ETL Final Workshop - Manuel Henao - Github Repo

Problem Description: The project aims to analyze the performance of a facial validation API using an ETL process. The API, built on a microservices architecture, includes face detection, accessory detection, image quality assessment, and Presentation Attack Detection (PAD) to prevent identity spoofing. Performance data is stored in a non-relational database. The primary objectives are to identify execution time bottlenecks of each microservice, analyze acceptance rates per client, and improve user experience by examining retry patterns.

Context: The facial validation API is designed to ensure secure and efficient identity verification by capturing two moments: one where the user is near the camera and another where the user is farther away. This process, known as liveness detection, aims to enhance the performance of the PAD microservice. Each microservice will have execution metrics for both images. The API comprises several microservices:

- Face Detection: Identifies and locates faces in images.
- Accessory Detection: Detects accessories like glasses or masks that might obscure the face.
- Image Quality Assessment: Evaluates the quality of the image to ensure it meets required standards using various methods.
- Presentation Attack Detection (PAD): Detects attempts to spoof the system using photos, videos, or masks.

Workflow Diagram of Facial Validation API by microservice
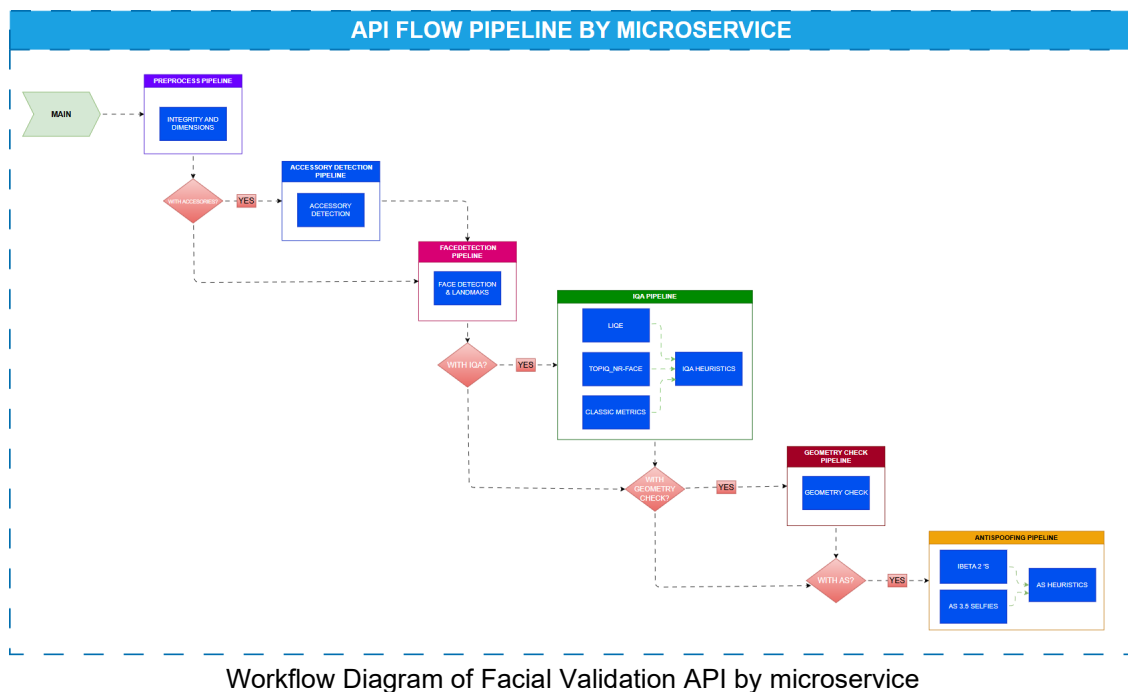
The API's performance data is stored in a non-relational database using MongoDB, capturing metrics such as execution times, acceptance rates, and retry counts for two attempts. The pipeline workflow executes the microservices in a serial manner. For example, if the accessory detection microservice detects an accessory, it will terminate the transaction without processing the subsequent microservices.

Dataset Description: The dataset includes various performance metrics for each microservice. Below is a model card for the variables of interest:

