YOUTUBE DATA INGESTION

Github: [YoutubeInsights](https://github.com/dibnyk/project/tree/main/Youtube%20data%20load)

* To do this project we are have used AWS EC2, S3 and airflow to create the pipeline.
* Using Youtube’s API we are downloading data of a channel and loading it into s3 in the inbound folder.
* Then to transform the data we are using pyspark.
* All this script execution will be done using EC2 instance.
* And to create the Pipeline Airflow would be used.

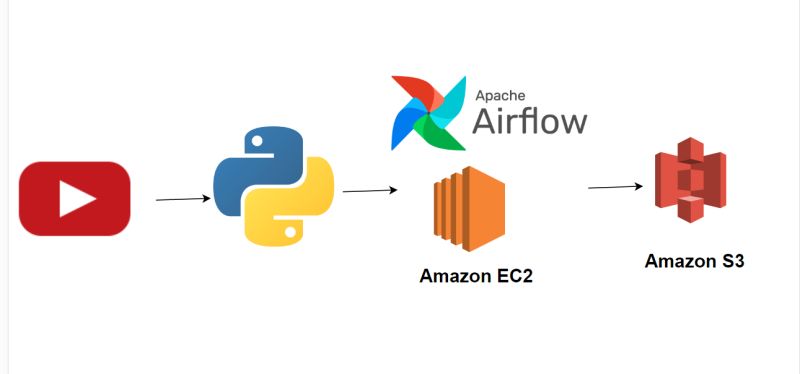
**Youtube API:**

* Youtube API can be found on google cloud console page from wehre we will get the credentials and authentication key to call our API.
* The API will be used to download data of a particular channel.



**Inbound:**

* When the API is called it will load data in the form of a json file, which will have the data of each video of the channel, its likes, subscribers count, video count and many more.
* Then from this json string we are gonna extract the required data and transform them.
* Then the required data is stored as a csv in S3 bucket for inbound files.



**Transform:**

* The inbound csv is sourced from the S3 bucket and used to do some transformations like getting the increase in subscribers count for each video and many more.
* Using spark we will transform the data and load it into the warehouse S3 bucket.
* To create the Pipeline we will use Airflow which will take the scripts as parameters and execute them one after the other.