PORTFOLIO

Bad Boys Club

Alan Valbuena – Ariel Buenfil – Damaris Dzul – Diego Monroy – Paulina Chiquete – Sergio Barrera

OVERVIEW

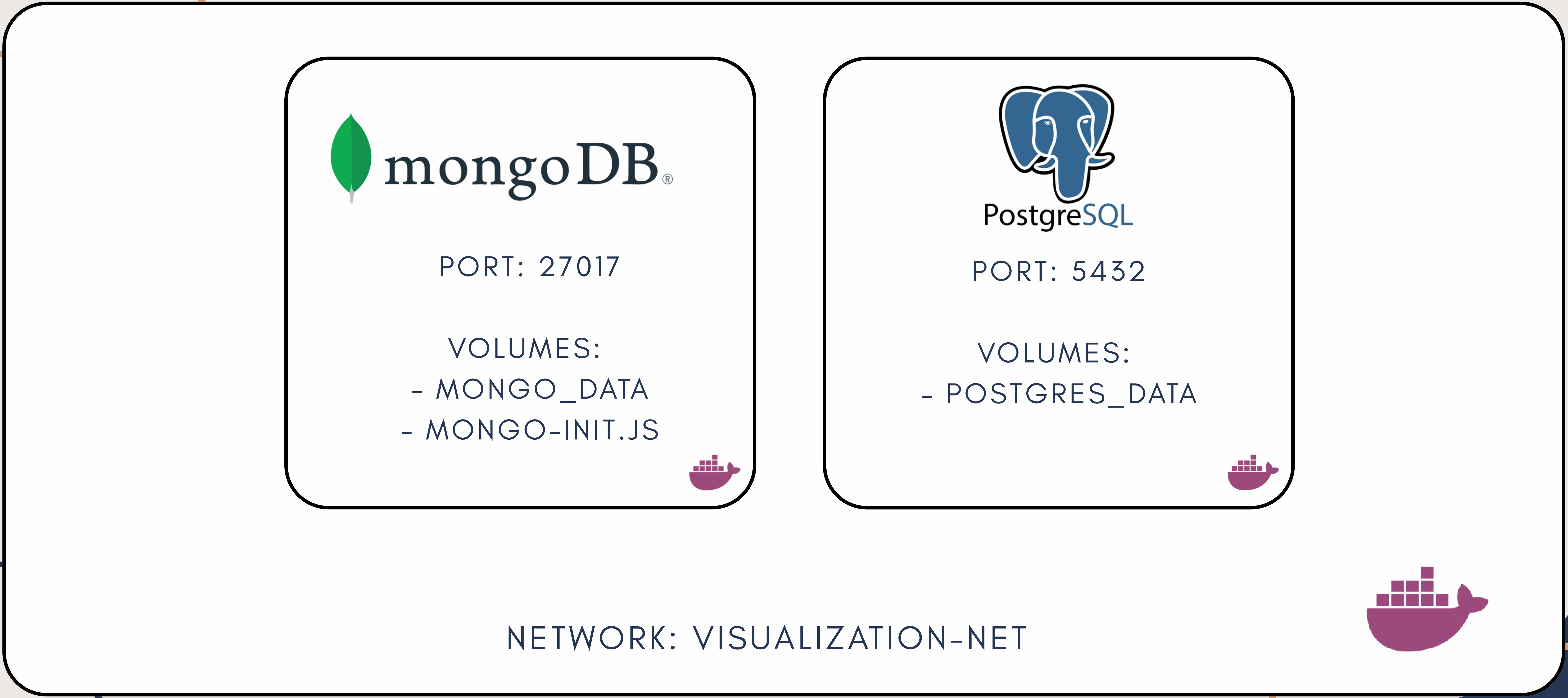
To design, implement, and deploy a multi-service data engineering solution that analyzes simulated video streaming data and presents key insights via an interactive dashboard. The objective of this analysis is to perform a comprehensive Exploratory Data Analysis (EDA) on three integrated datasets (users, viewing sessions, and content) to evaluate data quality, identify patterns in user behavior, and generate insights into demographics, engagement, and content consumption trends.

SYSTEM ARCHITECTURE

- API Service (Flask) → CRUD + analytics endpoints.
- Web Service (HTML) → Interactive dashboard.
- Injector Service → ETL pipelines (CSV/JSON processing).
- PostgreSQL → Structured data (salaries, users).
- MongoDB → Semi-structured data (metadata, sessions).

Benefits: scalability, modularity, and portable deployment.

