Data Vault Model Task:

1. What technology/technologies will be used to implement this storage solution?

Answer: Data storage solution, I recommend a cloud-based data warehouse like Google BigQuery as these data warehouses can handle large volumes of data for an event-driven architecture.

Data Ingestion: Apache Kafka/ Pub-sub for streaming events, Google Dataflow, Cloud composer Airflow for ETL pipelines.

2. Describe the table structure, attribute composition, and data types. The format of the description is open-ended; use whichever is most convenient or familiar for you.

Answer: Data Vault 2.0 Model involves Hubs, Links, and Satellites.

- Hubs: Represent core business entities.
- Links: Represent relationships between Hubs.
- Satellites: Store descriptive attributes about Hubs and Links.

Hubs

1. Hub User
    - user_hash_key (PK, VARCHAR)
    - user_id varchar (VARCHAR) from payload.uid
    - load_date (TIMESTAMP) Record ingestion time.
    - record_source (VARCHAR) Source system ("auth_event").
2. Hub Application
    - app_hash_key (PK, VARCHAR)
    - app_id (VARCHAR) from payload.app
    - load_date, record_source.

Links

Relationships between hubs. Each event type is a separate link.

1. Link Auth Event
    - auth_event_hash_key (PK, VARCHAR) msg_id.
    - user_hash_key (FK, VARCHAR)
    - app_hash_key (FK, VARCHAR).
    - load_date, record_source.
2. Link Spin Event
    - spin_event_hash_key (PK, VARCHAR) msg_id.
    - user_hash_key (FK, VARCHAR)
    - app_hash_key (FK, VARCHAR).
    - load_date, record_source.
3. Link Purchase Event
    - purchase_event_hash_key (PK, VARCHAR) msg_id.
    - user_hash_key (FK, VARCHAR)
    - app_hash_key (FK, VARCHAR).
    - load_date, record_source.

Satellites

1. Satellite User
    - user_uid (PK, VARCHAR)
    - user_hash_key (FK, VARCHAR), load_date (Timestamp).
    - email (VARCHAR), phone (VARCHAR): PII data from events
    - record_source

2. Satellite Auth Event
    - auth_sat_hash_key (PK, VARCHAR)
    - auth_event_hash_key (FK), load_date (Timestamp).
    - publish_ts (TIMESTAMP) from publish_ts.
    - email, phone
3. Satellite Spin Event
    - spin_sat_hash_key (PK, VARCHAR)
    - spin_event_hash_key (FK, VARCHAR), load_date (Timestamp)
    - publish_ts, spin_amount (INT)
4. Satellite Purchase Event
    - pur_sat_hash_key(PK, VARCHAR)
    - purchase_event_hash_key (FK), load_date(Timestamp)
    - publish_ts, amount (INT).

