# Project: Youtube

## Project Overview

In this project, I would like to analyse YouTube data to get more knowledgeable on factors that affect a popular video on YouTube.

To achieve my goal, I must design a data pipeline and the following criteria are required to be successful in my endeavour:

1. Data Ingestion: Ingest data from multiple sources, one-offs and incrementally.
2. Data Lake: Design and build a new data lake architecture.
3. AWS cloud: AWS as the cloud provider.
4. ETL design: Extract, transform and load data efficiently.
5. Scalability: The data architecture should scale efficiently.
6. Reporting: Build a business Intelligence tier, including a dashboard.

## 2. Required Skills

* Building a data lake from scratch in Amazon S3: Joining semi-structured and structured data
* Lake House architecture design: Best practices, cost and performance.
* ETL in AWS Glue Spark jobs
* Amazon SNS for alerting
* SQL using Amazon Athena and Spark SQL
* Script to ingest data incrementally and schema evolution
* Business Intelligence dashboard in Amazon QuickSight

## 3. Dataset

Kaggle is the source of the dataset used for this project. The data is collected using YouTube API to answer the following question:

1. Top Trending videos
2. What is “Trending”? : How can I use YouTube factors, including user interactions to make a video trending?

* Launching new data-driven campaign
* Main advertising channel: Youtube
* Initial questions to answer:
* “How to categorise videos, based on their comments and statistics”
* “What factors affect how popular a YouTube video will be”