**Event-Driven Data Pipeline for Healthcare Data in GCP BigQuery**

**Objective**

Design and implement an event-driven data pipeline in Google Cloud Platform (GCP) for processing healthcare data with a focus on cancer patients.

* **Requirements**
* **Data Source and Format:**
  + Daily receipt of clinical data as JSON files in a GCP bucket.
  + File Naming Convention: clinical\_data\_yyyy\_mm\_dd.json.
* **Event-Driven Ingestion:**
  + Automate data ingestion upon file arrival in the GCP bucket.
* **Target Storage:**
  + Data is to be loaded into a BigQuery table (You can create schema design as per your need and assumptions)
* **Data Filtering:**
  + Ingest only data related to cancer patients.
* **Data Validation:**
  + Ensure patient identifiers (like patient ID) are present and unique.
  + Validate date fields (e.g., date of diagnosis) for format correctness.
  + Check mandatory fields such as patient ID, diagnosis code, and hospital ID.
  + Confirm diagnosis codes are specific to cancer.
  + Verify data types and formats for all fields (e.g., numeric values, text strings).
  + Exclude records with incomplete or inconsistent data.
* **Sample Data File:**
  + A sample JSON file representing clinical data has been provided, refer that
  + Write script to generate 3-4 mock data files with random values to implement and test your logic
* **Hint - Suggested GCP Services:**
  + GCP Bucket
  + GCP Functions
  + Dataflow
  + BigQuery
* **Deliverables**
  + Code, Data files, Configurations, Metadata files
  + Documentation explaining the setup and complete execution of the application.