MICROSERVICE - DBS



PORT: 27017

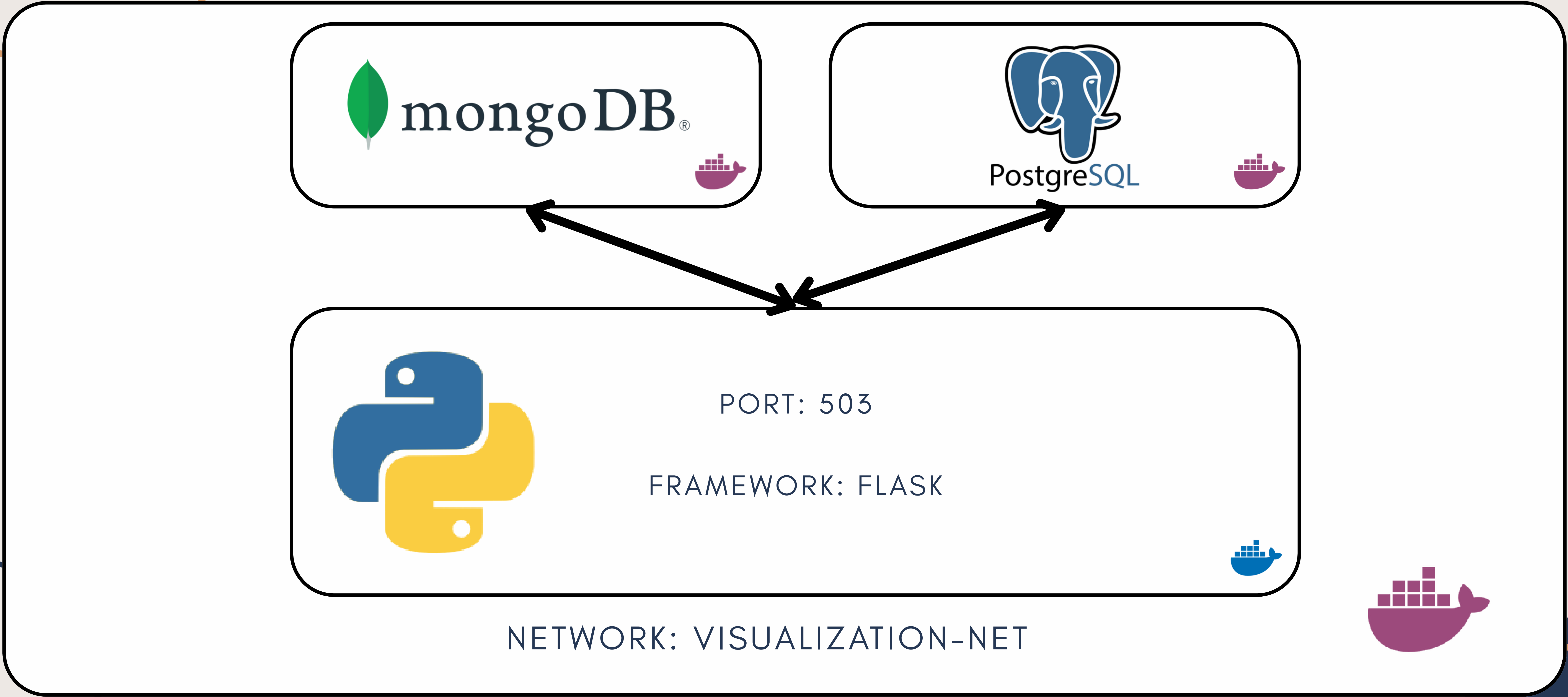
VOLUMES:
- MONGO_DATA
- MONGO-INIT.JS

PORT: 5432

VOLUMES:
- POSTGRES_DATA

NETWORK: VISUALIZATION-NET

MICROSERVICE - API

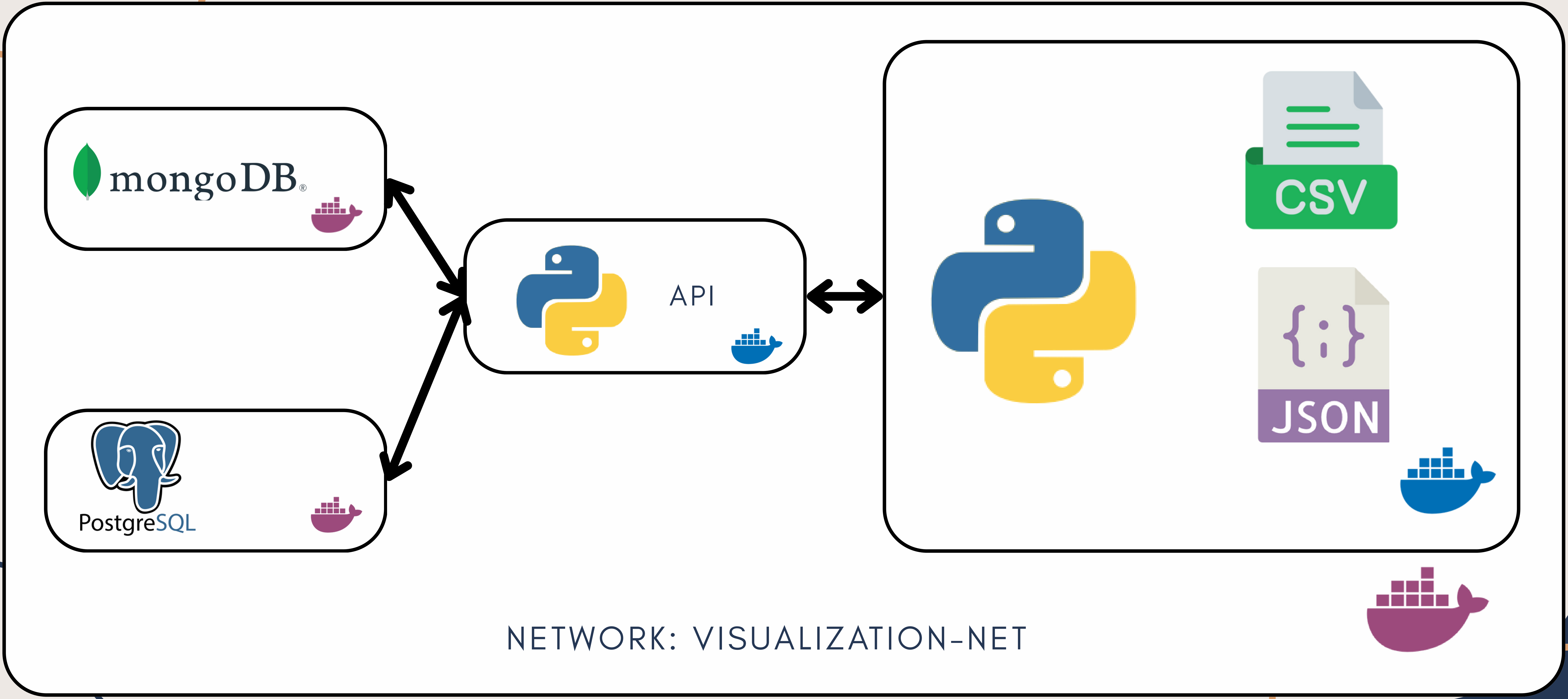


PORT: 503

FRAMEWORK: FLASK

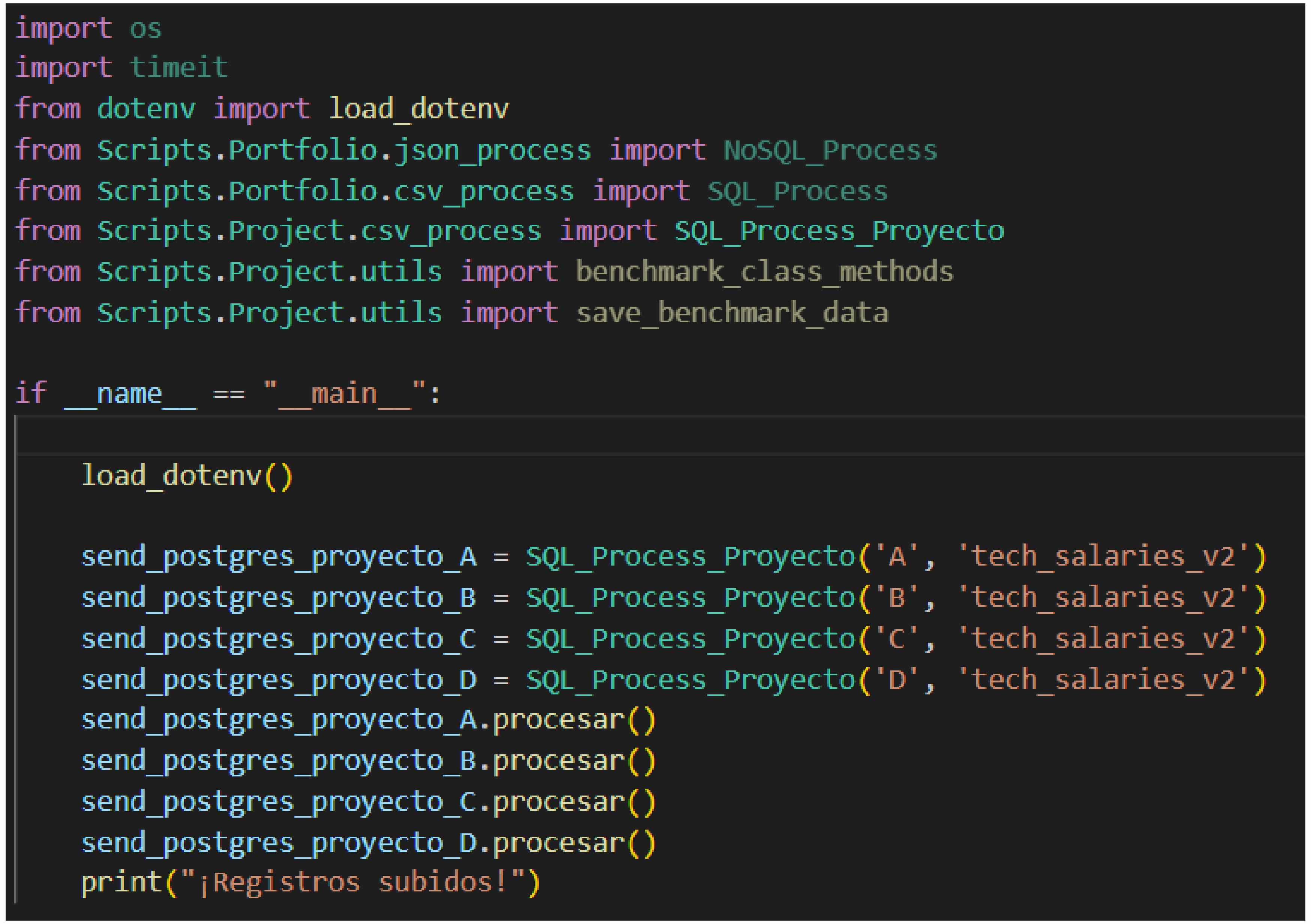
NETWORK: VISUALIZATION-NET

MICROSERVICE - INJECTOR



SUBMISSION METHODS

MICROSERVICE - INJECTOR



```
import os
import timeit
from dotenv import load_dotenv
from Scripts.Portfolio.json_process import NoSQL_Process
from Scripts.Portfolio.csv_process import SQL_Process
from Scripts.Project.csv_process import SQL_Process_Proyecto
from Scripts.Project.utils import benchmark_class_methods
