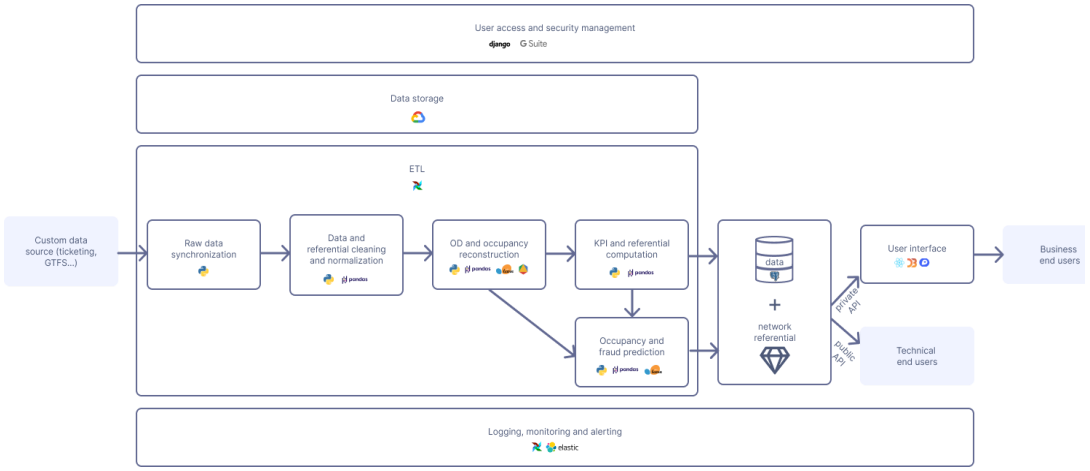
Technical Specification Double-side Page

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.

As a first step, a GTFS and a first set of one year of data (SAE, ticketing) have been sent by the dataprovider via FileSender, a tool certified by the CEA and the French government. They have been stored on a secured Datalake (Gcloud). Based on an audit carried out on these data, appropriate algorithmic treatments have been discussed with the data provider in order to increase the reliability of their data as much as possible. The code to integrate Bilbao's data has been written accordingly, transformed into CITiO's unique format and stored in Gitlab. In parallel, an API has been implemented by Bilbao's technical teams. It allows us to request the required data every morning without manual intervention (they are still stored on Gcloud). Now that the code is industrialised and using Airflow as the workflow management system, the expected KPIs (OD Matrix, Occupancy, Boarding/Alightings) and even more (operational KPIs) are automatically updated in CITiO's tools everyday. The KPIs are presented in graphical and cartographic dashboards. Spatial and temporal filters allow Bilbao to consult occupancy and OD on different levels, as demanded in the challenge: a specific date, a day of the week, a specific time slot, a single line, all the lines, an expedition/course, a vehicle, a stop. The occupancy prediction is also updated daily and its outputs are both returned in an interface that was developed for this project for on field tests and in an API that is meant to fuel the traveller application and the bus stop screens in Bilbao. The daily update of the data (and thus the daily update of the KPIs) depends on the performance of the API implemented by Bilbao's teams.

Our solution runs on a scalable infrastructure powered by Kubernetes on the Google Cloud Infrastructure and we use the ELK stack for logging and monitoring. Here is the full description of the tools that we use to run the proposed solutions:



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.

At the **data cleaning** step, besides other cleaning treatments applied to AVL/SAE and ticketing (management of inconsistent and missing data), careful consideration is given to ticketing location since a wrong location might trigger approximation in further steps of prediction. Several issues in raw data (wrong direction, delocalised validation...) are therefore fixed by **crossing information** from Ticketing and SAE data. This work on Bilbao allowed us to improve our data cleaning codebase.

As for O-D reconstitution, we used a multi-step algorithm: "**trip chaining**" exploits the fact that a smartcard may be validated several times consecutively: the boarding stop of the next validation is considered to be the alighting stop of the previous one. For validations that cannot be chained (single tickets, too long time between validations or stops that are not connected), we use two statistical models: a) the "**user-based model**" uses historical data from each individual smartcard to infer user's habits; b) the "**global statistical model**" uses historical data at a global level to infer travel probabilities and infer the most reliable exit stop. Once our platform has estimated the alighting stop of each trip, it can deduce the exact number of people that are in the bus, namely the occupancy.

