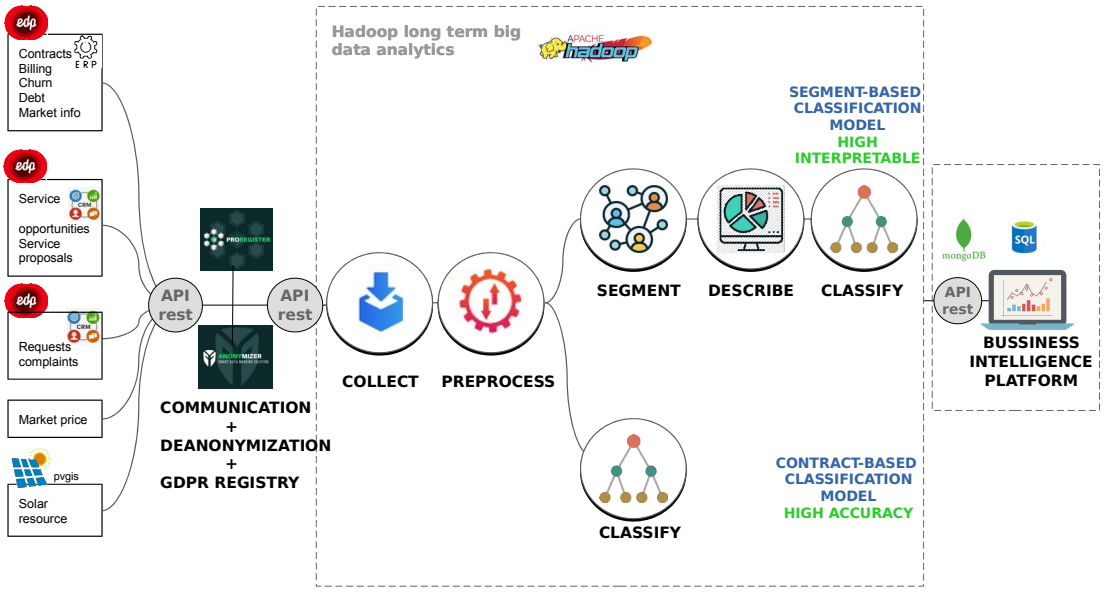
Technical Specification Double-side Page

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.

Beedata offers a customer analysis toolkit formed by a customer segmentation model and a behavioural model to predict churn and energy service acquisition. The proposed service is based on two different approaches or sub-models with different goals and time of use : a) a human-interpretable model based on customer segmentation to describe segment attributes and its impact on churn or service acquisition, and b) an automated high accuracy model to calculate churn and service acquisition probability per each contract. The results of model a) will be displayed on the EDP Celonis platform and the results of model b) will be presented on the EDP CRM platform. Model a) dashboard will resemble the smart business mock-up created by the following two web pages: https://marvelapp.com/prototype/6h94h37/screen/86734281/ https://marvelapp.com/prototype/6h94h37/screen/86734281/layer/157991376). In both models a) and b) the results of the model will be pushed to EDP via a secure API, or agreed channel, and integrated into customer analysis and attention processes. Service data is obtained from EDP's, ERP & CRM platforms via secure API, energy market price data is obtained from the Portuguese market operator (OMIP) and meteo data from the PVGIS database. Once data is obtained, deanonymized, and GDPR registered, are done to have data ready for decentralised big data efficient storage. The first step in analytics is data preprocessing to improve data quality. The steps that follow are dependent on the model. In model a), customer segmentation and segment prediction models are created. In model b), the creation of a single customer churn/service acquisition model is explored. The main innovations we introduced in this project are :
1) Combining customer behaviour interpretation and high-accuracy behaviour prediction in a single service. With a human-easy-to-interpret non-blackbox model, we provide insights to allow the company departments to understand their customers' behaviour via segmentation. With a high accuracy blackbox model, we provide contract specific behaviour prediction.
2) Using big data analytics to support analytics on EDP requirements (> 50K clients), providing simplified and meaningful insights via BI or CRM interface.