from Scripts.Project.utils import save_benchmark_data

if __name__ == "__main__":

    load_dotenv()

    send_postgres_proyecto_A = SQL_Process_Proyecto('A', 'tech_salaries_v2')
    send_postgres_proyecto_B = SQL_Process_Proyecto('B', 'tech_salaries_v2')
    send_postgres_proyecto_C = SQL_Process_Proyecto('C', 'tech_salaries_v2')
    send_postgres_proyecto_D = SQL_Process_Proyecto('D', 'tech_salaries_v2')
    send_postgres_proyecto_A.procesar()
    send_postgres_proyecto_B.procesar()
    send_postgres_proyecto_C.procesar()
    send_postgres_proyecto_D.procesar()
    print("¡Registros subidos!")
```

SUBMISSION METHODS

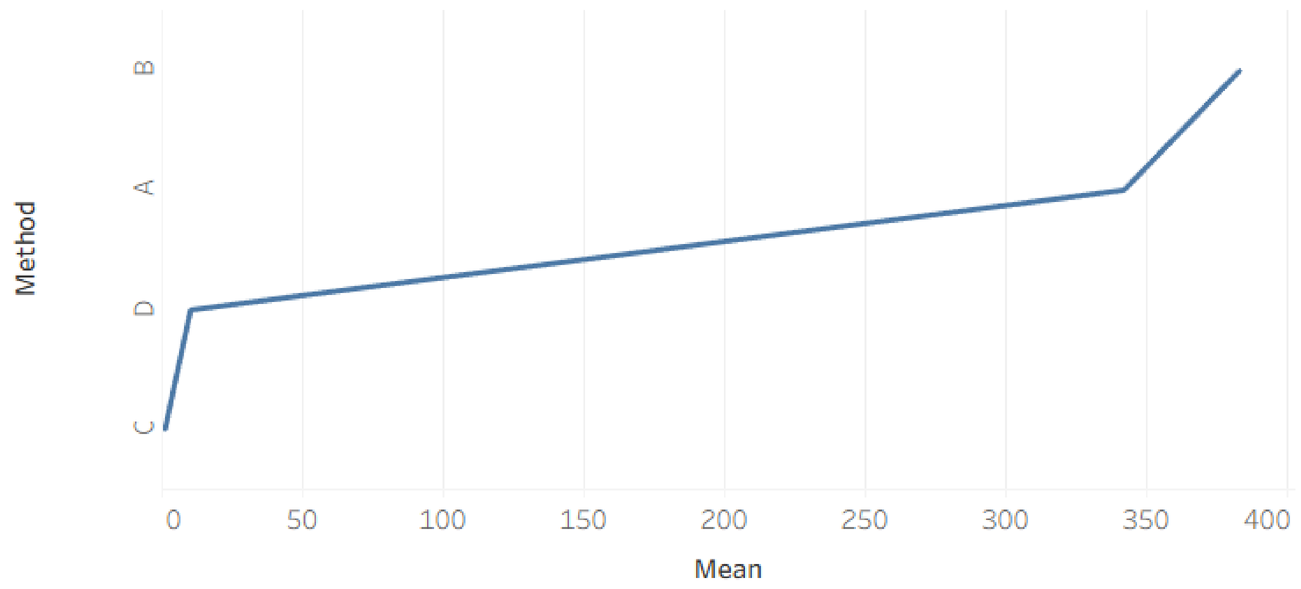
MICROSERVICE - INJECTOR

- **A:** PROCESS THE DATASET RECORD BY RECORD, VALIDATE THE DATA TYPE, AND IF EVERYTHING IS CORRECT, UPLOAD IT TO THE API RECORD BY RECORD.
- **B:** PROCESS THE DATASET AS A DATAFRAME, CONVERT THE COLUMN TO A SPECIFIC DATA TYPE, AND THEN PROCESS THE DATAFRAME ROW BY ROW TO UPLOAD IT TO THE API.
- **C:** PROCESS THE DATASET AS A DATAFRAME, CONVERT THE COLUMN TO A SPECIFIC DATA TYPE, UPLOAD IT TO THE API WITH A SINGLE CALL, UPLOADING THE ENTIRE DATAFRAME AS A LIST WITHIN THE PAYLOAD.
- **D:** PROCESS THE DATASET AS A DATAFRAME, CONVERT THE COLUMN TO A SPECIFIC DATA TYPE, AND UPLOAD IT TO THE API IN BLOCKS OF 50 RECORDS (FOR EXAMPLE).

