Shopware Enterprise Data Engineering Project

---

1. Project Overview

Shopware is a retail company with a growing need to consolidate and process data from multiple sources, both **batch** and **streaming**, to enable more efficient decision-making across various teams. This project involves building an **enterprise-level data pipeline** to transform and integrate raw data from 4 different sources into a streamlined, easy-to-access format for various business teams at Shopware. The final goal is to enable each team to leverage **key performance indicators (KPIs)** and actionable insights to guide decision-making.

Project Objectives

1. **Data Integration**: Integrate data from multiple sources, including both **batch** and **streaming data**.
2. **Data Transformation**: Clean and transform raw data to meet the business needs.
3. **Data Accessibility**: Provide different teams with access to data via **ad-hoc querying**, **dashboards**, or **data marts**.
4. **Data Storage**: Organize the data efficiently using **data lakes**, **data warehouses**, and **data marts**.
5. **KPI Tracking**: Provide teams with the ability to track KPIs relevant to their department's needs.

---

3. Data Sources

The following 10 data sources will be integrated into the project. These will include both batch data (periodic updates) and streaming data (real-time or near real-time updates).

Data Source	Type	Frequency	Description	Consumed By
POS Data	Batch	Daily	Sales transactions, including quantity, revenue, discount, etc	Sales, Operations, Finance
Inventory Management	Batch	Hourly	Real-time inventory levels and restocking data	Operations, Sales
Customer Interactions (CRM)	Streaming	Real-Time	Customer messages, loyalty events, support feedback	Marketing, Support
Web Traffic Logs	Streaming	Real-Time	Clicks, views, sessions, and engagement behavior	Marketing, Data Analysts

---

4. Team Responsibilities and Data Access

Sales Team

- **Data Needed**: POS, Inventory
- **KPIs**:
    - Total Sales by region/product
    - Stock Availability
    - Product Turnover Rate
- **Access**: Sales Data Mart

Marketing Team

- **Data Needed**: Web Traffic Logs, CRM Interactions
- **KPIs**:
    - Customer Engagement Score
    - Session Duration / Bounce Rate
    - Loyalty Activity Rate
- **Access**: Dashboard

Operations Team

- **Data Needed**: Inventory, POS
- **KPIs**:
    - Inventory Turnover
    - Restock Frequency
    - Stockout Alerts
- **Access**: Data Warehouse/ Lake house (Ad-hoc SQL access)

Customer Support

- **Data Needed**: CRM Interactions
- **KPIs**:
    - Feedback Score
    - Interaction Volume by Type
    - Time-to-Resolution
- **Access**: Dashboards + Alerting Systems

---

6. **Data Flow and Processing Pipeline**

1. **Data Ingestion**:
    - **Batch Data**: Extracted at regular intervals (e.g., daily, weekly) from systems like POS, ERP, and Customer Demographics.
    - **Streaming Data**: Continuously ingested from sources like web traffic, social media, and customer interactions.
2. **ETL (Extract, Transform, Load)**:
    - **Extract**: Data is pulled from multiple source systems.
    - **Transform**: Data is cleaned, validated, aggregated, and enriched as needed.
    - **Load**: Cleaned data is loaded into:
        - **Data Lake** for raw, unstructured data.
        - **Data Warehouse** for refined data.
        - **Data Marts** for team-specific, aggregated data.
3. **Data Storage**:
    - **Data Lake**: Raw data from all sources is stored for historical purposes and can be accessed for specific needs.
    - **Data Warehouse**: Clean, transformed data for business-wide queries and reporting.
    - **Data Marts**: Aggregated, team-specific data (e.g., sales, finance, marketing).
4. **Data Access**:
    - **Ad-hoc Queries**: Teams like Data Analysts and Operations have direct access to the Data Warehouse for customized querying.
    - **Dashboards**: Marketing and Sales teams will use predefined visualizations for key metrics like campaign performance, conversion rates, etc.
    - **Data Marts**: Sales, Marketing, and Finance teams have direct access to pre-aggregated KPIs relevant to their respective roles.

Security and Compliance

1. **Role-Based Access Control (RBAC)**:
    - Ensure that teams can only access data relevant to their function. For instance, the Sales team should only have access to sales-related data, not sensitive financial records.
2. **Data Encryption**:

- Ensure all data is encrypted both at rest and in transit to comply with industry standards and protect sensitive customer data.

3. **Compliance**:
    - Ensure that data handling is compliant with **GDPR**, **CCPA**, and other relevant regulations for data protection.

Monitoring and Logging

- **Pipeline Health Monitoring**: Automatically monitor the health of data pipelines. Alerts should be triggered for any failed processes or delays in data updates.
- **Data Quality Monitoring**: Validate data for completeness and correctness before it is loaded into data marts or data warehouses.
- **Audit Logging**: Maintain a comprehensive log of data access and transformations for compliance and troubleshooting purposes.

---

8. Additional Info: Schema for All 4 Data Sources

1. POS Data (Batch – Daily)

Field	Type	Nullable	Notes
transaction_id	String	No	Unique transaction ID
store_id	Integer	No	Store where purchase occurred
product_id	Integer	No	Product sold
quantity	Integer	No	Units sold
revenue	Float	No	Gross amount
discount_applied	Float	Yes	Discount applied
timestamp	Float (epoch)	No	Transaction time

---

2. Inventory Management Data (Batch – Hourly)

Field	Type	Nullable	Notes
inventory_id	Integer	No	Unique record ID
product_id	Integer	No	Product identifier
warehouse_id	Integer	No	Warehouse location

Field	Type	Nullable	Notes
stock_level	Integer	No	Current stock count
restock_threshold	Integer	Yes	Minimum level before restock
last_updated	Float (epoch)	No	Timestamp of last update

---

3. Customer Interactions (Streaming)

Field	Type	Nullable	Notes
customer_id	Integer	No	Unique ID
interaction_type	String	No	Feedback, Loyalty, Complaint
channel	String	Yes	Email, App, Phone
rating	Integer (1–5)	Yes	Optional satisfaction rating
message_excerpt	String	Yes	Excerpt from interaction
timestamp	Float (epoch)	No	Time of interaction

---

4. Web Traffic Logs (Streaming)

Field	Type	Nullable	Notes
session_id	String	No	Unique session ID
user_id	Integer	Yes	Nullable for anonymous users
page	String	No	Page visited
device_type	String	Yes	Desktop, Mobile, Tablet
browser	String	Yes	Chrome, Safari, etc.
event_type	String	Yes	Click, View, Scroll
timestamp	Float (epoch)	No	Event time