THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Advance ML Project: YouTube trending video prediction using AWS (USA Dataset)

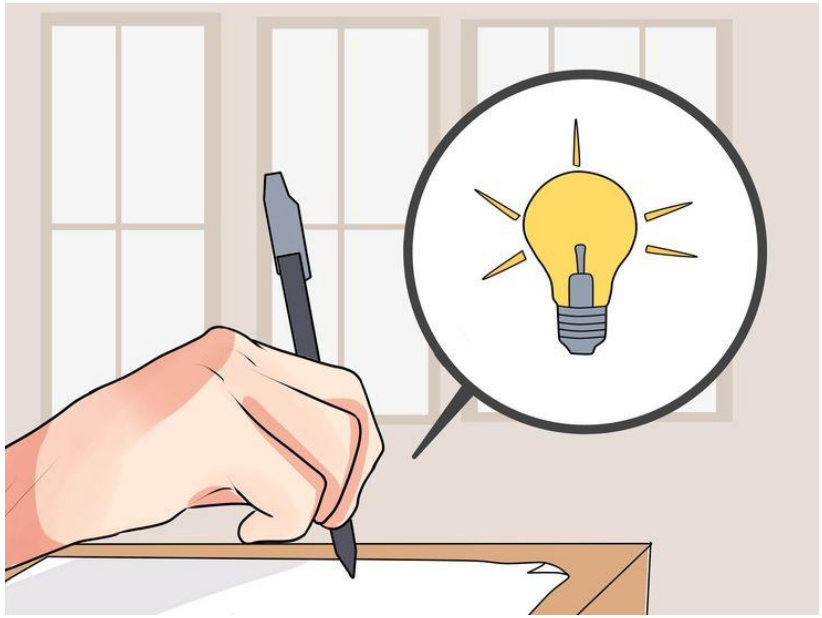
A Cloud-Based Machine Learning Project

Group – 6
Sirisha Ginnu (G27150918)
Sagar Shah (G26436634)
Chaya Chandana Doddaiggaluru Appajigowda (G35408608)

Cloud Computing - DATS 6450
George Washington University

Scope of the Project

- **Automating YouTube trending video analysis with AWS** to predict future trends.
- **Using AWS for data ingestion, cleaning, and NLP** on US YouTube Trending Data.
- **Deploying an ML model via API** for real-time trend prediction.
- **Dashboard Implementation:** A real-time interactive dashboard is being developed to visualize key insights from YouTube trending data.



Dataset Overview – US YouTube Trending Data

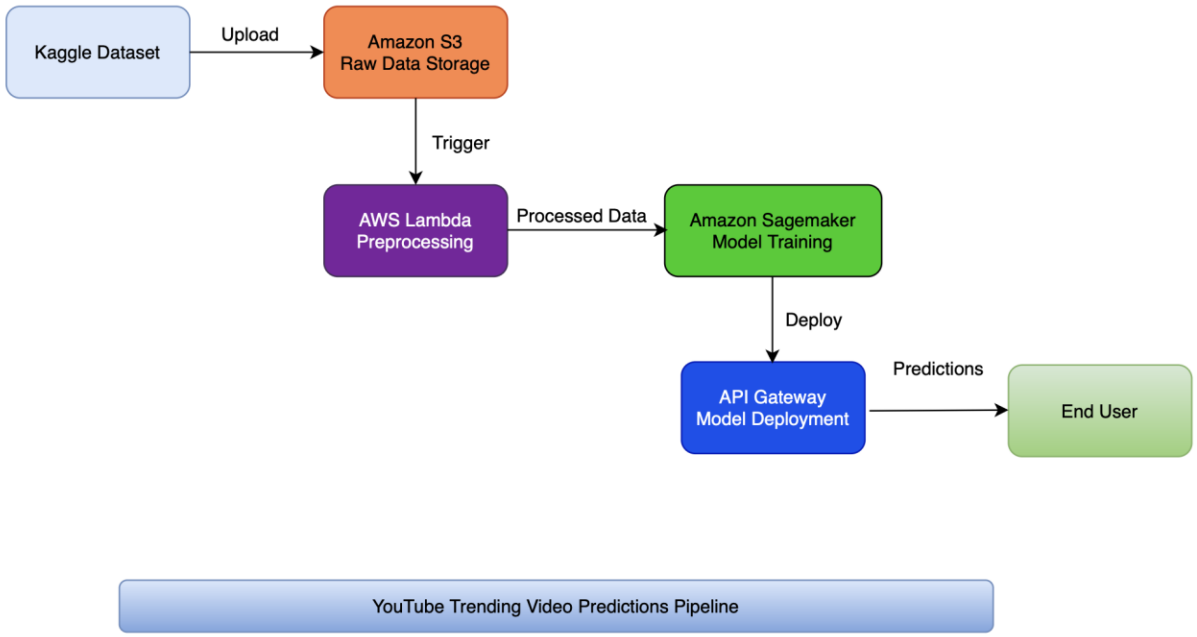
Dataset Summary -

- **Total Entries:** Large-scale YouTube video metadata (~100K+ records)
- **Data Source:** Kaggle - US YouTube Trending Dataset
- **Objective:** Analyze video trends and build an NLP model to predict trending content