.



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.

Data fetching via secure API is followed by deanonymyzation and GDPR registry, which will be executed with Deanonymizer and Proregister tools supplied by be-ys research. Data exploration and preparation is done on the EDP dataset, which is formed by 9 csv files related to billing, contracts, churn, service acquisition, debt, and complaints that add up to a total of 538 data fields per each service point. These fields are complemented with the monthly consumption (12 data values) and the hourly consumption (at least 8760 data values) per service point. Each client can have different service points, and the total number of clients, including existing contracts and those that left EDP, will be around 39400 clients (93405 service points). The NIPC (tax identification number) will be used as the common ID between all datasets, since it is the customer identifier. All of this data will be stored

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

and processed by Beedata's Hadoop architecture in a cloud offered by OVH company. During the 4 month project, we'll test the possibility of installing our system in the DEUSTO big data services. We'll use their Hadoop environment tools. Airflow and also the IDS Connector. Data preprocessing is done on the prepared data in order to improve quality by removing outliers (baseline model based, etc.) and normalising features (domain specific). Additional features are calculated to improve description and prediction: intervals of dates (i.e., num. of days of the contract), weather dependency (naïve linear weather dependency model) or solar potential (naïve solar potential model) are calculated to improve segment description. The next step in model a) is customer segmentation per each reference tariff via k-prototype clustering to support mixed type of attributes. ANOVA (chi square) analysis is used to check whether and which numeric attributes are statistically significant across clusters. After that, a highly interpretable prediction model is obtained per each segment via a CHAID decision tree to interpret and predict each segment's behaviour. The accuracy is obtained via PR AUC indicator. The clients that left EDP services and the list of complementary services that EDP is offering (solar panels, electric mobility, power factor corrector, energy certifications, efficient lighting, and voltage level increase service) will be used to calibrate the first model of churn and service acquisition prediction. The next step in model b) is a high accuracy classification model obtained per each contract via XGBOOST to predict each contract's behaviour. The evaluation of models will follow two levels of analysis: i) evaluation metrics of related churn segments. ii) Binary classification performance as measured by the F1-score Different classification evaluation, but the same method, will be done per use case (churn and service acquisition). The common approach in both use cases is to make it easy to implement and to increase calculation speed . Hadoop with python libraries (numpy, pandas, scipy, scikit-learn, kmodes, CHAID, and xgboost) will be used.

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

We use Hadoop as the dig data system because is designed for store the data on any number of data nodes that is fault-tolerant. We use MapReduce to perform the parallel processing required by the different analytics processes, while HDFS to perform the distributed computing, and YARN as the responsible for managing and monitoring workloads. We we'll use the beedata lab test cluster which is formed by. On top of HDFS we built Hive which is used to analyze structured data, and automatically translates queries into MapReduce jobs. so that, we can skip the requirement of the traditional approach of writing complex MapReduce programs, and increase scalability. We also built HBase to maintain the file system namespace, making it a scalable replacement for the HDFS's NameNode. With this approach, files and directories become rows of a very large HBase table representing the entire file system. Finally we use Ambari to manage, monitors, and provisions the health of Hadoop clusters. This configuration of the Hadoop ecosystem allows us high escalability possibilities. The analytics modules to be used are Python libraries (Pandas, SciPy, NumPy) that enables other alternative of R and permits to optimise the calculations and reduce the data processing time. In terms of data value chain, the solution we propose is addressed to 3 different users in a company, a)business developers, b) client attention teams, and c)commercial and sales teams, which will use the solution in different manners; from a business intelligence platform to an alarm for preventing churn of a client. In terms of other sectors, it will be e easily adaptable to any other utility related domains, with the same needs, such as water utilities, telecom, electric and gas distribution, energy service companies, electricity flexibility and aggregation companies, and facility management companies. These utilities operate in contexts where customers can change the services company or can ask for other related services and the churn prediction and service acquisition models are very useful. At the same time, since the raw data of the project is mainly based on typical CRM and ERP data related to client attention and complaints resolution, on regular bills, monthly and/or hourly/daily time series, and on typical tariff and costs schemes similar in all of these domains, the data processing is easy to adapt. Finally, the clustering and classification techniques implemented in the solution, can be trained with different variables than the original ones, and also give us the probability of churn an service acquisition according to other different parameters.

