Technical Specification Double-side Page

1. **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.

Travel2Fit aims to develop a cloud-based Revenue Management System (RMS) backed with customer demand forecasts and dynamic pricing capabilities for accommodation SMEs and their associated food-beverage facilities. The proposed system, named ROSIE, will be offered as a SaaS platform, with a flexible subscription-based model that its monthly fee is proportional to property size (number of rooms), providing an affordable solution even for micro-SMEs in the sector. The frontend of the platform allows its users (revenue managers) to login using their credentials. Then, they would be able to view regional indices, based on the location of the SME, such as booking probabilities at 10-day-ahead horizon (calendar-form), along with weekday and seasonal analytics. Each user will need to complete his/her profile, e.g. by setting the number of rooms, Average Daily Rate (ADR), and occupancy, thus allowing the RMS to apply an additional personalization logic into its approach. This will "unlock" more tailored outputs such as (i) price suggestions per night and (ii) estimated future KPIs, e.g. ADR, Revenue Per Available Room (RevPAR), and occupancy rates. From a technical perspective, the frontend of the system will be implemented using the Angular framework, while the backend will be based on Apache and Laravel PHP. Algorithms developed in Python will be exposed as REST services using the high-performance FastAPI running on Uvicorn (ASGI web server). The Machine Learning (ML) pipelines will be containerized with Docker, while Cassandra NoSQL will be used for big data stream handling. The system architecture will follow the Big Data Value (BDV) reference model, ensuring that the main concerns of BDV systems are depicted. ROSIE will be piloted at the area of Valencia, Spain, where data will be available from the REACH Data Providers. The system will be optimized for this area, taking into account several other regional factors like the period of the local events, climate conditions, etc. The pipelines will be automated since models will be re-trained every night to contain the current market status, and thus new recommendations will be daily available. Decision support indices along with advanced descriptive and predictive analytics will enable ROSIE's users to boost their operational capacity, offering a valuable data-driven resource & revenue optimisation tool. In order to validate the prototype the following KPIs have been set: KPI1: # of data sources integrated > 5, KPI2: %Increase in monthly revenues > 7%, KPI3: %Decrease in man hours for pricing-related tasks >50%. Architecture diagram is included in the Annex A.

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.

ROSIE aims to leverage on big data and AI to model customer demand in the travel sector. Then, the average room rate will be calculated as a typical dynamic pricing problem. From the available pricing strategies, ROSIE will focus on price discrimination (price is based on customer willingness to pay). Traditional RM systems rely on historical internal data to model demand. Scholars argue though that external data need to be acquired too to improve efficiency, which is what ROSIE will follow as an approach, exploiting among other open value chains, REACH datasets as well. This way, various independent variables can be used for modeling like: (i) property (e.g. location, additional revenue centres, reviews), (ii) room-specific (e.g. room rates, category, type of accommodation), (iii) demand (e.g. seasonality, events, marketing campaigns, reviews, satisfaction, ADR), (iv) customer (e.g. type, preferences, nationality, spending habits), (v) market dynamics (e.g. competition, destination's image). Proper price forecasting requires the application of suitable forecasting methods. ROSIE will examine 3 main approaches: (i) Historical models: Previous work has focused on statistical time series methods, e.g. Exponential Smoothing, ARIMA. ROSIE aims to use such algorithms as baseline and further proceed with multivariate time series analysis by applying a Vector Autoregressive model and Facebook's "Prophet". (ii) Regression models: Previous work has used Random Forests and SVR. State-of-the-art gradient boosting algorithms (e.g. CatBoost, LightGBM, XGBoost) will be exploited for better results. (iii) Deep learning models: ROSIE also aims to use the Long Short-Term Memory model (LSTM). The forecasts feed the mathematical models that produce recommendations for the revenue manager regarding the optimal levels of prices at regional-level and then at accommodation-level (providing his/her appropriate input). It should be noted that many forecasting models may use different demand and booking probability proxies, like online search trending data. Further, literature has shown that instead of training global models, a better performance can be achieved with a separate market-specific model. REACH datasets are expected to contribute towards this direction, providing valuable, region-specific data for Valencia.

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

ROSIE aims to produce a scalable solution that will be able to serve as a new (up-selling) product for Travel2Fit and leverage on heterogeneous data sources. Modern travelers leave digital traces for their stay on social media.

