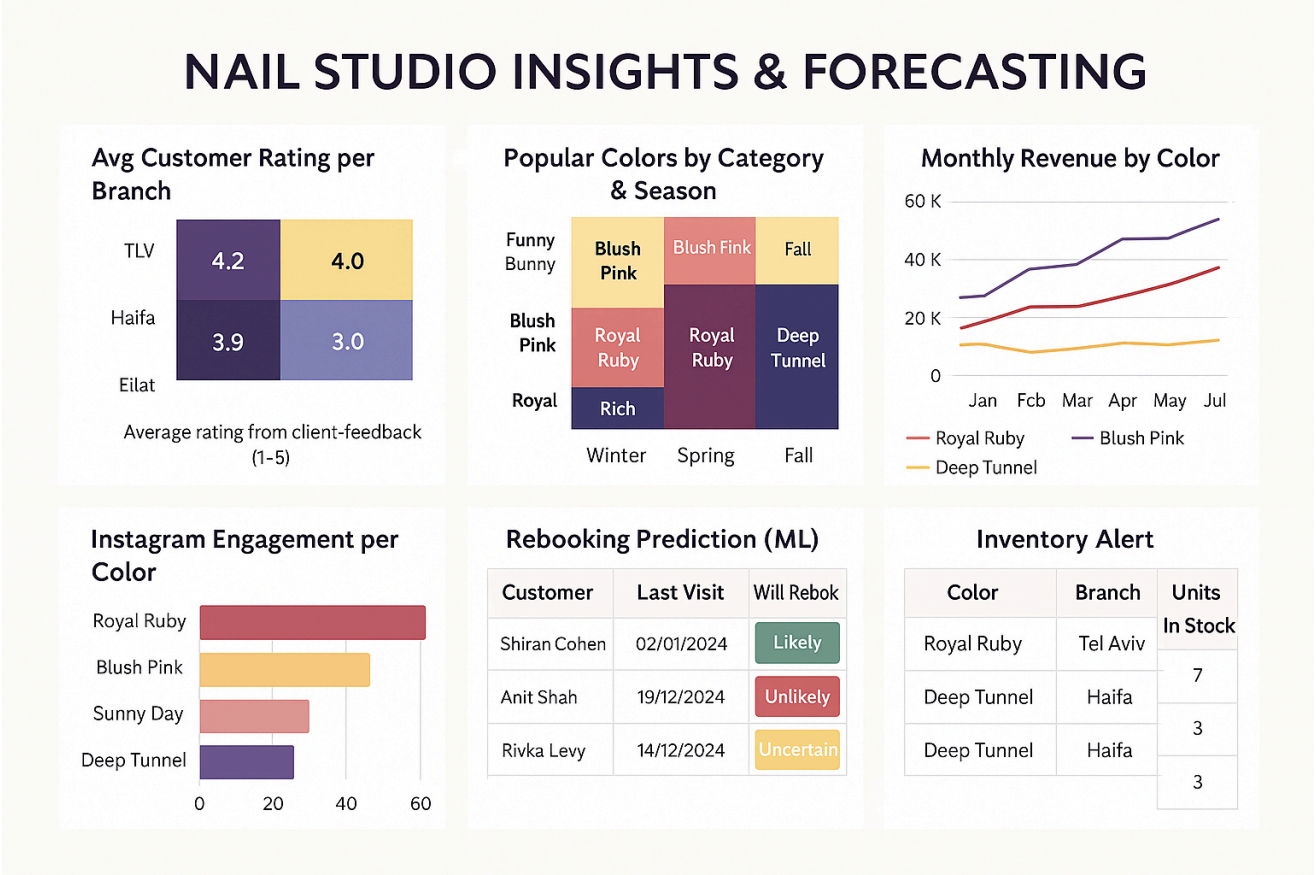
Logic:

- For each color:
    - Track its use over time
    - Join with session revenue (assumed from treatment cost or session-level revenue)
    - Identify top colors based on frequency and correlate with revenue over months

Slicers / Filters Panel

Based on Table	Filter
Branches Dimension Table	Branch
Derived from session timestamp	Season
Colors Dimension Table → Category	Color Category

Dashboard Data Sources & Table Design



Avg Customer Rating per Branch

- Fact Table: SESSIONS – contains rating, branch
- Dimension Table: BRANCHES – contains branch details

Popular Colors by Category & Season

- Fact Table: SESSIONS – tracks color and date
- Dimension Tables:
  - COLORS – for color, category
  - DATE_DIM – for season extraction

