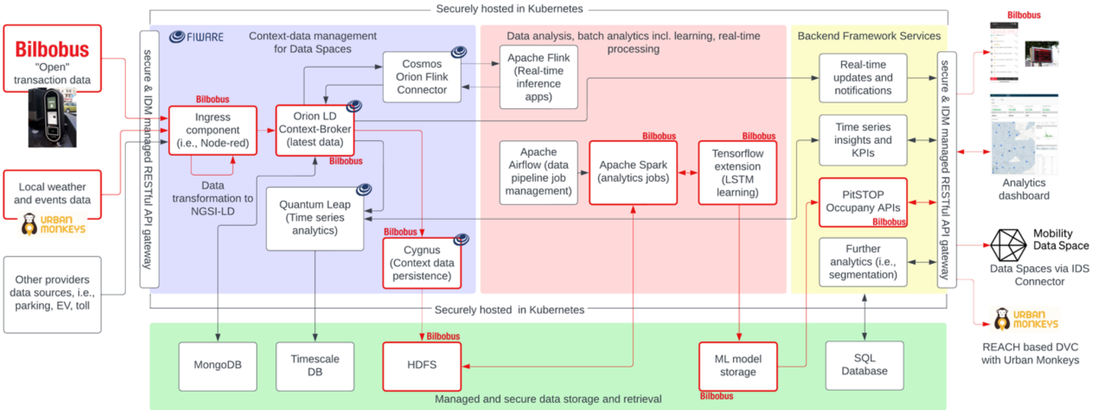
1. Technical Specification Double-side Page

2. **TECHNICAL SCOPE:** Summarize 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.

The Bilbao Bilbobus dataset consists of continuous recorded transaction (ticketing) data from passenger on-entry tap-ins that will be ingested into context-data management using Data Spaces compliant FIWARE components after being transformed and validated into the NGSI-LD format (blue box) and stored into Hadoop HDFS (green box). In the data analysis stage (red box) the solution addresses the challenge because it will help Bilbao to utilise existing transaction historical data in combination with event and weather data to predict occupancy on a route-based level that considers: 1) the "open" nature of the ticket data (no exit taps) using trip chaining, 2) it considers seasonality, weather and event data for learning and predictions, 3) continuous updates of predictions (relearning) based on new data. Finally, the solution provides an external API to integrate occupancy predictions with existing provider apps, digital signage, and 3rd party apps (such as provided by Urban Monkeys) (yellow box). The proposed solution provides transport operators with an additional information tool to encourage public transport usage and optimise demand distribution. Following DVC principles is envisioned to be able to utilise either Eclipse or IDS connectors into the Mobility Data space area (i.e., Mobility Data Space). The data provider can explore agreed KPIs for flexible time ranges in a Journey Analytics Dashboard (right hand side). The solution can be flexibly extended with additional algorithms, such as customer segmentation as developed previously.


Securely hosted in Kubernetes
Bilbobus
"Open" transaction data
Local weather and events data
URBAN MONKEYS
Other providers data sources, i.e., parking, EV, toll
secure & IDM managed RESTful API gateway
FIWARE
Context-data management for Data Spaces
Bilbobus
Ingress component (i.e., Node-red)
Data transformation to NGSI-LD
Orion LD Context-Broker (latest data)
Bilbobus
Cosmos Orion Flink Connector
Quantum Leap (Time series analytics)
Bilbobus
Cygnus (Context data persistence)
Data analysis, batch analytics incl. learning, real-time processing
Apache Flink (Real-time inference apps)
Apache Airflow (data pipeline job management)
Bilbobus
Apache Spark (analytics jobs)
Bilbobus
Tensorflow extension (LSTM learning)
Backend Framework Services
Real-time updates and notifications
Time series insights and KPIs
PitSTOP Occupany APIs
Bilbobus
Further analytics (i.e., segmentation)
secure & IDM managed RESTful API gateway
Bilbobus
Analytics dashboard
Mobility Data Space
Data Spaces via IDS Connector
URBAN MONKEYS
REACH based DVC with Urban Monkeys
Securely hosted in Kubernetes
Bilbobus
MongoDB
Timescale DB
HDFS
ML model storage
Bilbobus
SQL Database
Managed and secure data storage and retrieval

3. **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.

