Steps for Data Processing Workflow

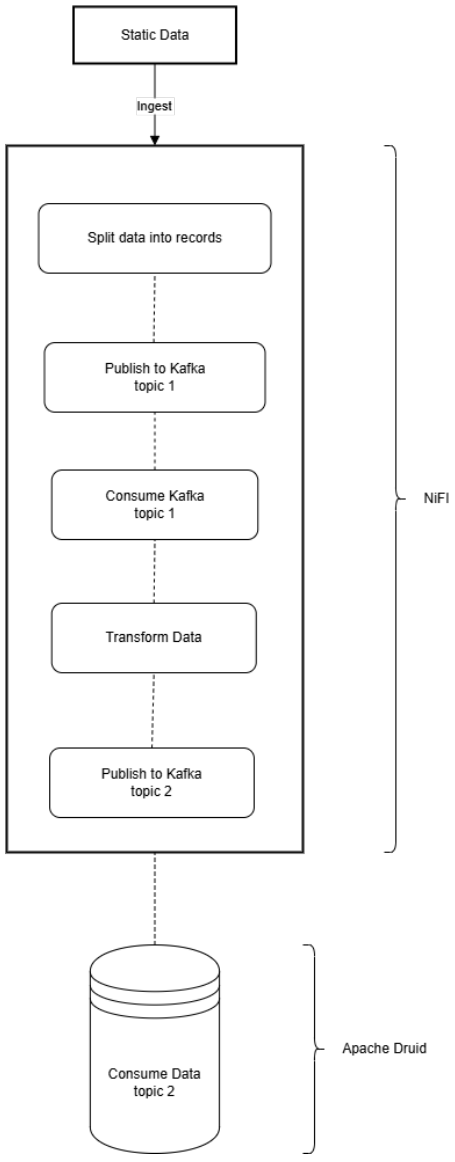

Figure 1: ETL Workflow Diagram

1. **Initial Ingestion of Static Data:**
   - NiFi ingests `tracking_data_x.xlsx` and `tracking_data_y.json` as static files.

- NiFi can read these files from the directory `data_input/incoming`.
- The xlsx file will be converted to json in Nifi

2. **Record-by-Record Processing with NiFi:**
   - After loading the entire dataset, NiFi splits each file into individual rows (or records in the case of JSON data).
   - Each row/record will be processed individually, treating each one as a separate message to simulate real-time data.
   - The FirstInFirstOut Prioritizer is used to move data in the order that it comes in.
3. **Add Delay Between Records:**
   - The `ControlRate` processor in NiFi is used to introduce a 1-second delay between records.
   - This delay simulates a real-time data stream by pacing the flow of messages instead of sending them all at once.
4. **Send Records to Kafka:**
   - NiFi's `PublishKafka` processor sends each delayed message to a Kafka topic (`tracking_data`).
   - This creates a continuous stream of data in Kafka, where each record arrives approximately every second.
5. **Data Transformation:**
   - NiFi's `ConsumeKafka` consumes messages from the `tracking_data` topic.
   - The dataset's schemas for both datasets are combined, and the newly formed dataset is transformed.
   - The transformed data is published to Kafka in a new topic(`publish_tracking_druid`).
6. **Load Data to Time-Series DB:**
   - Apache Druid consumes messages from the Kafka topic(`publish_tracking_druid`) in real-time data streaming.