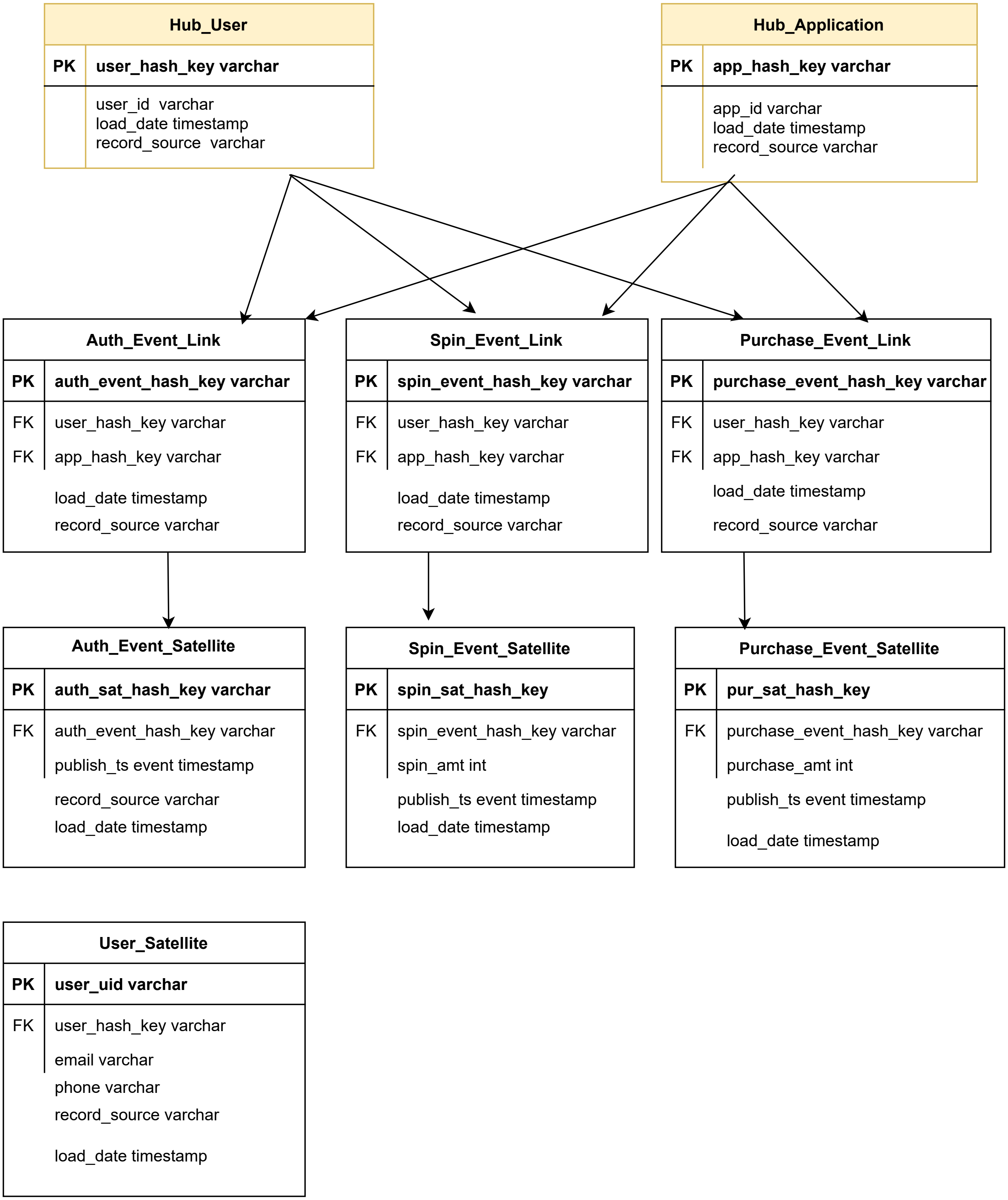
3.What additional components need to be developed to support your solution?

Answer:

- ETL Pipeline:

  Pipeline will extract data from the internal data bus transform it into the Data Vault format and load it into the data warehouse. Tools like Apache Kafka, Pub sub, Google Cloud Dataflow, Cloud composer Airflow for Orchestration can be used.

- PII Management: Encrypt sensitive fields
- Handle Late-Arriving Data: Windowing in BigQuery
- Data Quality Checks: Implement checks for null values and datatype validations to ensure data accuracy and consistency. Example: Great expectations



Hub_User

PK	**user_hash_key varchar**
	user_id varchar load_date timestamp record_source varchar

Hub_Application

PK	**app_hash_key varchar**
	app_id varchar load_date timestamp record_source varchar

Auth_Event_Link

PK	**auth_event_hash_key varchar**
FK	user_hash_key varchar
FK	app_hash_key varchar
	load_date timestamp
	record_source varchar

Spin_Event_Link

PK	**spin_event_hash_key varchar**
FK	user_hash_key varchar
FK	app_hash_key varchar
	load_date timestamp
	record_source varchar

Purchase_Event_Link

PK	**purchase_event_hash_key varchar**
FK	user_hash_key varchar
FK	app_hash_key varchar
	load_date timestamp
	record_source varchar

Auth_Event_Satellite

PK	**auth_sat_hash_key varchar**
FK	auth_event_hash_key varchar
	publish_ts event timestamp
	record_source varchar
	load_date timestamp

Spin_Event_Satellite

PK	**spin_sat_hash_key**
FK	spin_event_hash_key varchar
	spin_amt int
	publish_ts event timestamp
	load_date timestamp

Purchase_Event_Satellite

PK	**pur_sat_hash_key**
FK	purchase_event_hash_key varchar
	purchase_amt int
	publish_ts event timestamp
	load_date timestamp

User_Satellite

PK	**user_uid varchar**
FK	user_hash_key varchar
	email varchar
	phone varchar
	record_source varchar
	load_date timestamp