There are two expected project outcomes: O/D matrix inference, and occupancy prediction. For the former we will be utilising the trip-chaining method, as Bilbobus data is collected from an 'open AFC' and alighting stops will need to be inferred. Analysing the unique smart card numbers in the ticket information the system will find future journeys made by the same passenger, under the assumption that passengers will eventually travel back to their first origin destination we can infer alighting stops based on future boardings. We have demonstrated this method with the example 6-day dataset provided, of the 300,000+ ticket sales, we identified 171,693 smart card journeys, of which 87,559 contain more than one trip-leg and so are candidates for use in trip-chaining. This methodology has also been researched with improvements found that leverage machine learning and so provides opportunity for improvement at a later stage. For the second outcome we intend to use an LSTM neural network approach. These systems were specifically designed for time series forecasting problems due to their memory being able to learn long-term patterns, can be easily enriched with other data sources such as weather and events data, and allow us to easily handle splitting data into training, validation, and testing. Further after speaking to the data provider we can use camara based occupancy on a few buses to test for accuracy of the predictions. Another possible candidate approach is Random Forests which have shown good results on similar problems of occupancy prediction and would solve the potential 'overfit' problem LSTMs may encounter, however, are designed for classification rather than time series problems and so may be more complicated to work with. To realise these ideas, we will use Apache Spark to perform nightly batch processing on new ticket data to generate new O/D matrices. Tensorflow will be used for training and invoking our LSTM models through the Keras interface. Finally, we will create a web server that can be accessed via a secured API gateway to allow retrieval of O/D matrices or to invoke occupancy predictions on demand. We will experiment with Data Spaces connectors (like IDS offered by Deusto) to integrate the solution further seamlessly in data marketplaces.

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

4. **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).

Scalability of the solution is ensured through use of scalable open-source components. First the data storage needs to be suitable for continuous growing amount of data. Strategies for short, medium, and long-term data storage will be developed. Data processing based on the developed solution needs to be adequate in terms of processing time vs. velocity of incoming data. Processing of data in clusters, such as Spark will provide an analysis at scale. Most of the computational expensive learning will be delivered through batch processing, i.e., training up to once per day, as described in box 2 and 3. The developed solution will lend itself to be applied to other transport modes, trains, metro, trams with recorded electronic transactions. But we are also planning to train specific models for EV charger and parking occupancy forecasting in the future. Accelogress has also previously developed a parking reservation service called Save-a-Space (save-a-space.com), and more recently for EV charging space reservation (Reserve & Charge, no public presentation yet). This work extended the platform with real-time data analysis capabilities to monitor parking spaces using latest wireless sensor technologies in conjunction with EV charger states to assure advance availability of vehicle charging spaces for EV drivers and real-time operations management. This will be a B2B platform offering into Charging Station Management systems for Charge Point Operators and cities to enable advance reservation on their infrastructure. We plan to extend occupancy prediction to create demand-based pricing schemes for various mobility services, such as parking, EV and public transport (where applicable) that help better managing demand (one example of this is surge pricing as applied by Uber). The solution could be as flexible as to be applied to more generalized optimization use cases, where a demand sensitive resource with limited capacity is offered for a price. We will explore the use of smart data models from FIWARE and / or GTFS real-time occupancy data models, also used as part of smart data models to ensure easy integration into Data Spaces and into Google maps, as desired by the challenge provider.

5. **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).

On the outset ticketing data, or generally electronic payment data is usually provided in an anonymized form that would not allow the link between the actual person and the data using a random user ID (Satisfying GDPR, this is also the case for data from Bilbao). Furthermore, the data will be securely stored in a managed database that is hosted with a data security standards compliant cloud provider (i.e., all major providers, such as AWS, GCP or Azure will have the necessary data compliance). Any access to provided APIs is secured through authentication and data is securely distributed through HTTPS authorized use (i.e., Bilbao apps and display system only). If data should be made available to 3rd party app providers, we would aim to follow DVC principles and use connector implementations for official marketplace data spaces, i.e., the Mobility Data Space (https://mobility-dataspace.eu/). Data should be kept on European servers. Only relevant staff will have access to the data through according two-factor authentications. For EXPERIMENTATION, data could be provided onto HDFS stores of one of the experimentation infrastructures and utilized from there.

6. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planed 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.

Quality of the final product will be ensured through a combined use of an agile software development approach, testing and monitoring of the solution in deployments (both test and production deployments). Features will move from development into a staging environment and be tested based on defined validation targets. Risks per category are: 1) The selected LSTM algorithm does not perform as expected for occupancy prediction. Medium risk. We will ensure that in the early stages algorithms and parameters will be validated with test data sets and learning parameters optimized through various experimentations, and results are communicated with the challenge partner for a common judgement on progress, 2) Time for development takes longer than expected due to unexpected blockers. Medium risk. Agile project management method will ensure that blockers are detected at the earliest point in time and mitigated, 3) Testing delivers unsatisfactory quality results. Low risk. largely mitigated through agile approach and continuous testing, 4) Solution does not scale in deployment. Low risk. Scalability requirements will be assessed with challenge provider and policies will be developed and bottlenecks detected early.

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