REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

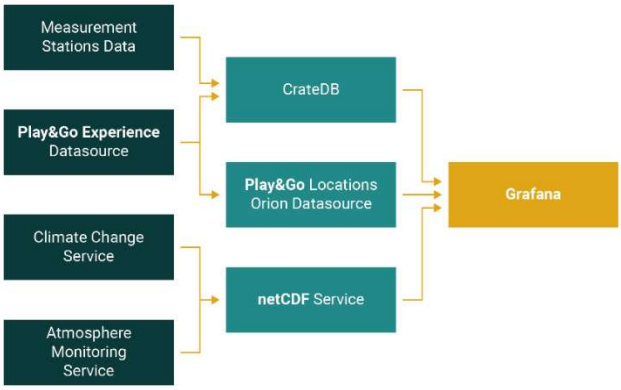
1. **TECHNICAL SCOPE:** The mock-up solution is suitable and correctly addresses the challenge/theme selected over the REACH dataset/s. The Big Data solution architecture proposed is adequate to tackle the data management issues associated to the solution in mind. "To what extent does the applications handle the data provided?".

MovInCity aims to empower public administration with contextualized urban mobility indicators to support decisions and actions to promote urban and sustainable infrastructures and sustainable tourism. The prototype solution would be developed in Valencia because it uses tourist data sets provided by Play&Go Experience, based on assessments of Points of Interest and objectives (check-in, selfie, play, etc.) achieved by citizens and tourists in Valencia, which provides us information about the flow and influx of people in this city. Along with these, it will use different datasets provided by the city of Valencia, thanks to an agreement reached with the Las Naves Foundation, the Visit València Foundation and other datasets provided by the Valencia Administration. The prototype will consist of 4 differentiated dashboards (Public Administration, Business, Tourism and Citizens) in which datasets, visualizations and specific recommendations for each target will be displayed.

This data is complemented by external data such as Copernicus Climate Data Store, Atmosphere Data Store and different measurement stations (CHIMERE) in the area.

The selected data sets are Sentinel 3 (land surface temperature to identify urban heat islands) and Sentinel 5P (air quality, vegetation, NO2, PM10, O3). These Copernicus data sets are obtained in netCDF format for each specific period and specific geographic area; the data is grouped by seasons (time) and regions (space). To these will be added the data collected from the measurement stations; therefore, the data set will have more precision. Together they will show the seasonal impact of climate change and urban heat islands. With the Play&Go Experience data, it will be possible to obtain a correlation between urban mobility and the impact of tourism in Valencia.

All data integration is done with Apache Big Data Stack and Apache NiFi. Finally, the correlations and contextual analysis are integrated and presented using the FIWARE platform and displayed in the Grafana dashboards.



2. **SELECTION OF ALGORITHMS AND TOOLS:** The indicated Data Science approach, i.e. algorithms chosen, and Big Data architecture approach, i.e. tools chosen may successfully accomplish the required data governance, processing and analysis. A clear understanding of the used REACH dataset/s is demonstrated.

The models/algorithms will be based on specific neural networks designed to handle sequence dependence as Long Short-Term Memory networks (LSTMs) Recurrent Neural Network used in deep learning because massive datasets can be successfully trained to identify the different anomalies/events that have increased or reduced the climate change impact. These models will allow mid-term forecasting impact and supporting strategies to encourage sustainability; providing evidence, indicators and models to support the action plans.

Another critical point of the proposed and evaluation of this solution is data quality, which is the main focus of HOPU leading the IEEE P2510 standard about data quality; which is based on data correlation and contextualization.

The AI/ML tools sci-kit learn, and Keras over Tensorflow. Besides, Python and R are used to create the different algorithms, pre-processing and integration of the datasets. For example, some pre-processing actions are to transform the Play&Go Experience dataset into a time-series dataset with geolocation. Due to the lack of geo-coordinates in the records of the Play&Go Experience data source, a new data source is created with the elements grouped by the geolocations. These new elements will have the location of the centre of each zone and the average of their values.

