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

DATA PROVIDERS

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?"

To address the Cross-Selling Challenge, we have executed an MVP, in line with our submitted proposal. It is capable of assimilating buying patterns based on consumers' purchase tickets and recommending cross-selling products based on similar entities' transaction history. The architecture adequately tackles the data management issues associated to this setup and follows the steps below:

1. The necessary data infrastructure was built, that allows for timely and consistent ingestion of data. A flexible set of REST APIs has been constructed that feeds a layer of tables in a designated data warehouse, where the data is stored
2. A daily routine script spawns a cluster where a Collaborative Filtering Machine Learning Algorithm runs and calculates the relevant cross-selling recommendations on updated data
3. After the calculations, the output is stored in a PostgreSQL instance where it is ready to be served by another layer of REST APIs to all consuming apps (OFs, Farmanager Portal, Aqurate Web App, etc)
4. To showcase the recommendations and monitor business KPIs that ensure the fulfilment of the objectives, a dedicated dashboard was built in our Aqurate Application. COFARES has easy access to the cross-selling recommendations at a pharmacy level and can monitor the volume and percentage of items recommended and purchased
5. Beside the monitoring of the business KPIs, CI/CD procedures were implemented, to ensure the monitoring of data quality throughout the Data Value Chain (DVC). These routines investigate distributional aspects of the data from ingestion to output and are centralized in Grafana for a complete overview.

The mock-up provides an end-to-end functional solution.

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 address the challenge, we have implemented a Collaborative Filtering (CF) Algorithm that can identify items that some consumers have bought and recommend them to similar, relevant users that haven't. This novel approach is superior in the cross-selling context to traditional Content-Based Filtering that only relies on item attributes for recommendations, as it is able to bring forward to OFs relevant items for their consumers that are in accordance with market demand.

Based on the sample data, enriched with simulations to account for multiple pharmacies and customers, we have implemented a non-parametrical model to account for the interactions between the pharmacies' customers and their items. The interactions stemming from consumers' tickets are first quantified within an Interaction Matrix that depicts these interdependecies. Based on the matrix, we calculate Jaccard metric distances that proxy for similarity in buying behaviour between pharmacies. We then recommend items for a pharmacy based on this proximity, taking into account the items that have already been bought. This approach is in line with described desired outcomes of the challenge. During the Experiment Phase, the full dataset will allow us to investigate further CF algorithms like Matrix Factorization, Factorization Machines or Neural Networks. We then choose the best performing model based on the models' performance with respect to recommender system metrics like MAE, RMSE, Hit Ratio, etc.

A mix of relevant, leading technologies was used to construct the DVC. The REST APIs were built using Fast API (Python Framework). The Data Warehouse relies on BigQuery, part of the Google Cloud Platform (GCP) for storage. This technology is also mentioned by COFARES as preference in the challenge description. The Compute Engine is also part of the GCP and the whole cloud orchestration is done through Airflow. The Relational Database that stores the recommendations is based on PostgreSQL. The Web App that contains the dashboard was built using the React framework in JavaScript. The CI/CD monitoring and alert procedures are implemented using GitLab, Grafana and Sentry.

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

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.

All parts of our solution have been designed to cope with the challenges of Big Data. The REST APIs have batch functionality (through Pagination), allowing them to receive and deliver sizeable amounts of data in a timely manner. We have also applied filtering for the fast retrieval of specific requests. BigQuery runs on a serverless architecture, allowing it to auto-scale with the increase in data flow. The cluster running the script of the model uses the Apache Spark framework (with Spark) to allow for distributed computing and fast scaling of the model.

The solution is well suited for a significant increase in data both on the provided data structure as well as for new data that needs to be integrated (data on stock, logistic restrictions, etc). This architecture has proven its efficiency in production for customers with >100 mil observations on a daily basis.

The solution is flexible for adapting to other domains, especially given the nature of the CF Algorithm. Because it quantifies patterns in user behaviour, it is not restricted to the item characteristics of a specific vertical. We have successfully deployed the CF Algorithm for customers in other fields like Ecommerce electronics, fashion, home & deco, etc. with proven results.

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).

The solution provides different measures to ensure the security of the data, throughout the whole Data Value Chain, following the best practices in terms of data security policies. The REST APIs follow an OAuth authentication protocol. The data is encrypted in transit and at rest. The application provides multi-factor authentication options and all communication with the UI is encrypted using SSL/TLS. To avoid data loss events, the GCP-powered data warehouse (BigQuery) ensures correct redundancy and data backups to account for soft & hard failures.

Given the sensitive nature of data in the pharmaceutical industry, our solution relies on strictly anonymized, non-PII (Personally Identifiable Information) data and respects the GDPR regulation. As requested by COFARES, if the substitution of the client id in the provided dataset is reversible, the traceability of usage of the client data can be ensured. This can be achieved by implementing the ProRegister Tool provided by the REACH Toolbox.

Aqurate has assigned a Data Protection Officer (DPO) to ensure ongoing compliance with GDPR provisions and address inquiries.

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.

One technical risk regarding the implementation of the CF model, that can arise during the Experiment Phase, could be the so called Cold Start Problem. Depending on the distribution of interactions between users and items, it might happen that some items are lacking or they appear in a low number of transactions, such that the model could encounter difficulties in properly assigning them to the right pattern. Given the small sample data, it is hard to infer to what degree this problem might occur for the whole data set. As we have previosuly dealt with this issue before, we aim to develop a fallback model (content-based or rule-based) to inform recommendations for items where data is scarce. After the items exhibit enough interactions, they are caught by the CF model and mapped out in the interaction network.

General data issues are caught through designed alerting flags in the overall CI/CD monitoring system.

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

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