Streaming Data Processing in Telecom Data Engineering

This document outlines the step-by-step approach for real-time streaming of telecom data using Apache Kafka, Databricks, and Power BI. The goal is to stream real-time network metrics such as signal strength and call drop rates, process them in Databricks, and visualize the data in Power BI.

# Tech Stack

1. Apache Kafka: Used for real-time stream ingestion of network performance metrics.  
2. Databricks: Used for processing streaming data with Structured Streaming.  
3. Delta Lake: Storage layer for real-time data with ACID properties.  
4. Power BI: Visualize streaming data in real-time dashboards.  
5. Python: Used for streaming ETL scripts.

# Step-by-Step Solution

## 1. Data Ingestion Using Kafka

Set up Apache Kafka to stream network performance data (e.g., signal strength, call drop rates) from various network devices. Kafka will act as the real-time ingestion layer for data sources such as network routers and devices.

## 2. Real-Time Data Processing in Databricks

Use Databricks Structured Streaming to process data as it arrives from Kafka. Apply business rules such as detecting network anomalies (e.g., signal drops or high call drop rates) and calculating metrics like average signal strength per region.

## 3. Storage in Delta Lake

Store the processed streaming data in Delta Lake. Use Delta Lake to manage real-time data with ACID transactions. This ensures data consistency while handling high-velocity data streams.

## 4. Sending Data to Power BI Using Streaming API

After processing the data, use the Power BI Streaming API to push real-time data to Power BI. Visualize metrics such as signal strength trends and call drop rates in real-time. Configure dashboards to update dynamically as new data is streamed into Power BI.

## Use Case: Real-Time Network Performance Monitoring

This use case involves streaming network performance metrics like signal strength and call drop rates. The data is processed in real-time and stored in Delta Lake. Power BI dashboards visualize the real-time metrics, providing insights into network health and potential outages.

# Streaming Processing Architecture

The streaming processing architecture is designed to handle real-time data ingestion, processing, and visualization. This architecture is suitable for real-time monitoring of telecom network performance metrics such as signal strength and call drop rates. Below is an overview of each component:

## 1. Data Sources

Real-time network performance metrics are collected from network devices, routers, and other equipment. These metrics include signal strength, call drop rates, and data transfer speeds. The data is streamed continuously from the devices into the streaming pipeline.

## 2. Apache Kafka (Stream Ingestion)

Apache Kafka is used as the ingestion layer to capture and stream data in real-time. Network devices push data to Kafka topics, which serve as message queues for the streaming data. Kafka ensures that data is ingested and processed in a fault-tolerant and scalable manner.

## 3. Databricks Structured Streaming (Real-Time Processing)

Databricks Structured Streaming is used to process the real-time data from Kafka in a continuous manner. Data is read from Kafka topics, transformed, and enriched using business logic. Databricks applies real-time transformations, such as calculating the average signal strength or detecting anomalies in network performance.

## 4. Delta Lake (Storage)

Processed real-time data is stored in Delta Lake, providing a reliable storage layer with ACID transaction support. Delta Lake enables streaming data to be stored in an organized format, allowing for both real-time and historical analysis. Delta Lake also supports time travel for querying older versions of the data.

## 5. Power BI Streaming Dataset (Real-Time Visualization)

Processed data is sent to Power BI using the Streaming API. Power BI dashboards are configured to update in real-time, displaying metrics such as signal strength trends, call drop rates, and network performance alerts. Dashboards are dynamically refreshed as new data is ingested and processed.

## Streaming Processing Workflow

1. Real-time data is ingested from network devices and streamed to Kafka topics.  
2. Databricks Structured Streaming reads data from Kafka and applies transformations.  
3. Transformed data is stored in Delta Lake in real-time.  
4. Data is pushed to Power BI using the Streaming API for real-time dashboard updates.