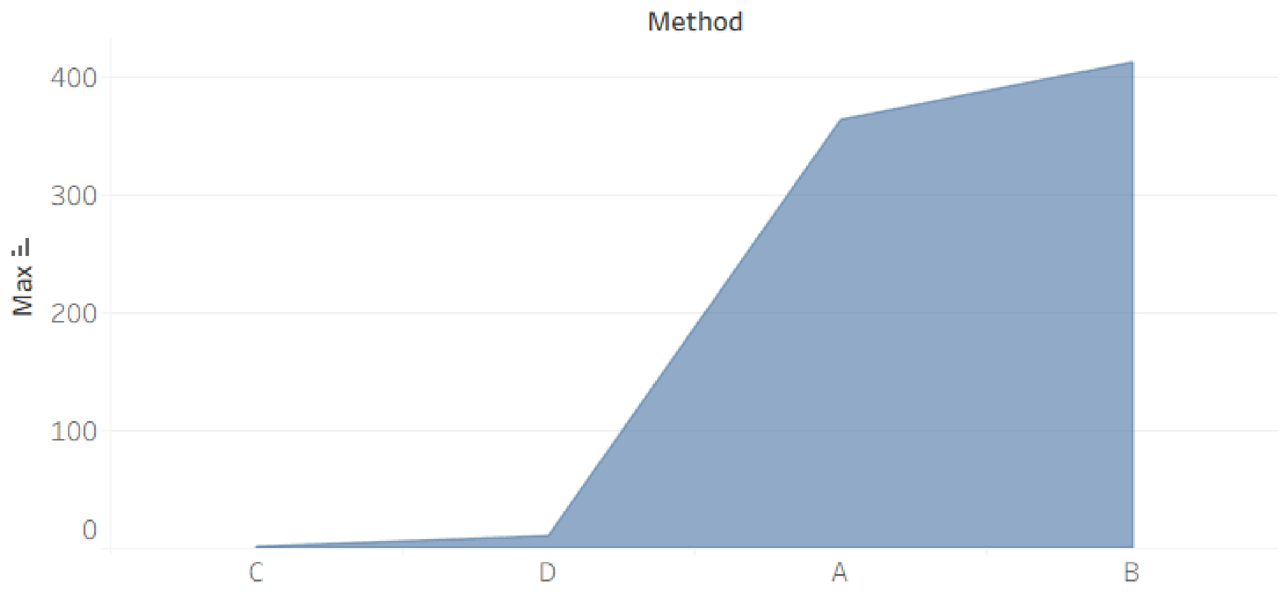
BENCHMARKING

MICROSERVICE - INJECTOR

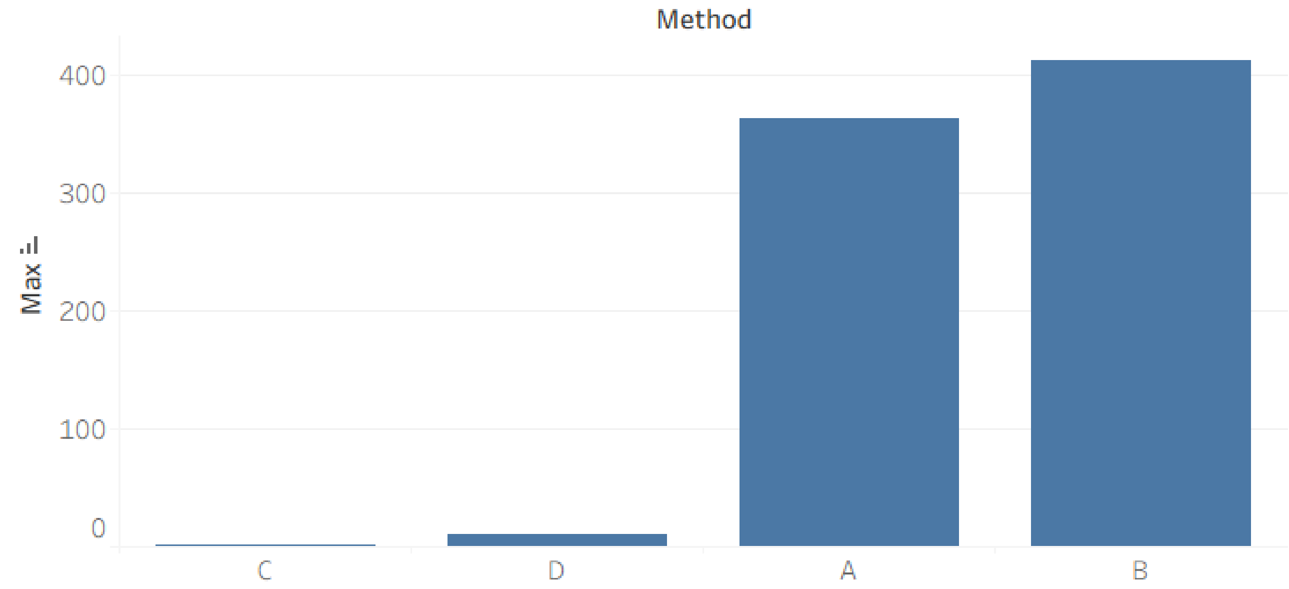
