**Health Condition Registry ETL Documentation and Guide**

**Objective:**The objective of this guide is to provide a comprehensive framework for integrating health condition registry data from various sources, ensuring consistency and accuracy through a well-defined ETL (Extract, Transform, Load) process. This will facilitate effective data management and improve the quality of patient information for analysis and reporting.

**1. Analyze and Document Dataset Structures**

* **Identify Common Fields:** Review and compare fields across different datasets to determine the common elements shared among them.
* **Identify Unique Fields:** Highlight fields that are present in some datasets but missing in others.
* **Standardize Field Names:** Align all field names to a standard set based on the master schema for consistency across datasets.

**Master schema:**

|  |  |  |
| --- | --- | --- |
| **Order** | **Field name** | **Type** |
| 1 | FullName | String |
| 2 | FirstName | String |
| 3 | LastName | String |
| 4 | BirthDate | Date |
| 5 | Gender | String |
| 6 | Condition | String |
| 7 | Treatment | String |
| 8 | AdmissionDate | Date |
| 9 | DischargeDate | Date |
| 10 | Address | String |
| 11 | State | Character (2) |
| 12 | Zipcode | Character (5) |
| 13 | Phone | String |
| 14 | Email | String |

**2. Implementing the ETL pipeline**

**2.1 Extract**

* **Merging Datasets:** Combine datasets from various sources to form a unified data set.
* **Mapping Fields:** Align the field names in each dataset to the corresponding fields in the master schema.

**2.2 Transform**

* **Verifying Data Integrity:** Ensure that the merged dataset accurately reflects the information from all source datasets.
* **Handling Missing Data:** Define a strategy to manage missing values (e.g., leave blank, use placeholders, or impute missing data).
* **Standardizing Field Formats:** Set consistent formats for fields such as:
* **Date format**: Use YYYY-MM-DD.
* **Phone number format**: Standardize phone numbers.
* **Address Validation:** Split the full address into components: address, state, and ZIP code.
* **Data Cleaning**: Detect and fix any inconsistencies or errors, such as:
* Incorrect or incomplete phone numbers.
* Misformatted dates.
* **Checking for Duplicates:** Identify and handle duplicate records based on unique identifiers (e.g., email addresses or phone numbers).
* **Checking for Consistency:** Ensure that all entries conform to the defined field formats and standards.

**2.3 Load**

* **Normalizing:** Creating the foreign keys in DataFrame for the Patient table in the database to ensure references to Condition, Treatment and Gender tables.
* **Record Verification:** check if the record already exists by comparing both the phone number and email with the data in the database.
* **Duplicate Prevention:** If a record with the same phone number or email is found, the insertion process is skipped to avoid creating duplicate entries.
* **Data Insertion:** If no matching record is identified, the new data is inserted into the database, ensuring that only unique records are added.

**3. Technical Tools:**

* **Python:** Used to implement and automate the ETL pipeline.
* **SQL Server:** Serves as the database management system to store and manage the processed data.

**Expected Outcomes:**

* **A fully automated ETL pipeline** that extracts, transforms, and loads health condition registry data on a scheduled basis, ensuring the database is continuously updated with accurate, consistent patient information.
* **A well-structured relational database** in SQL Server with a normalized schema, containing clean and consistent datasets that can be easily queried for insights into health conditions, treatments, and patient outcomes.
* **Comprehensive documentation** detailing the ETL process, schema design, field mapping, and usage instructions, ensuring that team members can effectively manage and maintain the system.
* **A system for data integrity and duplicate prevention**, including automated checks for missing, inconsistent, or duplicate data, ensuring high-quality, reliable records are maintained in the database.

**Benefits:**

* Enhanced data quality and consistency through standardized field names, formats, and validation, ensuring reliable patient information.
* Increased operational efficiency with an automated ETL pipeline, reducing manual efforts and ensuring timely data updates.
* Prevention of duplicate records by checking unique identifiers (e.g., phone number, email), ensuring accurate and unique patient records.
* Improved decision-making and reporting by providing clean, structured data that supports insightful analysis of health conditions, treatments, and outcomes.