Occupancy prediction is based on a **multicriteria average** of the historical occupancies, completed with **seasonality adjustments**. Indeed, this approach has shown to be the most efficient based on a comparative study between

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

several machine learning algorithms evaluated using standard metrics that exist in machine learning: MAE, MAPE, RMSE. The seasonality adjustments have been developed during the REACH experiment phase thanks to Bilbao data. It consists in computing weekly adjustment coefficients based on past biases on the prediction (occupancy over or underestimation).

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 CITIO solution can deal with tens of millions of data points daily by using on-demand clusters orchestrated by Airflow, and a strict division of atomic tasks, allowing the parallelisation of all computations. Smart monitoring allows us to keep track of memory and cpu consumption and adapt the requested resources to the size of the data we receive as input on a daily basis.

CITiO's hexagonal architecture and its flexible and agile development process allowed us to adapt to demands made by the Bilbao team both in terms of the user interface, API representations and specifics processing and algorithm adapted to the bilbobus network.

Our Domain Model (CTFS) is 100% compatible with GTFS and with industry standards allowing integration with other tools and solutions. All API data structures can be exposed using either system (CTFS or GTFS) to ease usage.

As a sign of the adaptability of our front-office interfaces, after evaluating a positive impact on end-users, we took into account changes suggested by the Bilbao Team, and were able to deploy it in production in a matter of days

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 will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Data provided by clients of CITiO are only ever available to said clients and their partners if they so chose. They are stored in cloud buckets provided by Google Cloud in a European (Brussels) datacenter, thus benefiting from their level of security, provided by strict roles and access lists.
Data access is restricted to vetted user accounts through an API exposed under a secure HTTPS connection.
Our front-facing applications are all OWASP compliant.

All data, especially validation data, are anonymized prior to any storage ensuring compliance with GDPR.
All data is deleted at the immediate request of any client and after any contractual obligation ends. In particular we do not use any data from former clients to benefit our algorithm or prediction models.

It is of utmost importance to us that no client can ever access any information about another client through the tools and services we expose. However, through secure channels and in respect with contractual law, we strive to ensure the data can flow freely in-between business partners (eg : between the Metropolitan Authority and the Transport Operator) through interoperable APIs and agile/lean access control.

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.

Engineers as well as the product team are involved at the earliest time to assess feasibility and complexity, plan and execute every phase of the final product, working step by step and delivering incremental releases. Years of experience in the data/transport industry give us the insurance of accurate planning time. Unit testing as well as code linting, integration testing, end-to-end testing, and continuous deployment are executed after each individual contribution is submitted and before any deployment. QA testing is done by the product team on a daily basis in a staging environment. The staging environment is not only a dev environment, it's also where our Ops and Data Analyst teams work daily, ensuring that no problems ever pushed to production and the final client.
If any problem were to happen, all systems can be rolled back in a matter of minutes to a previous, more stable state.

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

Means for accessing the MVP

Please, indicate in 1 page the means for accessing the MVP for a potential customer (login information, website address, link to a demo video or whatever means are needed to check that the MVP exists and works).

A specific URL is created for each new network we work with. A secured system of login ensures that no one but trustful users can access the platform. Bilbao's users can therefore monitor their KPIs by entering their credentials on this page :

https://bilbobus.cit.io/login

In order to ensure our data provider's data privacy, one should get their agreement before we create new access.

The following video shows how the MVP can be accessed and the different functionalities of the solutions :

https://drive.google.com/file/d/1L0cQamjlse3f9_1ZMTS1h7AqdoL_KOPP/view?usp=drive_link

For new potential customers, a demonstration environment based on simulated data is available under this URL : https://demo.cit.io/login. Prospects who request access can therefore navigate on the interface and assess its suitability to their needs.

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