PROMEDIO DE EJECUCIÓN DE MÉTODOS



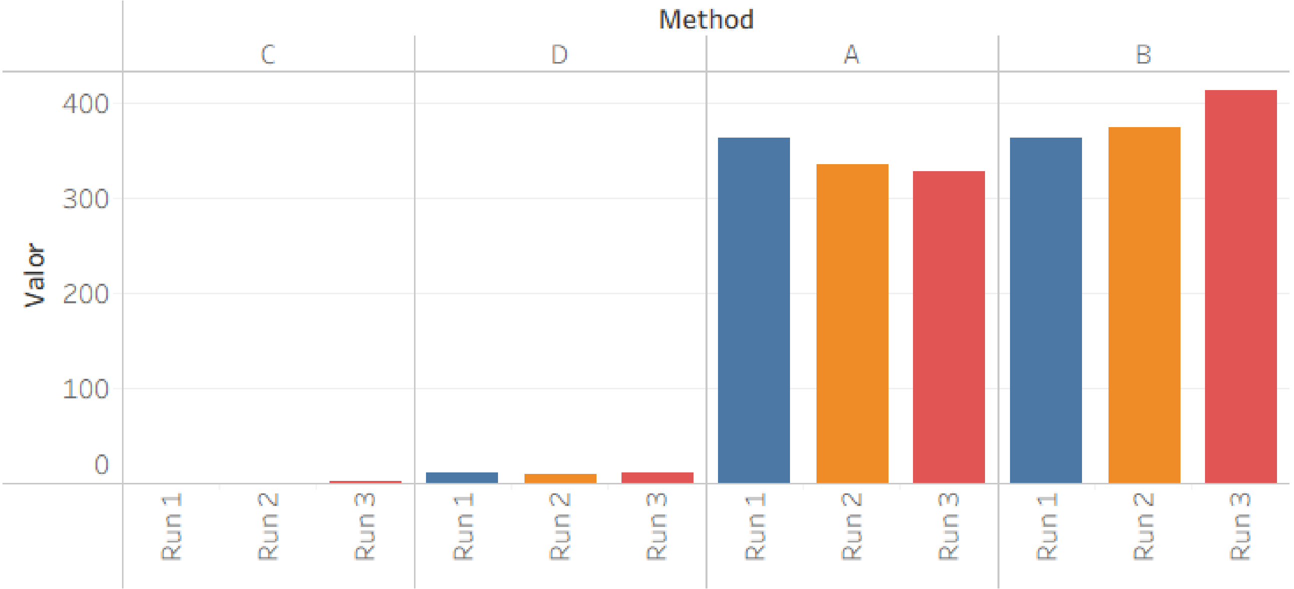
TIEMPO MÍNIMO DE CADA MÉTODO



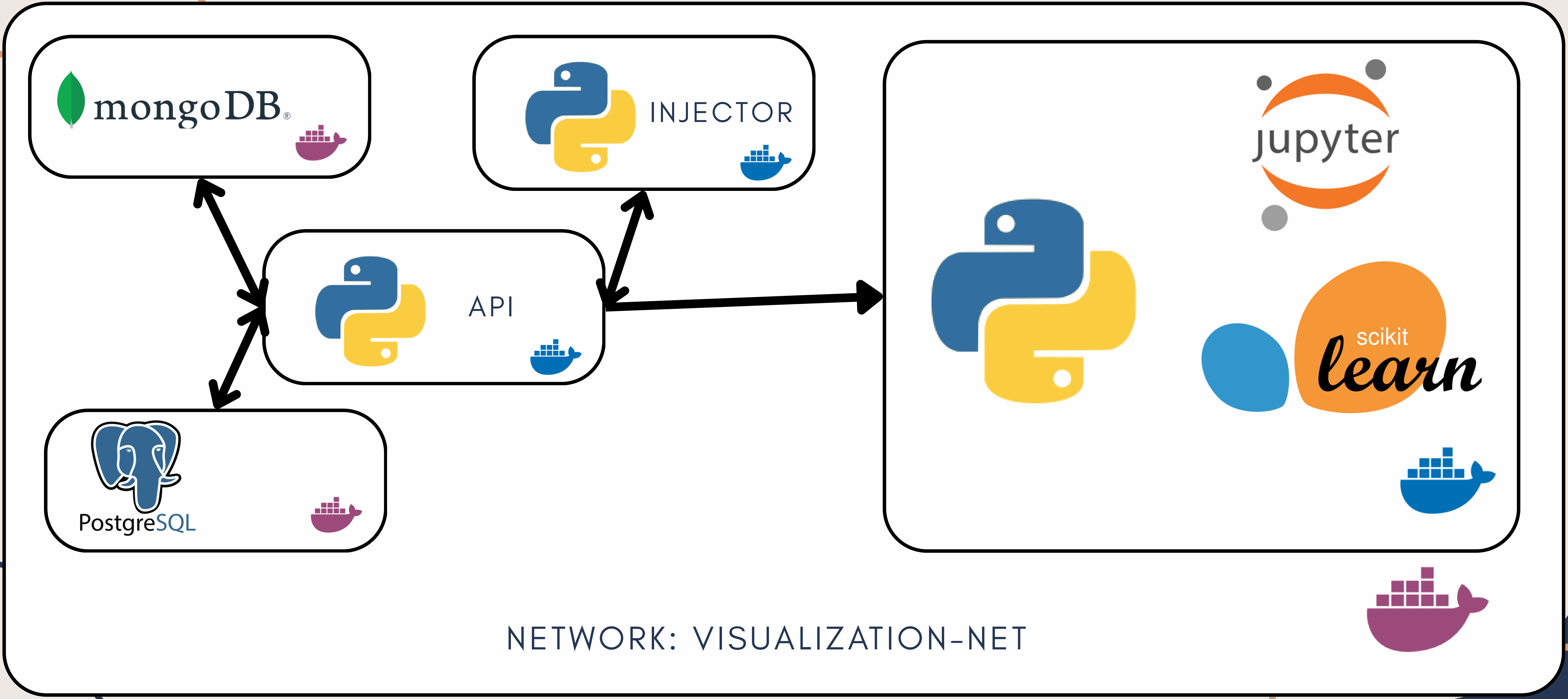