Category	Feature	Description
Metadata	`video_id`	Unique video ID
	`title`	Video title
	`publishedAt`	Upload date/time
	`channelTitle`	Channel name
	`categoryId`	Video category
Engagement	`view_count`	Total views
	`likes`, `dislikes`	User reactions
	`comment_count`	Number of comments
Trend Data	`tags`	Video keywords
	`trending_date`	Date video trended
Target	`IsTrending`	Trending status

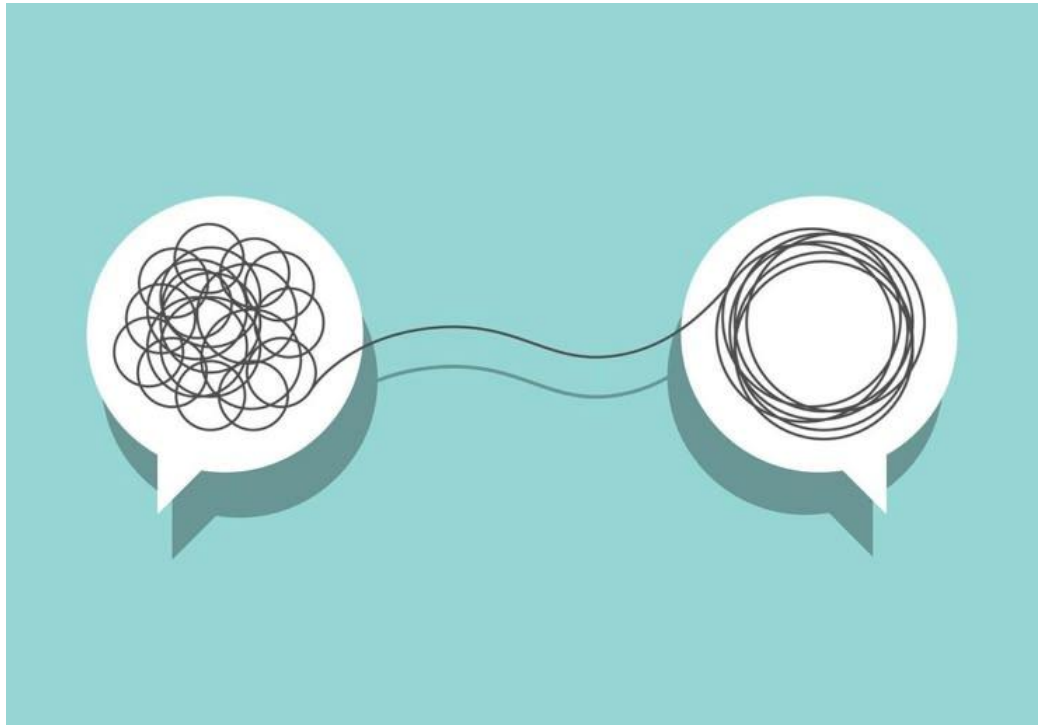
Features to be Implemented

- **Data Collection:** Downloading YouTube trending data from Kaggle.
- **Storage:** Uploading raw data to Amazon S3.
- **Data Preprocessing:** Using AWS Lambda to clean and transform the dataset.
- **Model Training:** Using Amazon Sage-Maker with pre-built NLP models.
- **Deployment:** Exposing the trained model as an API endpoint.



Expected Outcomes

- **Automated Data Pipeline:** A fully automated pipeline from data ingestion to ML predictions.
- **Scalable ML Deployment:** A serverless NLP model accessible via API.
- **Cloud-Based Solution:** Minimal maintenance using AWS-managed services.
- **Real-time Analysis:** Immediate insights from YouTube trending video data.



Data Flow Process

- **Data Collection:** Kaggle dataset uploaded to S3.
- **Data Cleaning:** AWS Lambda performs preprocessing (removing duplicates, filling missing values, etc.).
- **Data Transformation:** Preprocessed data sent to SageMaker for model training.
- **Model Output:** Trained NLP model predictions.
- **Deployment:** Model deployed as API using Lambda & API Gateway.

Interactive Dashboard for YouTube Trend Analysis

Purpose:

1. Provides real-time insights into YouTube video trends.
2. Helps visualize key metrics for trend prediction.

Dashboard Features:

1. **Trending Categories:** Identify the most popular video genres.
2. **Engagement Metrics:** Analyze likes, dislikes, comments, and shares.
3. **Trend Frequency:** Track daily/weekly trending patterns.
4. **Sentiment Analysis:** Evaluate audience reactions using NLP.

Thank you!