feature	description	unit
api_hash	API Version	string
user_name	Username consuming the API	string
file_token	Unique identifier (hash) of the transaction	string
service	API endpoint consumed (Service)	string
summary_code	Summary response code	int
summary_status	Summary status	string
summary_desc	Summary description	string
http_code	HTTP code of the transaction	int
total_time_ms	Total transaction time in ms	ms
date	Date	date
procesoConvenioGuid	GUID of the agreement process	string
convenio	Agreement (Client ID)	string
documento	Document Identifier	string
img_size_near_img_width	Width of the near image	px
img_size_near_img_height	Height of the near image	px
img_size_far_img_width	Width of the far image	px
img_size_far_img_height	Height of the far image	px
score_liqe_near_img	LIQE microservice score of the near image	double
score_liqe_far_img	LIQE microservice score of the far image	double
score_topiq_near_img	TOPIQ microservice score of the near image	double
score_topiq_far_img	TOPIQ microservice score of the far image	double
score_sharpness_classic_metrics_near_img	Classic sharpness microservice score of the near image	double
score_sharpness_classic_metrics_far_img	Classic sharpness microservice score of the far image	double
score_colorfulness_classic_metrics_near_img	Classic colorfulness microservice score of the near image	double
score_colorfulness_classic_metrics_far_img	Classic colorfulness microservice score of the far image	double
score_contrast_classic_metrics_near_img	Classic contrast microservice score of the near image	double
score_contrast_classic_metrics_far_img	Classic contrast microservice score of the far image	double
score_brightness_classic_metrics_near_img	Classic brightness microservice score of the near image	double
score_brightness_classic_metrics_far_img	Classic brightness microservice score of the far image	double
score_blur_score_classic_metrics_near_img	Classic blur microservice score of the near image	double
score_blur_score_classic_metrics_far_img	Classic blur microservice score of the far image	double
score_svd_score_classic_metrics_near_img	Classic SVD microservice score of the near image	double
score_svd_score_classic_metrics_far_img	Classic SVD microservice score of the far image	double
prob_as_35_selfies_near_img	Probability as 3.5 selfies microservice for the near image	double
prob_as_35_selfies_far_img	Probability as 3.5 selfies microservice for the far image	double
prob_ibeta2_crops_near_img	Probability ibeta2 crops microservice for the near image	double
prob_ibeta2_crops_far_img	Probability ibeta2 crops microservice for the far image	double
prob_ibeta2_full_near_img	Probability ibeta2 full microservice for the near image	double
prob_ibeta2_full_far_img	Probability ibeta2 full microservice for the far image	double
prob_ibeta2_clip_near_img	Probability ibeta2 clip microservice for the near image	double
prob_ibeta2_clip_far_img	Probability ibeta2 clip microservice for the far image	double
prob_as_heuristics_near_img	Heuristic probability microservice for the near image	double
prob_as_heuristics_far_img	Heuristic probability microservice for the far image	double
time_ms_accessory_detector_near_img	Time in ms of accessory detector microservice for the near image	ms
time_ms_accessory_detector_far_img	Time in ms of accessory detector microservice for the far image	ms
queue_time_ms_accessory_detector_near_img	Queue time in ms of accessory detector microservice for the near image	ms
queue_time_ms_accessory_detector_far_img	Queue time in ms of accessory detector microservice for the far image	ms
time_ms_face_detector_near_img	Time in ms of face detector microservice for the near image	ms
time_ms_face_detector_far_img	Time in ms of face detector microservice for the far image	ms
queue_time_ms_face_detector_near_img	Queue time in ms of face detector microservice for the near image	ms
queue_time_ms_face_detector_far_img	Queue time in ms of face detector microservice for the far image	ms
time_ms_liqe_near_img	Time in ms of LIQE microservice for the near image	ms
time_ms_liqe_far_img	Time in ms of LIQE microservice for the far image	ms
queue_time_ms_liqe_near_img	Queue time in ms of LIQE microservice for the near image	ms
queue_time_ms_liqe_far_img	Queue time in ms of LIQE microservice for the far image	ms
time_ms_topiq_near_img	Time in ms of TOPIQ microservice for the near image	ms
time_ms_topiq_far_img	Time in ms of TOPIQ microservice for the far image	ms
queue_time_ms_topiq_near_img	Queue time in ms of TOPIQ microservice for the near image	ms
queue_time_ms_topiq_far_img	Queue time in ms of TOPIQ microservice for the far image	ms
time_ms_classic_metrics_near_img	Time in ms of classic metrics microservice for the near image	ms
time_ms_classic_metrics_far_img	Time in ms of classic metrics microservice for the far image	ms
time_ms_as_35_selfies_near_img	Time in ms as 3.5 selfies for the near image	ms
time_ms_as_35_selfies_far_img	Time in ms as 3.5 selfies for the far image	ms
queue_time_ms_as_35_selfies_near_img	Queue time in ms as 3.5 selfies for the near image	ms
queue_time_ms_as_35_selfies_far_img	Queue time in ms as 3.5 selfies for the far image	ms
time_ms_ibeta2_crops_near_img	Time in ms of ibeta2 crops microservice for the near image	ms
time_ms_ibeta2_crops_far_img	Time in ms of ibeta2 crops microservice for the far image	ms
queue_time_ms_ibeta2_crops_near_img	Queue time in ms of ibeta2 crops microservice for the near image	ms
queue_time_ms_ibeta2_crops_far_img	Queue time in ms of ibeta2 crops microservice for the far image	ms
time_ms_ibeta2_full_near_img	Time in ms of ibeta2 full microservice for the near image	ms
time_ms_ibeta2_full_far_img	Time in ms of ibeta2 full microservice for the far image	ms
queue_time_ms_ibeta2_full_near_img	Queue time in ms of ibeta2 full microservice for the near image	ms
queue_time_ms_ibeta2_full_far_img	Queue time in ms of ibeta2 full microservice for the far image	ms
time_ms_ibeta2_clip_near_img	Time in ms of ibeta2 clip microservice for the near image	ms
time_ms_ibeta2_clip_far_img	Time in ms of ibeta2 clip microservice for the far image	ms
queue_time_ms_ibeta2_clip_near_img	Queue time in ms of ibeta2 clip microservice for the near image	ms
queue_time_ms_ibeta2_clip_far_img	Queue time in ms of ibeta2 clip microservice for the far image	ms
time_ms_geometry_check	Time in ms of geometry check microservice	ms
queue_time_ms_geometry_check	Queue time in ms of geometry check microservice	ms

Process

The ETL Workshop will be broken down into 3 main stages

1. **Extract Stage:** Involves the collection performance data from the API's non-relational database.
   a. Due to sensible data the raw data which is stored in a non-relational db in a json format some features were omitted and some others were anonymized because clients transactions behavior were involved. In order to achieve that as part of the ETL process the first stage called **000_anonymized.py** anonymizes and omits raw semi-structured data into a tabular format represented in a CSV file with the features mentioned in the model card.
   b. Following the ETL process the second stage is loading the data as part of the staging part that executed by **001_staging.ipynb** stages the raw data (after being anonymized) in a relational database.
2. **Transform Stage:** It Covers the cleaning, transformation and preprocessing of the data to ensure consistency and accuracy and relevant insights based on data.
3. **Load Stage:** Load the transformed data into an analytical database for further analysis

Evidences

Extract Stage Evidence after executing **001_staging.ipynb**

pgAdmin4 query access