Monthly Revenue by Color

- Fact Table: SESSIONS – includes payment_amount, color_id, date
- Dimension Tables:
    - COLORS – for color labeling
    - DATE_DIM – to get month

Instagram Engagement per Color

- Fact Table: INSTAGRAM_FACT – likes, comments, total engagement
- Dimension Table: COLORS – to get color names

Rebooking Prediction (ML)

- Fact Table: ML_FEATURE_TABLE – features like last_visit, will_rebook
- Dimension Tables:
    - CUSTOMERS – customer name, email, SCD2 history
    - DATE_DIM – for working with visit dates

Inventory Alert

- Fact Table: INVENTORY_FACT – stock quantity per color_id and branch_id
- Dimension Tables:
    - COLORS – for color labeling
    - BRANCHES – for branch location

Table Summary

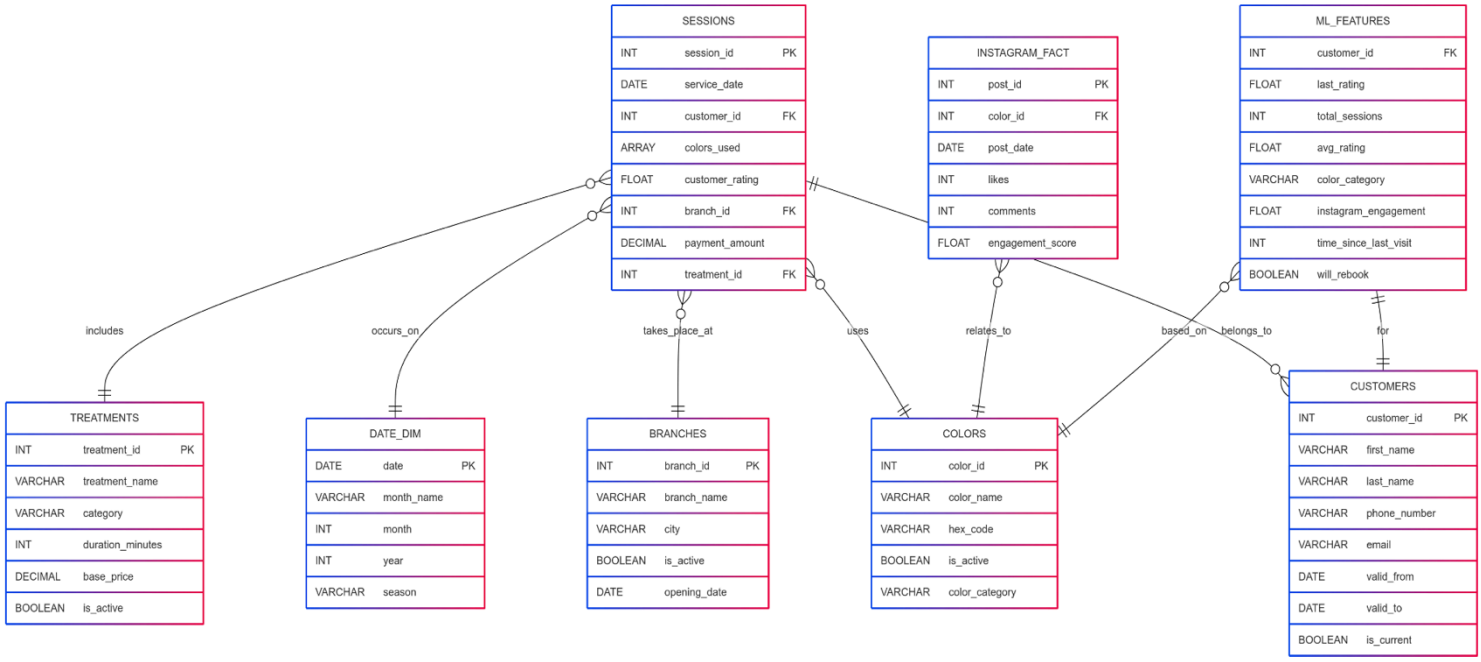
Fact Tables:

- SESSIONS
- INSTAGRAM_FACT
- ML_FEATURE_TABLE
- INVENTORY_FACT

Dimension Tables:

- COLORS
- BRANCHES
- CUSTOMERS (SCD Type 2)
- DATE_DIM

- Backtrack to design fact/dimension table(s) that will feed this dashboard



Design fact/dimension table :

Table: **SESSIONS**

- Used for: Analyzing popular colors, revenue trends, customer ratings, and connecting between entities (customers, colors, branches, dates, and treatments).