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

This way, tweets and Flickr geotagged photos can be used to understand travelers' preferences and satisfaction. Further, since social media can help to determine destinations of altering popularity, they can reveal trends and thus offer good variables for demand forecasting. More than that, points of interest and climatological characteristics are available for most touristic destinations. Regarding the use of internal data, Travel2Fit has a significant inherent advantage, since its proposal and quoting platform (already in TRL9) contains many attributes of a CRM system. This means that platform data can be used to extract property (e.g. amenities, stars, location), room-specific (e.g. price, type), and customer-related factors (e.g. nationality, age, interests). Regarding data capacity, ROSIE with (i) Apache Cassandra, a highly scalable distributed database designed to handle large amounts of data and providing high availability, and (ii) the use of cloud computing distributed servers, seems to have a solid vehicle for hosting big data workloads. Its modular architecture with "check points" ensures an effective error handling where errors do not propagate between layers. After piloting in Valencia, the aim of ROSIE is to cover many more destinations and support travel SMEs internationally. In order to scale fast and efficiently, tourism destination clusters (i.e. destinations sharing quite similar characteristics and tourism product) will be investigated, extending the initial area. Furthermore, the system will offer transferability potential to other accommodation revenue sources (e.g. food and beverage, spa & fitness facilities) that contribute to SMEs' total revenues. This could be achieved using the innovative transfer learning approach (with scaling, shifting, surrogate data generation, reinforcement learning, etc.). Finally, the provision of tailored services to more target users (e.g. destination managers, policy makers) and target sectors (e.g. agri-tourism, wineries, breweries) is also planned, as the system can also be an effective tourism intelligence solution, impacting wider regional and sectorial actors.

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).

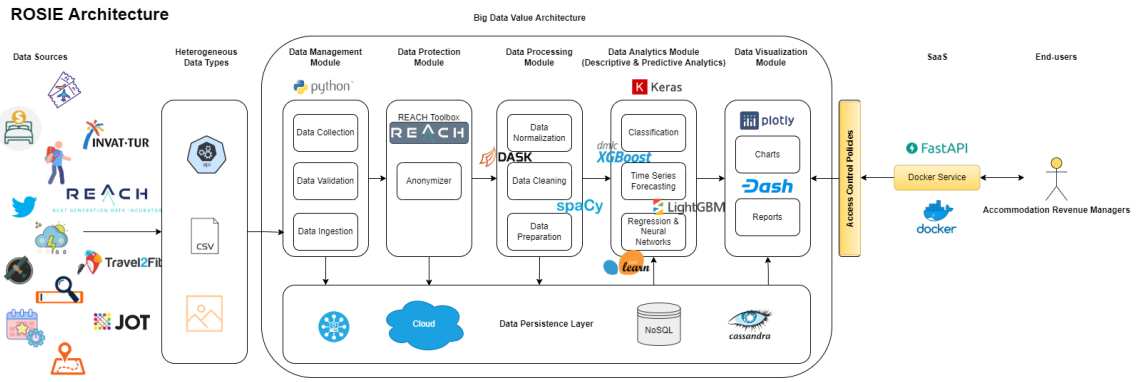
From a legal perspective, ROSIE is built around a "Privacy by Design" approach. Travel2Fit has a strong dedication to GDPR, ethical principles, and EU laws satisfaction, and data used in ROSIE will be compliant accordingly. Regarding data governance, Travel2Fit adheres the core facets of data management for a Responsible AI by implementing: (i) Data discovery: Preserve the privacy of end-user data (already signed Articles 6 & 7 of the GA). (ii) Governance controls: Manage different user roles and multiple Data Providers sensitive data by the use of the Anonymizer (provided by REACH Toolbox). (iii) Policy creation and enforcement: Enforce an authentication protocol (and test it regularly) to prove identities in a secure manner. Raw data will be safely stored at cloud-based servers. Security is enabled through protected access to the cloud platform (SSL encryption) and data transferring via FTPS/SFTP. Every user service will be protected by an authentication gateway, where each one can access only with his/her credentials. Each user owns his data and may at any time contact Travel2Fit to extract or delete all the data associated with him/her in our system. All Travel2Fit's employees will sign a confidentiality agreement (NDA) that prevents them from sharing the information with third parties.

5. **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.

The main data-related challenges are of three kinds: (i) Data sparseness: Most accommodation SMEs do not vary their nightly prices dramatically. As a result, price samples are usually close to the "base" price, which makes price extrapolation difficult. (ii) Sample uniqueness: The uniqueness of hotels makes it hard to generalize the learning. (iii) Feature dependency: Some raw features are price dependent, for example searches are usually negatively correlated with the price feature. Therefore, multiple models may be required to reveal the true demand curve. Proprietary data and models for dynamic pricing are more regularly closed rather than open for empirical research, but ROSIE will try to aggregate and use available open-source big data. Since the dynamic pricing problem depends on multiple factors, flawed assumptions impose risks in feature engineering. This way, business logic and strong pilot feedback and input may be needed in order to ensure genuine client satisfaction. In line with agile principles, regular meetings with the Data Providers and KPI-driven development will help towards this direction. The main algorithm design-related risk is that LSTM may not give satisfactory results. Nevertheless, there are more conventional time series and ML algorithms that can be used as an alternative. Furthermore, models may be time demanding for operational level. In this case, parallel processing and code speed-up techniques will allow a faster execution. Evaluating price suggestions is a non-trivial problem, since there is no ground-truth of "optimal" price. This way, a set of evaluation metrics derived from intuitions will be examined. For example, a bad suggestion is implied, if the night was booked and the suggested price was less than the booked price. Deployed models need to stay accurate at operational level. Thus, a module specialized on automated quality check and model update will be examined (e.g. based on time series anomaly detection).

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

Annex A - Architecture diagram



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