Finally, the data analysis/visualization tool for dashboards creation and business intelligence Grafana (e.g. https://opendata.hopu.eu/), which is used in cities as Madrid, Valencia, Utrera, and will facilitate the integration new services as part of our business proposition for climate change mitigation services.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Explain how the solution copes with the challenge/Theme Challenges requirements and how can it be adapted to other similar problems. What work is still pending to create a real/stable product if any? What TRL level is it in?

The solution will be provided as SaaS (software as a service) to facilitate replicability. Also it offers the scalability from Apache Big Data Stack and flexibility from Open Source tools like Grafana and FIWARE ecosystem. First, the way to integrate datasets is well defined using key components as Apache NiFi and agents in charge of processing the dataset as Apache HBASE (Cygnus), and add new datasets from Copernicus, netCDF or Open Data source.All the Fiware and Apache cloud services use Redhat OpenShift to facilitate the flexibility and scalability due to Dockers.

Grafana can support the creation of new dashboards and offers high flexibility to be able to create any graph to visualize the correlation of the data. The user with an appropriate role can modify, add or delete any panel of a dashboard in a simple way, being able to show the desired data at any time.

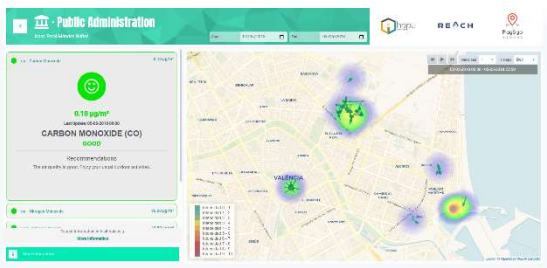
4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Data sharing challenges, data governance and legal compliance, must be observed. Selected methods and technologies to access and manage data assets have to be described. The security level of the solution, i.e. how authentication, authorization policies encryption or other approaches are used to keep data secure, are well explained. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

HOPU has a strong dedication to GDPR satisfaction and experience. The solution will attend the ethical principles and EU laws (Article 23 of the Annotated Model GA), GDPR: Art. 5.1a, 5.2,25,84. Data processing and management plans based on ITU-T regulations where HOPU participates in the Data Management Working Group. The strict control of privacy protection/confidentiality, informed consent, transparency and risk assessment (HOPU has subscribed insurance for civil/data issues).

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Feasible and credible quality process followed for the final product generation. The potential risks in all the phases of the project (design of the solution, development, testing, deployment…) are identified and convincing mitigation plans put in places.

The main risks are related to the ability to identify and to generate the classification on the historical surface temperature (Urban Heat Islands) due to the lack of correlation between the urban mobility data and air quality data, which helps to understand the flow of citizens and tourist within the city and define the action plan or urban mobility to mitigate these challenges of climate change in addition to giving clues to change people's behavior, choosing more optimal routes or suggesting routes in less-visited spots. Initial risks have been related to the accuracy and relevance of the data, which have been resolved using external data sources. Additionally, data quality and model accuracy are crucial, and that's key for HOPU, which presides over the IEEE P2510 standard on data quality for IoT data sources (https://ieee-sensors.org/industry-liaisons), along with the mapping of different data sources (https://www.fiware.org/wp-content/uploads/2020/06/FF_ImpactStories_Hopu.pdf).

Therefore, the project proposes an initial phase to validate the risk reduction hypotheses and correlate the impact of tourism with increases in urban mobility or crowds and indicators such as temperature and air quality in urban areas for identification and space classification. The model will be analyzed, and it will be calculated using a correlation index. The following mockups based on the initially available data set provide the initial correlations and dependencies.



The next phase, one of the most interesting characteristics, would be to include data in real time, both from the Data Provider and those that can be provided by the Public Administration of Valencia in relation to the flow of people, and its subsequent filtering in the visualization.

REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS