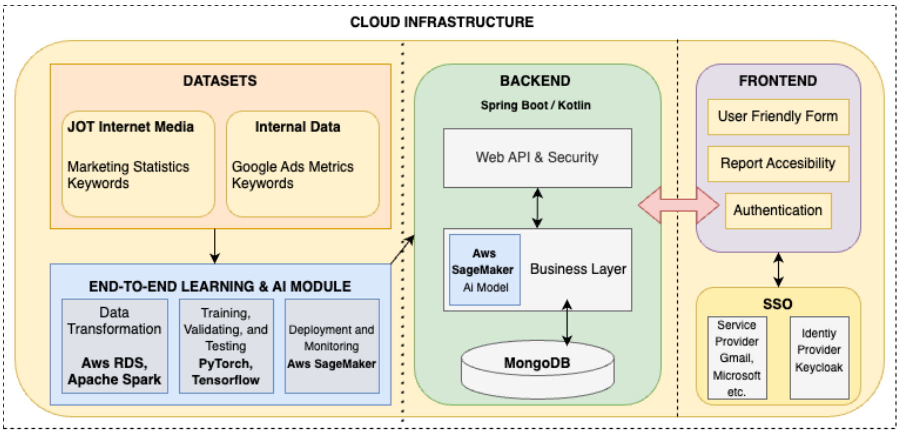
Technical Specification Double-side Page

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 recommendation system that can predict keywords temporal patterns and category behaviour for companies and agencies that uses digital advertisement. The solution will be built on advanced end-to-end learning architecture, enriched with efficient CTR and click predictions, and estimate explanations. **Jot Internet Media's** marketing statistics and campaigns data will be enriched with our internal datasets which will be used to increase the accuracy of our deep model. We will be using Adsbot's Google ads data as **Internal Data.** Thus, a richer data pool will be created in terms of language and country. As a result, accurate selected metrics (click, ctr e.g) will show the implicit feature interactions behind the user query and click behaviours. Results of our model will be displayed as the relevance in identifying datetimes, geographical, country, language economical information.



In order to build the backend system, we will be using Kotlin as programming language Spring framework, Kotlin and MongoDB, whereas React JS will being used for the frontend module. In machine learning layer python as programming language, keras, and pytorch, tensorflow as libraries will being used.

We will be using an AWS based pipeline flow, using Docker's containers and these containers will being orchestrated by Kubernetes. Tools to be used as cloud infrastructure are 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, validated, and compared. Comparison will be done using the accuracy result of the used networks. We will also consider the benchmark results of state-of-the-art networks. Moreover, the methods will be used as **regression algorithms** & create the function for predicting patterns for given Date/Time, Country, Language, Category, and other Campaign info, we will train an ML model which estimate the click, CTR or any other metric using:

A Tree based methods: XGBoost, Random Forest, CatBoost generally used for tabular data learning. Selects global features with the most statistical information gain. To improve the performance, dimensionality reduction of the input data is required. Otherwise, methods will suffer from the curse of the dimensionality.

A Neural Network based methods: Fully Connected Neural Network, TabNet are efficiently encode multiple data types, alleviating the need for feature engineering. Unlike Tree based methods, NN based methods are not domain specific.

Then, using estimated click, CTR and other metrics, with user preferences, find the coefficients (M, N, …) which optimizes following equation:

$$argmax(M\hat{y}_{Click} + \mathrm{N}\hat{y}_{CTR} + \dots)$$

To find the optimum coefficients, a linear programming optimization will be used. For instance, the cost of the campaign might be more important than impression for a customer, then the coefficient of the cost metric must be negative and very low while the coefficient of impression is + 1.

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

The outcome of these function is a **discrete optimization** method of patterns of keywords & categories behaviour based on week, day & time.

ML-ToAST needs to set the configurable parameters for using Google BigQuery. The biggest advantage is the support of the multilingual keywords. ML-Toast uses UMAP for dimensionality reduction and unsupervised learning method (K-means) for labelling the data. This method will be used for **clustering patterns of keywords**.

Since we utilize structured data, incorporating manual features will enhance the accuracy of the models. Additional features (e.g:seasonality, public holidays) will be added during **feature engineering**.

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 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.

Apache Spark (AWS EMR), AWS ECR, AWS ELB, AWS RDS and S3, AWS Glue, AWS Athena and SageMaker are designed to provide a robust, scalable infrastructure. They can automatically manage the underlying resources based on the requirements of your algorithms. This allows your algorithms to perform at their best, irrespective of the size of the workload.The cost of using AWS for big data increases logarithmically, not linearly or exponentially , we will optimize costs by leveraging AWS Spot pricing.

Model's training will be split based on output, with XGBoost, RandomForest, CatBoost models trained incrementally. We can use the traditional data partitioning; 80% training or 20% testing. However, with the increasing computational power we can train our model in the cloud like in Google Colab, Amazon Cloud etc. Training neural network may require additional techniques like dropout, changing the activating functions or layers to increase accuracy. Regarding flexibility, we can use or modify the state-of-the-art networks, like TabNet and ML-ToAST.

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).

Systems will require https, SSL/TLS protocols for the communication. Accessing to the web framework will be done by the HTTPS. Accessing to the private and restricted environments such as development and database will require 2 step authentication and single sign-on. Data harvesting and integration are not an issue in our current system. We will focus on assuring that only the right data is accessible by the right user by implementing an IdAM framework.
Moreover, we will temporarily save the customer data that Jot **Internet Media** sends as an input data in our platform as an anonymous data; so does our internal data. We will implement continuous monitoring and auditing of these processes. We will work towards ISO 27001 certification.
It is also important to note that, following the EU's GDPR, no personal, identifiable information will be shared

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 text dependent models/canonical deep tabular data learning model that we will train for keywords require multilingual implementation. Due to the multilanguage challenge, different tuning and testing will be required for different models. We will train our model with the English, Turkish and German languages separately. Then we will validate and test our models. If the accuracy of our model satisfactory we will save it as pretrained model.

GitHub will be used for version control and collaboration. Model's training resources are already referred to in point 2. If the selected model is not accurately performed on our test set, then we will apply the standard techniques for continuous training. Our team will adapt the state-of-the-art techniques to our business model. **Mean square error** for Neural Network based & Tree based methods, **Spearman's rank correlation coefficient** for discrete optimization method will be used for benchmark& accuracy comparison. **Rand score** will be used for ML-Toast algorithms.

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