Assignment: Data Flow, ETL Pipeline & Reporting for Stream's B2B Operations

Company Context:

Stream is a B2B company that manages customer data, sales processes, and marketing efforts through various systems. Stream's in-house system, **Nessy**, acts as the company's backend database and source of truth for customer details, including membership, plans, and organizational data. Nessy integrates with other systems to support marketing, sales, and subscription management efforts.

Here's a breakdown of Stream's tools and systems:

- **CRM & Sales:** Salesforce, HubSpot, Nessy (in-house tool)
- **Marketing & Analytics:** Google Analytics, Google Ads, GitHub
- **Data Warehousing & Integration:** BigQuery (data warehouse), Fivetran (data integration), Zapier (automation)
- **Reporting & Lead Enrichment:** Looker (reporting), Qualified (lead enrichment and scoring)
- **Subscription Management:** Stripe (connected directly to Nessy for managing subscriptions)

Key Entities in Stream's B2B Operations:

- **Leads:** Prospective clients from marketing efforts (e.g., Google Ads) and website activity tracked by Google Analytics. Leads are enriched with tools like Qualified and MadKudu before entering the CRM.
- **Contacts:** Once a lead is qualified, they're converted into a contact in the CRM (e.g., Salesforce, HubSpot). Multiple contacts can be linked to a single opportunity, representing different stakeholders in the sales process.
- **Opportunities:** These are potential deals in the CRM, with multiple contacts connected. Opportunities move through various sales stages.
- **Organizations:** When a user signs up via Nessy, an organization is created. Members (users) can belong to one or more organizations.
- **Plans:** New organizations are assigned a default free legacy plan unless another is selected. Plan data is stored in Nessy and tied to the organization.
- **Campaigns:** Marketing efforts (e.g., Google Ads) are tracked in the CRM and linked to leads, contacts, and opportunities.
- **Subscription Records:** Managed through Stripe, subscription details for organizations (e.g., active plans, overages) are stored in Nessy and linked back to the CRM via HubSpot.

Assignment Overview

You will complete three main tasks:

1. **Create a Data Flow Diagram (DFD)** to represent how data moves through Stream's systems.
2. **Develop an ETL Pipeline** to process and transform a set of provided datasets.
3. **Generate a Marketing Performance Report** using Streamlit based on the processed data.

Task 1: Data Flow Diagram (DFD)

Create a DFD that represents how data moves through various systems at Stream, focusing on these key processes:

- **Lead Generation and Enrichment:** Show how leads are generated from marketing efforts (Google Ads, LinkedIn, GitHub, etc.) and enriched using Qualified and MadKudu before being passed to the CRM.
- **Trial Conversion Process:** Highlight how a lead might initially sign up for a trial with a personal or throwaway email, then create a second trial with a corporate email, transitioning into an opportunity in Salesforce.
- **Nessy's Role:** Show how Nessy acts as the source of truth for customer and plan data, integrating with HubSpot to feed CRM information into Salesforce. Nessy also manages subscriptions and connects directly to Stripe for billing and plan data.
- **Data Storage and Reporting:** Detail how CRM data (Salesforce, HubSpot, Nessy) is stored in BigQuery via Fivetran, and how it is utilized in Looker for reporting.
- **Automation:** Include how Zapier automates tasks such as syncing data between tools (e.g., Salesforce, HubSpot).

Task 2: Build an ETL Pipeline

You are provided with the following datasets:

Download Dataset HERE

- **transactions.csv** — Contains transaction data (e.g., transaction id, customer id, subscription it, total transaction amount, currency, and date of transaction).
- **users.csv** — Contains user information (e.g., customer ID, customer name, customer email)

- **products.csv** — Contains subscription plan and product details (e.g., subscription ID, plan ID, product plan plan, interval of billing, contract amount, and billing status)

Instructions:

1. Data Ingestion:
   Write a Python script to ingest data from the provided CSV files.
2. Data Transformation:
   - Standardize date formats.
   - Handle missing or duplicate records.
   - Create a relational schema by establishing primary and foreign keys.
   - Calculate total spending per user.
3. Data Loading:
   Load the transformed data into a relational database (e.g., PostgreSQL, MySQL).
4. Optimization and Analysis:
   Write three SQL queries to answer:
   - Top 5 users by total transaction amount.
   - Product category with the highest total sales.
   - Monthly revenue growth for the last 6 months.

Task 3: Generate a Performance Report

After building the ETL pipeline and loading the data, you will create a report that provides insights into Stream's performance using the datasets from Task #2: users.csv, transactions.csv, and products.csv.

Instructions:

1. **Ingest and Transform:**
   - Ensure you have ingested the users.csv, transactions.csv, and products.csv datasets into your ETL pipeline.
   - Clean and transform the data as necessary, ensuring that user_id links the users to their transactions and products.
2. **Performance Report:** Using Streamlit, create a report that visually answers the following questions:
   - Average User Transaction Amount for the Last 6 Months
   - Product Category with the Highest Total Sales
   - Monthly Revenue Growth for the Last 6 Months

Submission Guidelines:

- **DFD Diagram:** Submit an image or PDF of your DFD diagrams.
- **Codebase:** Submit a well-documented codebase for your ETL pipeline, with clear instructions on how to run it.
- **Marketing Performance Report:** Submit a Streamlit-generated report that visualizes key marketing performance metrics and insights.

Estimated Time:
This assessment is designed to take approximately 2-3 hours, with a focus on clear representation of data flow, development of the ETL pipeline, and generating actionable business reports.

Evaluation Criteria:

- **Clarity:** Diagrams, reports, and code should be easy to follow, well-organized, and clearly documented.
- **Detail:** Provide enough detail in your diagrams, reports, and code to demonstrate a strong understanding of data flow, ETL processes, and marketing performance reporting for a B2B company.
- **Technical Expertise:** Show proficiency in handling B2B data flows, transforming and optimizing datasets, and delivering meaningful insights through reporting tools.