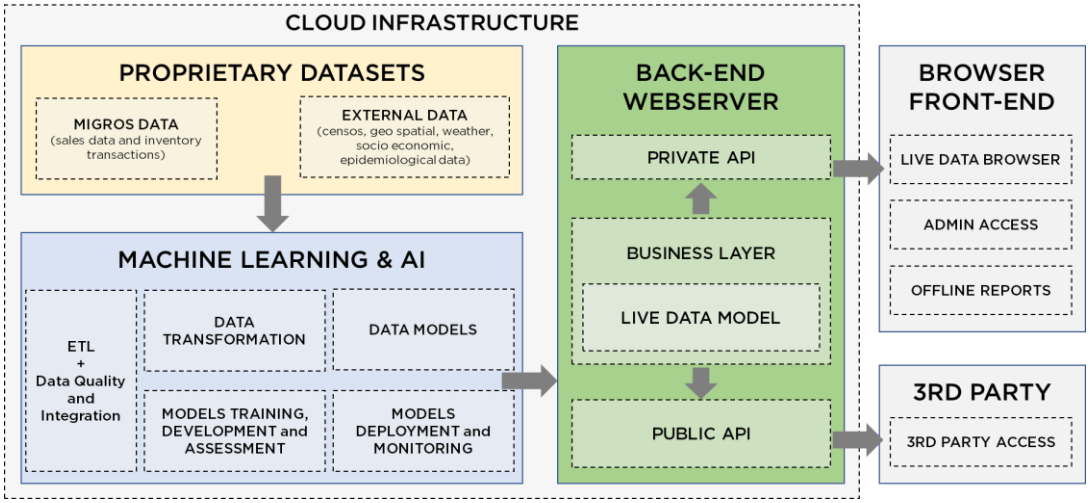
Technical Specification Double-side Page

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.

The proposed solution is a visual data browsing platform for retail store managers and decision makers that uses business data and a game-based visual language, built on a state-of-the-art forecasting engine, enriched with estimate explanation and what-if scenario simulation. **Migros'** sales data and inventory transactions data will be enriched with other external datasets which increase the human relevance in identifying purchase drivers and geographical relevance, such as population census, geo spatial data, socio and economic information, weather forecasts, and epidemiological data.



We will be using an AWS based pipeline flow, using Docker's containers, Apache Spark (AWS EMR), AWS ECR, AWS ELB, AWS RDS and S3, AWS Glue, AWS Athena and SageMaker.

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.

Several models will be trained, assessed and compared. For the different outputs (forecast, explanation, simulation) different models will be required.

Random Forest needs data transformation to account for time correlations. Together with **ARIMA** it will be the benchmark to other models.

Data hungry **Recurrent Neural Networks** usually have an advantage on the estimates. With more parameters than a normal NN they have an added risk of overfitting and are more time consuming to train. This requires added effort in cross-validation.

Bayesian Hierarchical is the best model to help explain effects, and more importantly, it gives the cloud of probability of each point and not only a single point estimate. The biggest drawback? The explosion of training time. A good business knowledge is paramount to manage variable interaction, non-linearity, and each parameters' priors.

XGBoost is a fast algorithm with very good estimates, having many hyperparameters to adjust by Bayesian optimization. **LightGBM** is similar to XGBoost but faster and more prone to overfitting.

Shapley additive explanations algorithm will be used to measure the factor contribution of a blackbox such as XGBoost.

With the selected model, a prediction is made for each combination of variables that intercepts the universe of business context, to get the one that maximizes the business objective.

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

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

The external data sources will grow only marginally, while the internal Migros data sources will grow much faster. With 2 000 stores, the daily increase in the transactional data could reach millions of new rows. Datasets will be structured by relevance and age, for storage and modeling. Aggregation and partitioning will be used on training models. Incremental training will be required.

The architecture described in 1. was designed with scalability in mind, AWS ELB manages scalability as required. As the retail network increases, data is easily incorporated on the pipeline, as it will have the same format and variables. The increased volume will be incremental on the previous approach.

If older data is tested as not bringing signal to models, then older time frames will be frozen.

Model's training will be split based on output, with Bayesian and XGBoost models trained incrementally. Regarding flexibility, by default, we are addressing several business scenarios and needs, with proper output parametrization new features can be rapidly deployed. Added analyses or outputs are feasibly implemented.

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).

All data will be encrypted while rested and in transit. Communications between systems will require SSL/TLS. Frontend dashboards will require password hashing, 2 step authentication and single sign-on. Access to dev and prod environments, databases and visualizations are independent from each other. Web access is done using HTTPS. Most relevant issues of data harvesting and integration are already solved by our Data Provider, as they already are GDPR compliant. We will focus on assuring that only the right data is accessible by the right user, either by aggregation (regional focus, or C-level user) or anonymization (not disclosing competitors' data), by implementing an IdAM framework.

We will implement continuous monitoring and auditing of these processes. We will work towards ISO 27001 certification.

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.

Data quality from our Data Provider is partially assured: data from sales has a high-quality standard, but one of the original issues is in fact the quality of the data related to shipments and warehouses.

On the initial proposal, we listed some risks based on the sample data that could impair the project's goals, among those, the low quality of the data. Those risks and eventual changes on the proposed solution and outputs, must be and can only be addressed directly with our Data Provider (the initial stock and sell-in data, for example). Nevertheless, we are already designing strategies to be proposed and discussed with our Data Provider, correlating sales with warehouse data to level out outliers.

External data sources have also their own quality assurance methodologies. The major data risk is in the integration phase where data from different sources and different levels of aggregation must be properly addressed. A four hands programming approach will be used in all data pipelines. GitHub will be used for version control and collaboration. Different environments for development and production will be implemented.

Model's training resources are already referred to in point 2. Model's accuracy is maintained and improved through continuous monitoring and re-training. We have a dedicated team of data scientists and machine learning engineers to improve on current models, adapt to business needs and implement future algorithms as science advance.

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