REACH

NEXT GENERATION DATA INCUBATOR

EXPLORE PHASE

TECHNICAL SPECIFICATIONS

11/05/2023

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

CORE PARTNERS

FGS brpx cea ESTBAN zabala innovation consulting Systematic Paris Region Digital Ecosystem gnúbila Deusto Universidad de Deusto University of Deusto CERTH ITI INVESTIGATE TO INNOVATE

DATA PROVIDERS

YapıKredi Teknoloji vrt be|almerys Bizkaia Play&go experience JOT MİGROS TİCARET A.Ş. idea75 SONAE MC

1 ANNEX I. Technical Specification Double-side Page

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?”

The proposed solution (Visight) comprises a monitoring and maintenance platform and a data acquisition platform that allows (i) asset management and (ii) real-time asset monitoring, analytics, and predictive maintenance. The backend system of Visight is based on ASP.net. The database and data model are based on the PostgreSQL. The frontend system consists of a web application (Angular), a mobile application (Xamarin), and an application for smart glasses (Xamarin). The platform’s main modules are the Assets, Maintenance, Analytics, and Work orders modules. The Assets module allows an efficient and centralized overview of the available assets (i.e. machines). The maintenance module’s main functionalities are implemented through the Data Acquisition, Execution, and Diagnostic tool which takes machine information gathered from smart sensors and sends it to the DataLake for processing and analysis. Based on the processing of data, work orders are created. The smart glasses module allows for faster information flow and better access to centralized documentation.

The Data Acquisition, Execution and Diagnostic tool presents a separate platform that is embedded in the Maintenance module of the Visight platform. The architecture of the application is designed with a focus on the genericity of the data model and scalability of the solution, in terms of the data quantity as well as new instances of databases. Services needed for functioning are separated into three layers: API layer (DataImport), Service layer (Azure Kubernetes Service, Azure AD), and Data Layer (PostgreSQL Azure Service, Data Lake).

Data collected from sensors are sent through generic endpoint to ESB Talend where they are transformed. The data are sent to the Redis client and DataLake for storage. Data are sent through Redis client to DataImport where they are validated and alarms are created in the case of deviation from set KPIs and data are saved in the Postgre TimescaleDB, with adjustable data retention time. Data are also sent to the DataLake in a raw format where they are transformed to parquet format within Spark tool that allows for complex data transformation, analysis as well as the import of developed AI algorithms in Python. ML algorithm which is part of the platform will be developed in this proposal based on the available dataset and configured in the Spark tool. After the data is processed, maintenance work orders will be created and sent to the asset management platform as a work order (notification). An alarming system allows that in the case of uncontrolled deviations from the predicted limit values the machines will be allowed to preventively shut down, allowing for services and quick reaction minimizing the potential damage. Integration of Grafana allows for data visualization and Datawarehouse (DWH) allows the creation of reports for business analyses.

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.

To successfully tackle the proposed challenge of Production optimization with the dataset provided by IDEA, the provided dataset will be used to do preliminary studies in terms of the selection of AI approach. For predictive maintenance, the classification approach and regression approach will be studied and different models for both approaches will be tested. The first is used to predict whether there is a possibility of failure in the next n-steps and the second, predicts remaining useful life (RUL) – how much time is left before the next failure. Firstly, algorithms will be developed and deployed in the Spark tool with the use of Python using data from DataLake. Docker containers are run in a Kubernetes cluster and can be scaled up if necessary. Different AI algorithms/approaches will be deployed to select the most suitable one. Upon connecting the machine/sensor, the AI will be validated and tested on real-time data. Modeling will be implemented using Spark's machine-learning library. The proposed dataset will be tested; however, it is expected that additional data will be needed to properly train AI. Furthermore, there is a need to clearly define threshold values and KPIs (key performance indicators) to successfully establish predictive maintenance (alarming in the case of deviations). The project utilizes a Big Data approach and employs Big Data infrastructure. The platform is hosted on the Microsoft Azure cloud. Data sent from the machines (sensors) will be sent to Redis where it will be stored temporarily, and the task queue is built. The Data import is responsible for de-queuing of the data. The service is run in Kubernetes allowing clusterization thus providing scalability as new instances can be added if necessary. A denormalized data model will be used (the same data model for all time series) providing genericity and allowing for configuration of multiple types of machines and several instances of machines. Spark tool allows processing and data analysis that can be visualized in Grafana. Datawarehouse (DWH) allows reporting and business analysis realizing data value chain.

3. **TECHNICAL SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** The solution can truly cope with humongous and increasing datasets, potentially from diverse data providers, and is flexible it to adapt to other related domains.

Visight asset management platform and Data Analytics platform that is integrated can be used jointly or separately. The platform was designed with flexibility in mind. The solution's data model is generic and offers great scalability which is supported by Spark Tool and clustering. Scalability is achieved through horizontal scaling. The computing-intensive components are packed in docker containers that can be deployed on-demand and run parallel to successfully cope with high loads. Scalability is ensured using microservices approach based on Docker technology which communicates using REST API protocols for data transmission. Docker containers run in a Kubernetes cluster allowing the increase of processing powers by adding more nodes if necessary.

Four ensuring flexibility, Azure cloud infrastructure is used as a service orchestrator and the main entry point to the frontend. This allows us to separate the processing and extraction layers and scale them based on demand. Data interoperability is handled differently based on the layer. The top layer, where external data integration is handled, communicates using open data formats such as JSON. The cloud architecture provides a way to manage time series data and can be configured for the data sets from the industry or energy sector ensuring flexibility. Certain customization and adaptation are needed, but the genericity of the data model allows configuration regardless of the sector or the data that needs to be stored. Furthermore, the companies that already have asset management platforms can integrate only the Data Acquisition, Execution and Diagnostic tool.

For flexibility of data visualization Grafana tool is used, as the user with the appropriate rights can create graphs to visualize the data based on their specific needs, ensuring flexibility of the data visualization.

To ensure data governance NDA will be signed with Data Provider to ensure that the company's data is not shared with any 3rd party. These datasets do not include any personal information, only data about machine states. All data will be safely stored at Inden's cloud-based servers and all Inden's data analysts and software developers will sign the confidentiality agreement that prevents them from sharing the classified information with a third party. Furthermore, raw data processing will be conducted only within the production system allowing only authorized users to access and process the secured data. Security is enabled through protected access to the administration interface with SSL encryption via https protocol and transferring data (up, and downloads) via FTPS/SFTP. Furthermore, to ensure the security of data, SSL encryption protocols will be used as well as Microsoft Azure AD for authentication and authorization. Two-factor authentication is applied for important entries (e.g. cloud login). In terms of secure data sharing, available tools from REACH Toolbox will be assessed to see the possibility of applying them in our solution like Data sharing platform solution. API token authentication will be used to limit access to data transmission.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Data sharing challenges, data governance and legal compliance, must be observed. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

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 place.

The accuracy and reliability of predictive maintenance can only be as good as the data it is built on. The biggest risks we foresee is data-related risks. Currently, the available amount of data is not adequate to develop an accurate AI algorithm. It is necessary to work with DP to gain access to a larger dataset that will be pivotal for effective model training (access to historical data). KPI and limit threshold values need to be defined to successfully implement alarming and thus predictive maintenance with the automatization of work orders. Our team follows structured project management where clear assignments of responsibilities and roles are defined. For effective planning we use Redmine, and complex tasks are split into smaller, more manageable tasks. The work is done in sprints and iteratively to produce MVP and continuously improve it based on the feedback received. To ensure good cooperation and follow the work regular meetings with mentors/DP are planned.

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