**Customers Table**

* **customer\_id**  
  A unique identifier for each customer. This numeric key is used to link customer records to policies, claims, and communications.
* **first\_name**  
  The customer's given name. It helps personalize interactions and serves for identification purposes in reports and dashboards.
* **last\_name**  
  The customer's family name. Used together with the first name for complete identification and record-keeping.
* **age**  
  The customer's age at the time of data generation. Useful for segmentation, risk analysis, and demographic reporting.
* **gender**  
  The customer’s gender (e.g., 'M' or 'F'). This can be leveraged for demographic breakdowns and trend analysis.
* **region**  
  The geographic location where the customer resides. In our case, it represents a county in Kenya—with a higher probability for major urban counties (Nairobi, Mombasa, Kisumu, Eldoret, Nakuru) and a smaller chance for other counties. This field is key for location-based segmentation and regional analysis.
* **vip\_status**  
  A boolean flag indicating if the customer is considered a VIP. This could affect prioritization in service, eligibility for premium services, or targeted marketing strategies.

**Policies Table**

* **policy\_id**  
  A unique identifier for each insurance policy. This ID differentiates individual policies and is crucial for linking policy records with claims and communications.
* **customer\_id**  
  The identifier of the customer who owns the policy. It connects the policy back to the corresponding customer in the Customers table.
* **policy\_type**  
  The type of insurance policy (for example, Health, Motor, or Life). This column helps categorize policies and tailor analysis and reporting to specific segments.
* **premium**  
  The amount that the customer pays for the policy (could be monthly, yearly, etc.). It’s important for revenue calculations and risk assessments.
* **sum\_insured**  
  The total coverage amount provided by the policy. It indicates the maximum financial liability of the insurance provider in case of a claim and is used in premium and risk calculations.
* **start\_date**  
  The effective date when the policy begins. This helps in tracking policy durations, renewal dates, and eligibility periods for claims.
* **end\_date**  
  The date when the policy expires. Together with the start\_date, it defines the policy’s active period and can be used to trigger renewals or expirations in analyses.

**Claims Table**

* **claim\_id**  
  A unique identifier for each claim record. It distinguishes each claim filed against a policy.
* **policy\_id**  
  The identifier of the policy associated with the claim. This links a claim directly to the policy it’s filed under.
* **customer\_id**  
  The identifier of the customer who filed the claim. This direct linkage helps streamline customer-level analyses without always having to traverse through the policy.
* **claim\_date**  
  The date on which the claim was filed. It’s used to track claim trends over time and assess policy performance.
* **claim\_amount**  
  The monetary value being claimed. It is a key metric for assessing the financial impact of claims and managing risk.
* **claim\_status**  
  The current status of the claim, such as Approved, Rejected, or Pending. This column helps track the progress and resolution of claims, which is critical for operational and customer service performance.

**Communications Table**

* **communication\_id**  
  A unique identifier for each communication record. This field is used to track individual interactions between the customer and the insurer.
* **customer\_id**  
  The identifier of the customer who initiated the communication. It links the communication record back to the customer, enabling a comprehensive view of all customer interactions.
* **timestamp**  
  The date and time when the communication occurred. This is vital for time-series analysis, measuring response times, and tracking communication history.
* **channel**  
  The medium through which the communication took place (e.g., Phone, Email, In-Person, Chatbot, Social Media). This helps in understanding and analyzing customer preferences for communication channels.
* **reason\_category**  
  The broad category that defines the purpose of the communication. Examples include:
  + **Complaint**: Issues or dissatisfaction expressed by the customer.
  + **Feedback**: Suggestions, compliments, or general comments about service or products.
  + **Call Log**: Interactions that mimic the conversational or transactional nature of a phone call—even if conducted via a Chatbot, for instance.

Using “Call Log” for a Chatbot interaction recognizes that many chatbot exchanges resemble traditional call logs in that they are quick, transactional interactions.

* **message\_text**  
  A short descriptive text capturing the essence of the communication. It might include details like the nature of the complaint or the subject of the inquiry, which could be later used for sentiment analysis or natural language processing.
* **sentiment**  
  A basic sentiment indicator (e.g., Positive, Neutral, Negative) assigned to the message\_text. Although generated randomly for our synthetic dataset, this field would ideally come from an NLP sentiment analysis in a production environment.
* **outcome**  
  The result or resolution of the communication (such as Resolved, Escalated, Pending Info, or Acknowledged). This indicates how the issue was addressed and helps measure customer satisfaction and service efficiency.
* **related\_claim\_id**  
  An optional field that links the communication to a specific claim, if applicable. For example, if a customer files a complaint or inquiry that directly references a claim, this column captures that relationship. It’s particularly useful in scenarios where a communication is intended to follow up on a claim, ensuring all related interactions can be easily joined and analyzed.