# Credit Card Fraud Detection Data Pipeline

**Problem Statement:** A financial institution wants to analyze credit card transaction data to detect fraudulent activities and understand patterns related to fraud. They need to build a data pipeline that collects, processes, and analyzes data from various sources to generate meaningful insights for decision-making. As a data engineer, your task is to design and implement the data pipeline to support this analysis.

**Dataset:**

The dataset for this project will consist of the Credit Card Fraud Detection dataset, which includes the following information:

* Time: Number of seconds elapsed between this transaction and the first transaction in the dataset.
* V1 to V28: These are the principal components obtained with PCA (Principal Component Analysis). The exact meaning of these features is not provided due to privacy reasons.
* Amount: Transaction amount.
* Class: 1 for fraudulent transactions, 0 otherwise.

## Project Steps:

**Data Collection:**

Download the dataset from Kaggle.

**Data Ingestion:**

Create an ingestion process to receive and store the raw data from the CSV file.

Use tools like Apache Sqoop for batch data ingestion.

Store the data in a data lake or distributed file system (e.g., HDFS).

**Data Processing:**

Design ETL (Extract, Transform, Load) processes to cleanse and transform the raw data.

Implement data quality checks and filtering to ensure data integrity.

Utilize Apache Spark for distributed data processing and transformation.

Apply data modeling techniques (e.g., data normalization, denormalization) as per the analysis requirements.

**Data Storage:**

Choose a suitable database or data warehouse (e.g., Apache Hive, Apache HBase) for storing processed data.

Create optimized tables and partitions for efficient querying and analysis.

Ensure data security and privacy measures are in place.

**Data Analysis and Visualization:**

Use SQL queries or Spark SQL to extract relevant insights from the processed data.

Perform fraud detection analysis based on the provided features.

Generate reports, dashboards, and visualizations using tools like Apache Superset, Tableau, or Power BI.

**Overall architecture flow:**

****

**Task 1: Data Ingestion and Storage**

Outcome: Ingest the credit card transaction data into Hadoop using Apache Sqoop, process it using Apache Spark, and store the results in HBase.

Deliverables:

Sqoop command to ingest data from the CSV file into Hadoop.

Spark code to process the ingested data.

HBase commands to store the processed data.

**Task 2: Fraud Detection Analysis**

Outcome: Analyze the credit card transaction data to identify fraudulent activities using Apache Spark and SQL queries.

Deliverables:

Spark code and SQL queries to analyze the credit card transaction data.

A report detailing the patterns related to fraudulent activities and any interesting trends in the data.

**Task 3: Data Visualization**

Outcome: Create visualizations of the fraudulent activities using a tool like Apache Superset, Tableau, or Power BI.

Deliverables:

Visual representations (e.g., line charts, bar charts, heatmaps) of the fraudulent activities.

A dashboard that displays the visualizations and allows users to interact with the data.

* Number of fraudulent transactions over time
* Correlation between transaction amount and fraud likelihood
* Distribution of fraudulent transactions by card type or country
* Time of day or day of week with the most fraudulent transactions

For all these tasks, you can use the [Credit Card Fraud Detection dataset](https://www.kaggle.com/mlg-ulb/creditcardfraud) mentioned above. Please note that you might need to preprocess the dataset to fit the exact requirements of the assignment. Also, remember to respect the terms of use for the dataset.

**Tools required to achieve the end dashboard:**

1. **Hadoop ( hdfs , hive , spark )**
2. **Tableau/Power BI**