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?"

One of the best features of DPella solution is that it does not require seeing customer data to parametrize and deliver our product: the DPella privacy-preserving analytics engine. DPella needs to know only the database schema to produce the deliverable. Despite that, for the scope of this project where we will develop a new feature -privacy-preserving social network Analysis- for DPella solution, we need Joint's data (data provider). After successfully finishing this project we will not require customer data to parametrize and deliver our product on social network analyses.

In this project, we will (i) evaluate DPella's current analytics for HR-related metrics and (ii) develop a new feature: privacy algorithms for social network analytics, a type of analytics that today DPella does not currently support.

We need data provided by Jolint to validate our value proposition through experimentation. In this project, we will create synthetic data for testing purposes out of a description of Jolint's data, e.g., using tables schemas. Some time will be invested into calibrating the privacy/accuracy parameters of DPella's engine to the proposed analytics.

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.

We use differentially private algorithms; our data engine has different user roles to facilitate data governance. It separates the privileges of data owners from those of data analysts. At the moment, the engine connects to file-based databases. Our architecture design is based on containerization and runs on customer premises – a conscious design decision that protects customers from transferring sensitive data to DPella, which could potentially demand a considerable legal paperwork.

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.

This project scope is to elaborate several HR-related scores (e.g., inclusion scores, burnout scores, etc.) per team based on organization's data extracted from Microsoft Teams' metadata. The analytics that we will develop in this project can be generalized to other social network analyses, to understand the interaction among different parties. We envision applying this solution to segments like healthcare, social services, financial data, marketing, social media.

Differential privacy works best with large datasets, and as such, it is an intrinsically scalable technology.

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 core value of DPella is preserving privacy of individuals when publishing or sharing data. DPella applies Differential Privacy, one of the Privacy Enhancing Technologies (PETs). PETs are a set of tools and techniques designed to protect individuals' privacy while using digital services and handling personal data. These technologies aim to ensure that data is collected, stored, processed, and shared in a privacy-preserving manner, while still allowing for the necessary functionalities and services.

Differential Privacy is one of the PETs that can be utilized in various contexts to protect personal data while allowing for data analysis and insights. In this context, GDPR says that data insights are not considered personal data as long is not possible to identify one individual on the data set (Recital 26, Recital 162). This is what Differential Privacy offers, therefore when applying DP, the output is GDPR compliant.

Not only us believe that Differential Privacy provides stronger privacy protection than, for instance, anonymization, but also the EU Commission in the (recent) Data Governance Act where it recommends the use of Differential Privacy.

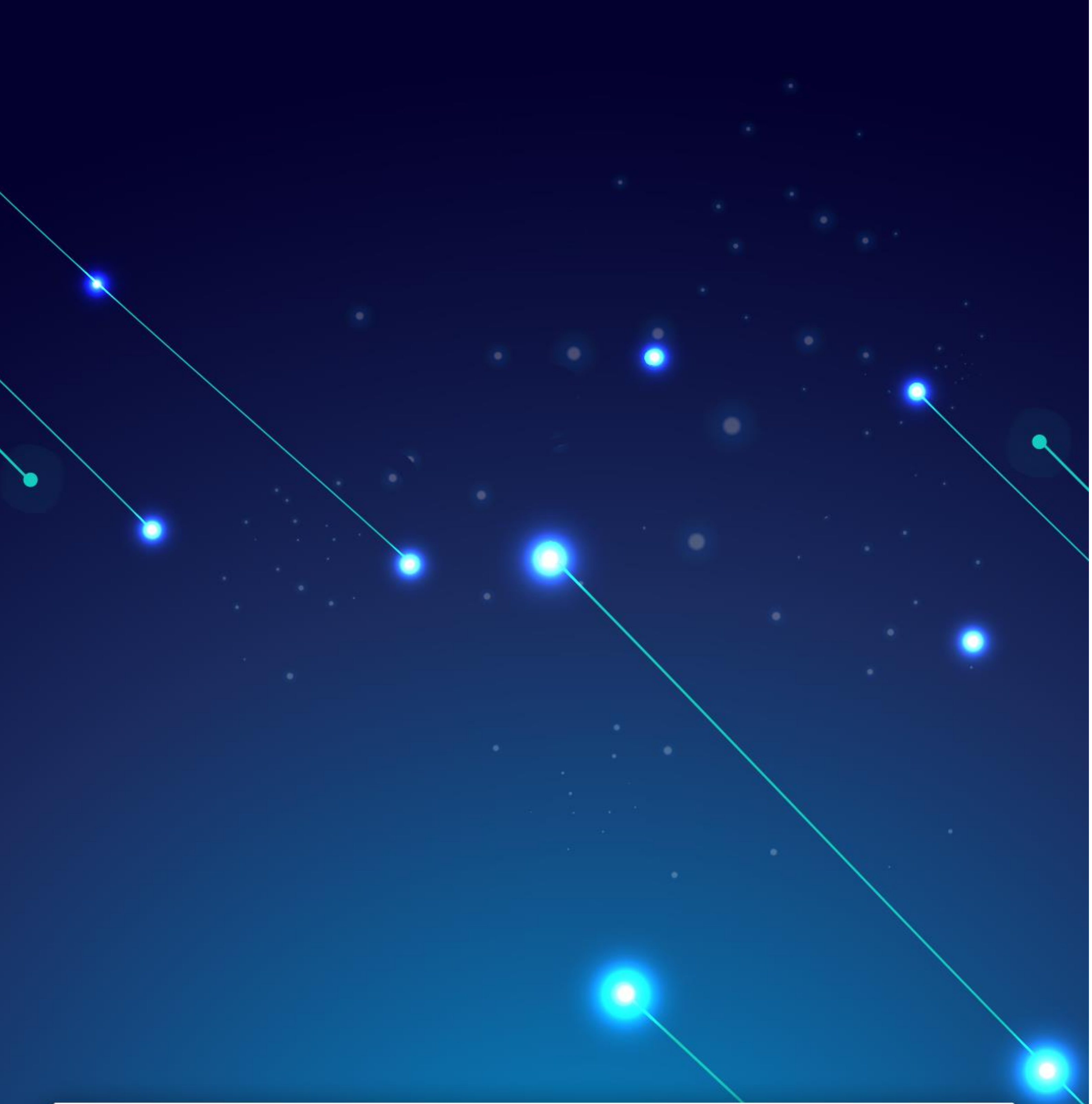
5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Feasible and credible quality process followed for the final product generation. The potential risks in all the phases of the project (design of the solution, development, testing, deployment…) are identified and convincing mitigation plans put in place.

- Design: The risk in the design of the algorithms is that of getting privacy protection at the expense of a big degradation of the quality of the results. To mitigate this, we might need to reduce the scope of our solutions to analyze medium to large-sized teams. We could also fine-tune existing algorithms to Jolint's specific needs to increase accuracy.
- Development: Currently, the engine is executing the analytics together with the privacy protection. One of the risks is that the engine does not scale for big datasets and computational expensive analytics. In this case, we could mitigate these risks by offloading the computation of the analytics to industrial database engines; thus leaving DPella's engine only responsible for making the results privacy-preserving.
- Testing: We do not foresee major risks on this stage as we use advanced programming language techniques to ensure correctness by construction, e.g., strongly typed programming languages.
- Deployment: There is a minimal risk of obtaining non-deterministic builds which might introduce unforeseen runtime failures in our customer's execution environment. Currently, we use Docker containers to address this problem.

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