Customer Data Platform(CDP) for Enhanced Customer Segmentation

ABC Company is experiencing rapid growth in its digital channels and needs to better understand its customers across multiple touchpoints. Currently, customer data is siloed across various systems, making it difficult to create a unified view of each customer and effectively segment the audience for marketing purposes. The company requires a **Customer Data Platform (CDP)** to collect, unify, and activate customer data from various sources with the following goals:

- Create a single source of truth for customer profiles
- Enable effective customer segmentation for targeted marketing
- Provide easy access to customer data and segments via APIs

Goals

1. **Collect user information**
    - Use a cookie as the primary key if an email is not present. Otherwise, use the email as the primary key.
2. **User identification and merging**
    - When a user visits the website, a unique cookie is assigned via the website.
    - On registration or login, the system captures the user's email.
    - If the email is not already present, update the user profile by associating the email with the existing cookie.
    - If the email is already linked to a different cookie, merge both records into a single user profile.
3. **User segmentation**
    - Classify users into segments based on their interests.
    - Interests can include categories such as sports, technology, finance, entertainment, fashion, etc.
    - Example: If a user is interested in "Tom Brady," they belong to the **Sports Cohort** rather than a movie segment.
    - Example: If a user is interested in "Nike shoes," they belong to the **Fashion Cohort**
4. **Develop a scalable system**
    - Implement an ETL pipeline to process and transform raw data.
    - Enable real-time data collection.

    - Store user data in different stages:
        - Raw Data Storage: Maintain all collected user data in its original format.
        - Structured Data Storage: Store processed data after passing through the ETL pipeline.
        - Segmented Data Storage: Maintain information about user and their cohort assignments.

Requirements:

1. User Data Collection

User Profile Schema (Sample Data)

Mandatory Fields:

- Cookie (unique identifier)

Optional Fields:

- Email
- Phone Number
- Location (State, Country, City)
- Demographics (Age, Gender, Income, Education)
- Interests (array/list)

User Data Management Rules

- Use **cookie** as the primary identifier for creating user profiles.
- If a **cookie** is already present, update the existing user information.
- If a **cookie** is not present, create a new user profile.
- If the same **email** is found with different cookies, merge the user profiles and maintain **email** as the primary key.

2. Data Collection

Raw Data Storage (For new Entries)

- Store all incoming new data in its original format.
- Maintain timestamps and sources.

Example:

```
{
  "data": {
    "cookie": "dk5h2p9s7f3j8m1n4v0x",
    "email": "john.smith@example.com",
    "phone_number": "+17325559876",
    "location": {
      "state": "New York",
      "country": "USA",
      "city": "Brooklyn"
    },
    "demographics": {
      "age": 42,
      "gender": "Male",
      "income": "$100,000-$149,999",
      "education": "Bachelor's Degree"
    },
    "interests": ["cricket", "football", "tom brady"]
  }
}
```

Data Structuring

- Transform raw data into a structured format for analysis.
- Store data in either:
  - **SQL** (for relational data with a defined schema)
  - **NoSQL** (for flexible schema and high-volume data)

User Segmentation

- Process new data to identify user interests and behaviors.
- Segment users into cohorts based on identified interests. (Should develop a ml based algo for classification based on user interests)
- Implement **dynamic segmentation algorithms** that update user cohorts as new data arrives.
- Store segment/cohort assignments with user profiles.

Example Cohorts:

- **Sports Cohort** (Users interested in Cricket, Football, Basketball, etc.)
- **Technology Cohort** (Users interested in AI, Programming, Gadgets, etc.)

- **Finance Cohort** (Users interested in Investment, Banking, Crypto, etc.)

3. API Endpoints

User Details Endpoint

Endpoint: `/api/user` **Method:** `GET`

Parameters:

- `cookie` or `email` (for lookup)

Response:

```
{
  "user_profile": {
    "cookie": "dk5h2p9s7f3j8m1n4v0x",
    "email": "john.smith@example.com",
    "location": "Brooklyn, New York, USA",
    "demographics": {
      "age": 42,
      "gender": "Male"
    },
    "interests": ["crcket", "ai", "football"],
    "cohorts": ["Sports", "Technology"]
  }
}
```

Data Ingestion Endpoint

Endpoint: `/api/ingest` **Method:** `POST`

Parameters:

- JSON or CSV payload with user data

Response:

```
{
  "status": "success",
  "records_processed": 500,
  "errors": []
}
```

Similar Users Endpoint

Endpoint: `/api/cohort/users` **Method:** `GET`

Parameters:

- `cohort` (required, e.g., "sports", "technology", "finance")
- `limit` (optional)
- `offset` (optional)

Response:

```
{
  "cohort": "Sports",
  "users": [
    {
      "email": "john.smith@example.com",
      "similarity_score": 0.95
    },
    {
      "email": "jane.doe@example.com",
      "similarity_score": 0.89
    }
  ]
}
```

4. Technical Considerations

Performance

- High-throughput data ingestion.
- Scalable data storage.
- Optimized query performance for real-time segmentation.

Integration

- Batch processing capabilities for large datasets.
- ETL pipeline for data transformation and loading.

Analytics & Segmentation

- ML models for segmentation.
- Rules engine for business-defined segmentation.
- Real-time classification services.

5. Project Deliverables

- **Comprehensive documentation** covering:
  - User data ingestion.
  - Cohort query API endpoints (with examples).
  - System architecture.
  - API specifications.
- **ETL**: Data transformation processes with batch import and quality validation.
- **VIDEO DEMO**: 5-minute walkthrough showcasing system functionality, API usage, data ingestion, and segmentation.
- **GitHub repo** with README, setup instructions, and codebase.
- **Demo URL** with test credentials and example API calls.