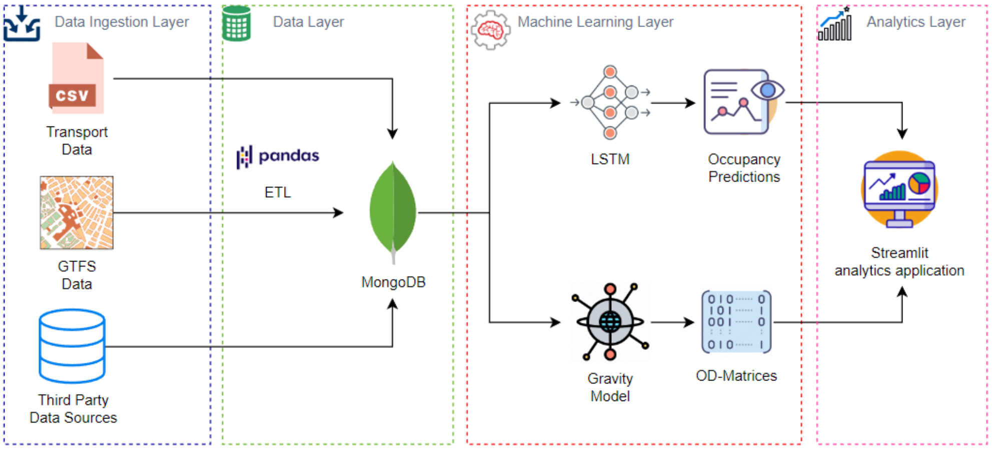
Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarise the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

MOVE-BILBO is a web-based application which allows administrators and transport operators to track and predict passenger demand for bus transportation. The users are able to review occupancy and boardings by bus line and stop, possible transfers between the lines, as well as the Origin-Destination (OD) matrices explaining the number of boardings and exits at stop level. In addition, the application visualises occupancy predictions for the next day, which are based on next stop and next expedition predictions. During the EXPLORE phase we managed to:

1. **Build scalable pipelines** that help us to clean the data and integrate multiple sources of data;
2. Utilise **MongoDB** to store all of the incoming data in unstructured format and all of the intermediate calculations;
3. Construct reliable **OD matrices** that reflect people flows within Bilbao. Those were enhanced with the help of **Gravity Model** based on demographics data and recorded boardings;
4. Trained **Long short-term memory (LSTM) neural network** which predicts the boardings based on previous ticketing data and with the help of probabilistic OD matrices yields **predicted bus occupancy** per line for next stop and next expedition;
5. **Flexible analytics application** which includes all of the KPIs, interactive maps with current and predicted traffic and stops information, as well as a sandbox for simulating occupancy based on transfer algorithms when changes into the system are introduced.



2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

In terms of modelling, our prototype utilises 2 algorithms to yield accurate forecasts for boardings and occupancy:

1. **LSTM** with 4 layers with a dense output layer. The algorithm is used for predicting the next stop and next expedition boardings with respective RMSE of 3.38 and 1.31. The input to the model consists of day of the week, boardings and time of departure. The results from the models are later used to calculate the boardings for the next day. Since we have incomplete information for the exits from the bus we use probabilistic OD matrices to obtain the expected occupancy.
2. **Gravity model** - The gravity model helps explain the flow of the population between two areas as proportional to the population in those areas. This reflects the notion that the greater the population, the more the area has gravitational power and reasons for people to move to. The model we're using is the **constrained gravity model** with the number of people at the origin being known and at the destination - unknown.

For data cleaning, mapping, and storage, we primarily used **Pandas** and **MongoDB**. Given the challenging nature of the data, which was inconsistent and lacked clear business context, we developed scalable pipelines in Pandas to process and clean the unstructured data, making it suitable for modelling and visualisations. Our next step is to incorporate a **PostGIS** database to securely store all GIS and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

ticketing data in a structured format, acting as a middle layer between the source data and the frontend. For the prototype's frontend, we opted for **Streamlit**, an open-source framework that enables rapid development of analytics web applications. Streamlit provides great flexibility and can be easily deployed in new environments.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains and integrate into Data Value Chains (DVC).

The solution we devise follows the **Big Data Value reference model** which ensures flexibility and scalability. In place of the Data layer, multiple open-source components can be used to provide secure and cost-effective solutions. Depending on the needs of the data provider the solution can be easily deployed **on-premise or in cloud** with very little effort.

The current architecture of the solution operates on the **Plug-and-play** principle. If there are changes in the data source, the only effort required is to adjust the ETL pipelines to accurately map the data points. Presently, we have a **robust ETL process**, allowing for seamless scalability and adaptability. To further streamline the process, we plan to incorporate **Airflow**, which will enable us to easily manipulate and monitor the pipelines.

Each of the services we use can be easily containerized with **Docker**, which will ensure a portable and easily manageable solution.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how it will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

For the current prototype, no personal data has been utilised. The only sensitive information we possess consists of ticket ID numbers, which we neither store nor retain. In our setup, we anticipate receiving and storing only aggregated data, and we do not retain any Personally Identifiable Information (PII). Therefore, our solution is **compliant with GDPR** regulations. If the data provider maintains any other personal data and shares it with us, we will utilise **anonymization software** before the ETL process.

In the event that end users require the integration of predictions and analytics into third-party software, we are prepared to facilitate this through a streamlined data transfer via **REST API**. To ensure data security, we will employ **HTTPS and TLS protocols**.

In addition, containerization via **Docker** allows us to isolate and encapsulate the application components, ensuring that the solution operates in a secure and controlled environment.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planned for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment…) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

Data Quality - One critical risk we must tackle is missing data, which can arise from unintentional exclusions due to limited public sources or insufficient knowledge, as well as from delays in data ingestion that can disrupt real-time forecasts. To mitigate this risk, we have identified key groups of metrics that hold significant value for our models and subsequent decision-making processes, such as the number of boardings. Throughout the development and deployment stages, it will be of utmost importance to obtain these metrics effectively. Additionally, we will implement alerting and monitoring processes for these metrics in **Airflow** to ensure the solution's health is continuously monitored.

Model Performance - The proposed solution consists of multiple interconnected algorithms designed to provide the most accurate estimation of bus occupancy. It is important to address any inaccuracies in the estimations to avoid issues such as bias, overfitting, or underfitting of the model. To mitigate inaccurate estimates, we will implement reliable tests and thorough validation processes. For easy tracking of the models accuracy and evaluating the possible bias we plan to use **ML Flow**. Another anticipated challenge is the potential shortage of computational power, particularly for near real-time models, which we plan to address by implementing a suitable retraining schedule to ensure system reliability.

Deployment and testing - During the development of the prototype, we utilised version control in **GitLab**. As we progress to later stages of development, we are prepared to implement a robust CI/CD (Continuous Integration/Continuous Deployment) process and create automated tests.

Internal communication - We follow **Agile methodology** in order to ensure an incremental development process. In order to develop the best possible solution for our clients needs, we plan working closely with the stakeholders and will propose weekly meetings for in-depth discussions and status updates. This approach will enable us to rapidly implement our solution.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981