TIEMPO MÁXIMO DE CADA MÉTODO



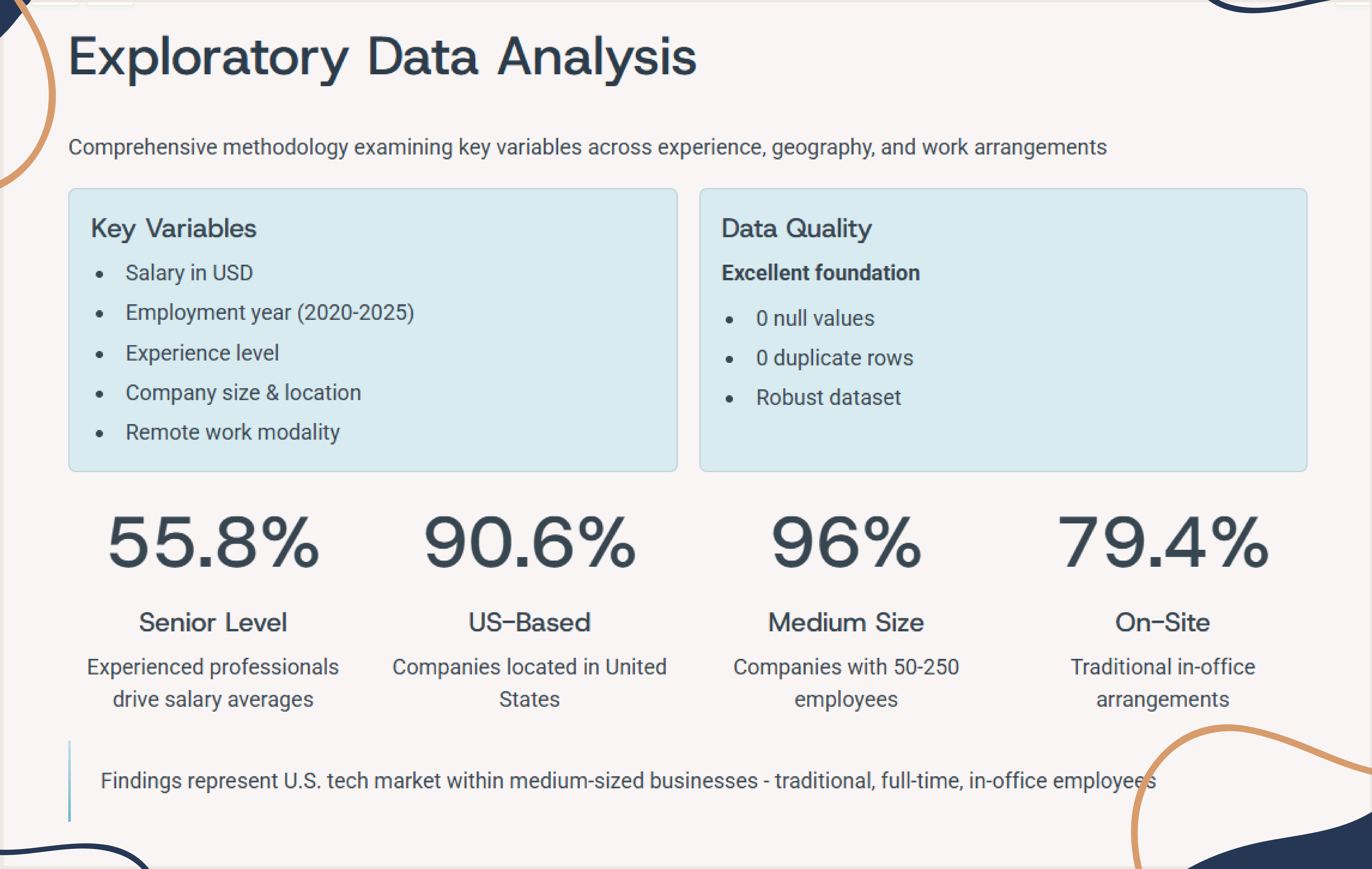
COMPARACIÓN DE CADA EJECUCIÓN POR MÉTODO