Table: **CUSTOMERS**

- Used for: Storing historical customer profile changes (SCD Type 2), tracking satisfaction over time, and supporting personalized prediction models (ML).

Table: **COLORS**

- Used for: Displaying color names and categories in reports, segmenting colors by type, and analyzing popular trends on social media.

Table: **BRANCHES**

- Geographic segmentation, analyzing revenue and ratings by branch, and managing operations based on branch status and availability.

Table: **DATE_DIM**

- Used for: Segmenting by months and seasons, calculating date intervals, and identifying time-based trends.

Table: **TREATMENTS**

- Used for: Analyzing revenue by treatment type, tracking treatment popularity, and pricing services in BI reports.

Table: **INSTAGRAM_FACT**

- Used for: Measuring engagement per color on social media, generating marketing insights, and analyzing campaign effectiveness.

Table: **ML_FEATURES**

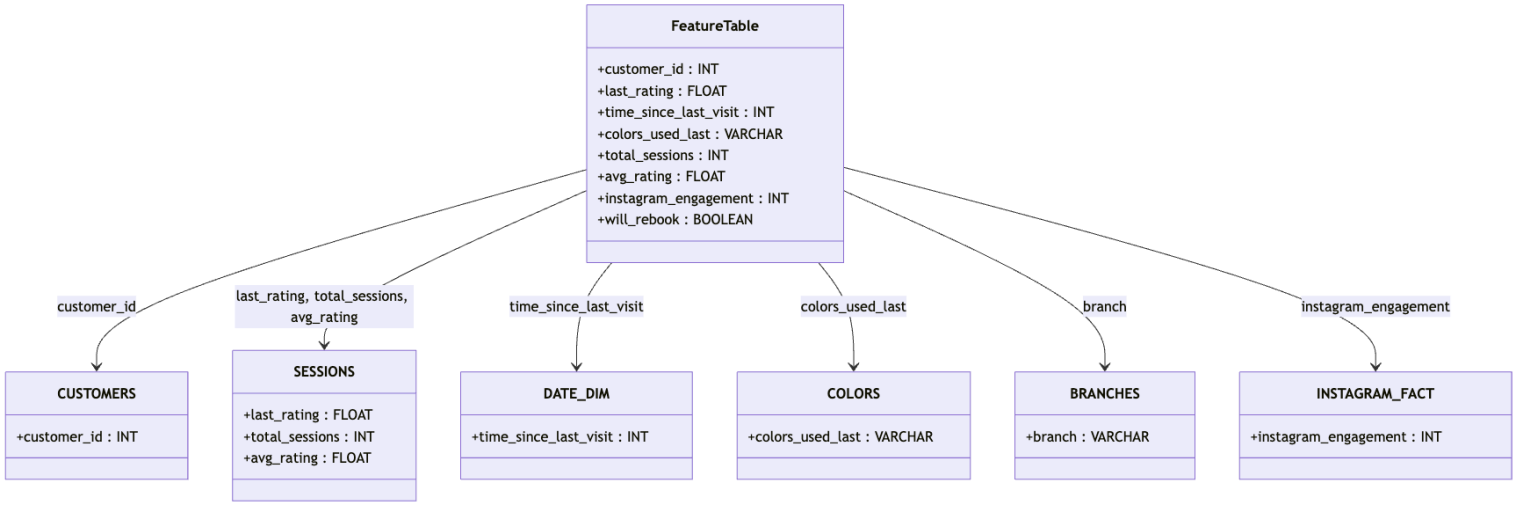
- Used for: A feature table used by the prediction model to assess the likelihood of customer rebooking. Includes ratings, interactions, colors, and timing.

2. Machine Learning Requirements
    - Business Problem suitable for ML

Key Business Problem Question – "Can we accurately predict whether a customer will schedule another appointment within 30 days of their last visit? "

This problem is highly relevant to customer relationship management, retention, and personalized marketing. Using the model, the system can proactively identify customers at risk of not returning and trigger reminders, promotions, or a follow-up service call.

