Node Descriptions (3-line explanation for each)

1️⃣ Data Ingestion (🟢 Start Node)

- Extracts raw data from multiple sources like **SQL databases, S3 buckets, and campaign logs**.
- Includes **direct features** such as impressions, clicks, conversions, and ad spend.
- Stores structured data in a **centralized repository** for further processing.

2️⃣ Exploratory Data Analysis (EDA) (🔵)

- Performs **data cleaning**, missing value handling, and outlier detection.
- Uses **visualizations (univariate, bivariate, multivariate)** to understand feature distributions.
- Identifies **trends, patterns, and potential correlations** between features.

3️⃣ Data Preprocessing (🟡)

- Standardizes and normalizes numerical features for better model performance.
- Handles **time-series data** by aggregating, smoothing, or detrending.
- Converts categorical data into numerical form using encoding techniques.

4️⃣ Feature Engineering & Selection (🟠)

- Extracts **new KPI-driven metrics** like CTR, ROI, CPC from raw features.
- Performs **correlation analysis & VIF check** to select the most relevant features.
- Reduces dimensionality using **PCA or feature elimination techniques**.

5️⃣ Model Training & Evaluation (🔴)

- Splits data into training and testing sets for unbiased learning.
- Trains **statistical models (ARIMA, SARIMAX), ML models (XGBoost, Random Forest), or deep learning models (LSTMs, Transformers)**.
- Evaluates models using **KPI-focused metrics** like RMSE, R^2, Precision-Recall.

6️⃣ Model Deployment & Optimization (🟣 End Node)

- Saves trained models as **Pickle (.pkl), H5 (.h5), or ONNX** formats.
- Deploys using **Docker, FastAPI, AWS Lambda, or Kubernetes**.
- Monitors performance via **Grafana, Prometheus, or MLflow**, and iteratively improves.