MICROSERVICE - ML



NETWORK: VISUALIZATION-NET



Exploratory Data Analysis

Comprehensive methodology examining key variables across experience, geography, and work arrangements

Key Variables

- Salary in USD
- Employment year (2020-2025)
- Experience level
- Company size & location
- Remote work modality

Data Quality

Excellent foundation

- 0 null values
- 0 duplicate rows
- Robust dataset

55.8%

Senior Level

Experienced professionals drive salary averages

90.6%

US-Based

Companies located in United States

96%

Medium Size

Companies with 50-250 employees

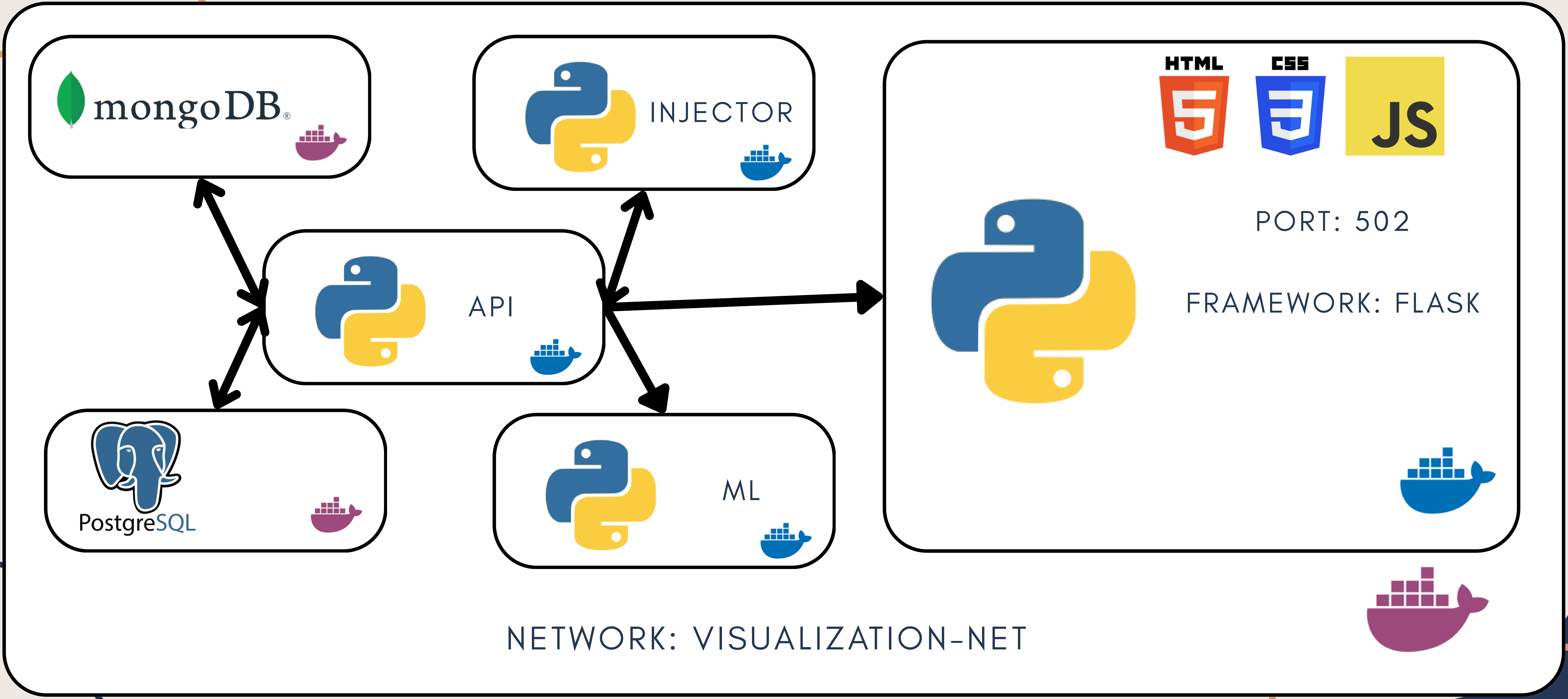
79.4%

On-Site

Traditional in-office arrangements

> Findings represent U.S. tech market within medium-sized businesses - traditional, full-time, in-office employees

MICROSERVICE - WEB



PORT: 502

FRAMEWORK: FLASK

NETWORK: VISUALIZATION-NET

DEMO

Conclusions

The EDA confirmed that the dataset is robust, well-structured, and of high quality, with no missing key values and consistent data types across variables. The unified DataFrame allowed for meaningful insights: users are diverse in age, with strong engagement indicators such as high completion rates and long average viewing sessions. Premium users, Mexico, and Colombia emerged as key engagement drivers, while Smart TVs are the dominant device for viewing. Although variability exists in total watch time due to a small group of “super-users,” their behavior represents a valuable segment for business strategy. Overall, the findings highlight a healthy platform with engaged audiences, clear consumption patterns, and a strong basis for further predictive modeling and strategic decision-making.

Thank you.