- Design a dataset/feature table for this ML model



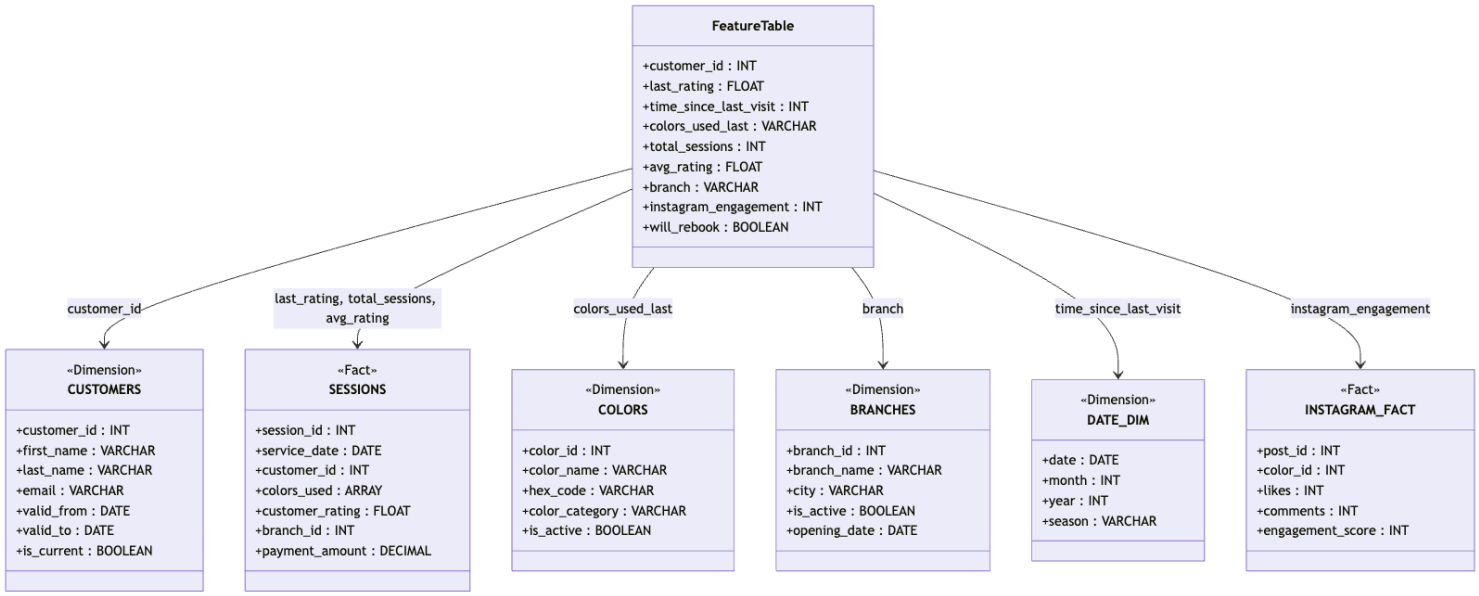
The diagram presents the structure of the Feature Table used by the Machine Learning model for predicting whether a customer will return for another appointment within 30 days.
The table includes key features such as customer_id, last rating (last_rating), time since the last visit (time_since_last_visit), color category (colors_used_last), total number of sessions, average rating, social media engagement (instagram_engagement), and most importantly – the target variable will_rebook.

Each feature in the table is derived from a different source within the system:

- CUSTOMERS – Customer identifier
- SESSIONS – Ratings and number of visits

- DATE_DIM – Time calculation between sessions
- COLORS – Colors used during the last session
- BRANCHES – Identification of the last branch visited
- INSTAGRAM_FACT – Interactions with social media posts

- Backtrack to design fact/dimension table(s) needed for this dataset



SESSIONS (Fact Table)
Contains all information related to service sessions:

- Service date, selected colors, customer rating, branch, payment amount
- Used to calculate features such as last_rating, total_sessions, and avg_rating

INSTAGRAM_FACT (Fact Table)
Stores data from Instagram marketing posts:

- Linked color, number of likes and comments, total engagement score
- Provides the instagram_engagement feature

CUSTOMERS (Dimension Table – SCD2)
Stores customer profile information:

- Identifier, name, email, validity range (SCD Type 2)
- Provides the customer_id and enables historical matching over time

COLORS
Contains color details:

- Color ID, name, category, active status
- Links colors_used_last to its corresponding category or visual characteristic

BRANCHES
Branch metadata:

- Branch ID, name, city, and active status
- Used for the branch feature

DATE_DIM
Calendar dimension table:

- Date, month, year, season
- Enables calculation of time_since_last_visit and supports seasonal analysis if needed