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).

We use OVH as a hosting provider, and according to GDPR is defined as a "processor". OVH offers a guarantee of security and privacy which is documented in OVH General Terms and Conditions of Service, and can be summarized as follow: I) processing personal or anonymized data solely for the purposes of carrying out the services correctly: OVH will never process your information for any other purposes (marketing, etc.), ii) Keeping your data inside the EU and only in countries recognised by the European Union as offering a sufficient degree of protection (cloud services, which is not the case), iii) no services involving any access to data you have stored as part of the service subcontracted

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

outside the OVH Group. In addition to that all communication with datasets and with BI for the results will be done vi secure and encrypted APIs. In case of other ways os communication of data will be used, all of them will incorporate SSL protocol for user and password sending, X.509 certificate in server and client, and OpenAM: identification and access management platform. On top of that, as we'll use the Anonymizer and Proregister software for data adquisition and for results offering, all security requirements and GDPR standards will be achieved. The reference architecture will be secure by accomplishing the specifications of secure-by-design (SbD) and secure by default approach.

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 RISKS: i) Amount of data: not enough training data. Solution: Ask EDP for more data or apply data augmentation techniques, ii) Non representative training data of service acquisition or churn evaluation. Solution: Ask EDP for more data. Iii) Imbalanced data: Apply data augmentation techniques iv) Missing documentation of current datasets provided by EDP. Solution: Identify whether they're supposed to be important for clustering or prediction during data exploration. v) Poor quality of data (outliers ,..) . Solution: feature preprocessing and feature selection – domain specific. vi) meter poor quality of data (outliers, ..). Solution: feature preprocessing and baseline based outlier detection and imputation vii) Irrelevant features. Solution: feature preprocessing and feature selection in data exploration to reduce the amount of features.
MODEL LIMITATION RISKS: I) Under fitting the training data. Solution: Maximize the training time, enhance the complexity of the model, add more features to the data, reduce regular parameters . ii) Over fitting the training data. Solution: Gather more data or use data augmentation technique, and review data errors in the training. Iii) Imbalanced data: Pick models and metrics prepared to manage imbalanced data (PR AUC, xgboost, ..) iv) Different accuracy and results between models a) and b) Solution: We expected different results between models but the goal is also different (model a is interpretable, model b is accuracy). In case of significant difference model a) attribute could be adjusted or weighted considering model b) attribute importance analysis.
PERFORMANCE RISKS: I) Low performance in computation. Solution: As using hadoop platform it would be easy to increase computing nodes to fulfil computing requirements. We willadd/remove nodes on-demand using cluster manager Ambari.
INTEGRATION RISKS: I) API not available. Solution: Implementation of specific database import/export protocols or provide CSV files with data.
SECURITY RISKS: I) Security vulnerabilities in provider. Follow cluster provider notifications in order to identify potential security issues asap ii) Security vulnerabilities in the software. Follow software provider and community notifications in order to identify potential security issues asap iii) System security issues. Enable cluster provided security tools (DDoS attack, firewall configuration, etc) in order to limit input/output traffic to the one required for the service. iv) EDP-Beedata Interception of communications. Use available secure connections via SSL connections between EDP and beedata. v) Beedata internals interception of communications. Use VPN between nodes in the hadoop cluster

The main criteria to define the quality of results of the project will be the accuracy of the 2 prediction/classification models, and because of that, all control and measurement of accuracy will be done based on techniques showed before, and at least 2 iterative process of accuracy validation against real customers that left EDP, and real customers that bought new services, will be implemented. At the same time, within these iteration processes, usage validation from the different stakeholder will be done.

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