1. Business Problem

Customer churn, where customers stop using a company's services or products, leads to revenue losses and increases competitive pressure. Given that acquiring new customers is costly, businesses focus on retaining existing customers through predictive analytics. The goal is to leverage data-driven insights to proactively identify at-risk customers and implement targeted interventions to reduce churn.

2. Business Objectives

- Develop an automated data pipeline to process raw customer data from multiple sources.
- Enable predictive modelling to identify customers at risk of churning.
- Improve customer retention strategies by providing actionable insights.
- Reduce revenue loss by implementing data-driven interventions.

3. Key Data Sources & Attributes

The pipeline will process data from the following sources:

1. **Open Banking Transactions Dataset**
    a. **Source**: Kaggle - Bank Customer Churn Dataset
    b. **Description**: This dataset contains customer transaction history, demographics, and banking behaviour.
    c. **Attributes:** customer_id, credit_score, country, gender, age, tenure, balance, products_number, credit_card, active_member, estimated_salary, churn

2. **Bank Customer Churn Dataset**
    a. **Source**: Maven - Bank Customer Churn Dataset
    b. **Description**: This dataset contains behavioural factors like purchase frequency and credit score.
    c. **Attributes**: CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited

4. Expected Pipeline Outputs

1. **Clean Datasets for Exploratory Data Analysis (EDA)**
    a. Pre-processed data, handling missing values, outliers, and inconsistencies.
    b. Merged dataset from various sources for a holistic customer view.
2. **Transformed Features for Machine Learning**

    a. Feature engineering: customer tenure, engagement scores, recency-frequency-monetary (RFM) metrics.
    b. Normalized, encoded, and aggregated data for model training.
3. **Deployable Machine Learning Model**
    a. Trained and validated churn prediction model.
    b. Model API or batch predictions for integration with business applications.

5. Evaluation Metrics

To measure model performance, the following metrics will be used:

- **Accuracy**: Overall correctness of predictions.
- **Precision & Recall**: Balance between identifying churners correctly (precision) and capturing all churners (recall).
- **F1 Score**: Harmonic mean of precision and recall for balanced evaluation.
- **ROC-AUC Score**: Measures the model's ability to distinguish between churners and non-churners.
- **Confusion Matrix**: Provides insights into false positives and false negatives.