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?"

We have built GeneFinder / GPPEM, a marketplace where firms and national digital sequencing initiatives can buy and sell access to digital sequence information and access "Digital Product Passport" information for such data with a focuse on enzyme manufacturing. Our marketplace is unique in that it allows firms to make interesting genetic sequences searchable without the need to fully open-source their IP - consumers can see the 'score 'that a hit has and based on this execute a Material Transfer Agreement (MTA) that defines next steps in a collaboration (e.g. licensing terms) without having yet seen the actual sequence. This allows supplier firms to discover which of their IP assets might be monetizable without having to undergo an expensive IP protection (e.g. patent) process.

One example use case for this platform could be Firm A, which sells an RNA polymerase enzyme (relevant for mRNA vaccine manufacturing), and has gathered a large dataset of viral proteins during their R&D process. This dataset will likely include other viral proteins of interest (for instance DNA polymerases) that Firm B may be interested in developing, but for which Firm A has no current use. By becoming a Supplier on the GeneFinder network, Firm A's novel DNA polymerases might be found when Firm B queries the network for similar hits to their current-best DNA polymerase.

To kick-start the marketplace we have built a state-of-the-art platform that facilitates discovery of public sequence data from ENA and NIH/SRA that we annotate with interesting metadata in the form of a standardized "Product Passport" that is enriched by the outputs of ML models; these calculate industrially-relevant properties such as thermostability and expression in a desired host. Having drawn consumers to the platform, we will then start to onboard suppliers who trust our data security procedures and are willing to upload sequence data to the platform directly. In the final phase we will work with firms holding large volumes of sequence data and who wish not to relinquish full control of the data, but are willing to opt into an intelligent federated search platform.

We utilise an architecture which consists of careful data normalisation in a PostgreSQL database as well as storage and search in industry-standard compressed formats and tools for sequence data, either in protein or nucleotide form depending on the provider. We provide a GUI that helps users make sense of the volume of data via intelligent visualisations such as showing distributions for thermostability or expression, allowing users to quickly navigate amongst a vast dataset to their are of interest.

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.

As the amount of publicly-searchable genetic sequence information is already >10 PB in size when considering unassembled data, GeneFinder clearly relies on Big Data technologies. Among those currently in use are: PyTorch (for ML model development and execution), PostgreSQL (for normalized data and management), Spot CPU and GPU instances at various cloud providers, OCI containers, as well as Pandas and Biopython for numerous ETL and data management tasks.

The overall orchestration of various components of our stack can be managed via Kubernetes to allow for cloud neutrality. Industry-standard OSS such as the Kubernetes Cluster Autoscaler allow us to scale up and down compute as new data sources are added or new queries are entered into the system.

For the search of a protein sequence database, Diamond (OSS Tool) is used. For searching in a DNA sequence database, tblastn (OSS Tool) is used.

We have programmed the responsive front-end for GeneFinder/GPPEM with ReactJS. We use ExpressJS for the back-end along with PostgreSQL for the core database which stores "product-passport" relevant data such as ML model outputs, outputs on query-hit similarity, literature references, etc. We use a REST API that we document via Swagger so that customers can choose whether they interact with the system via API or via the GUI. We expect the GUI to be the predominant means that users interact with the application

We have validated that our Open Data Sources from NIH and ENA could be integrated into our prototype application with product passport information calculated by ML models run on our backend.

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

CORE PARTNERS

DATA PROVIDERS

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.

We have validated the scalability of our tool on multi-petabyte datasets and this work has been featured in numerous German newspapers for a commercial polymerase discovery project. We achieve this scale via a combination of spot instances, OCI containers, as well as careful job queue management in a postgreSQL backend.

We have validated the flexibility of our data model via integrating multiple types of digital sequence information - currently over 10 different providers are supported with a number of different focuses (e.g. protein datasets, nucleotide datasets, metagenomic datasets, etc). We achieve this flexibility by supporting industry-standard dataset formats and supporting a standardised ETL pipeline that allows us to integrate new data sources with minimal effort.

Currently, we assume that most data providers have far less than 1 PiB a sequence data currently, and have data generation capabilities of lower than 100 TiB/Y.

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).

We are careful only to process data for allowed purposes and to not request unnecessary data to ensure GDPR compliance via following industry best practices in this domain.

For those suppliers who wish to keep their sequence information entirely in their own custody, we plan to support them to keep their data outside of our control. They can participate in the network by receiving queries and uploading result IDs and scores to our REST API; they can verify that only the IDs and scores of the top N hits are returned, ensuring that (from an information-theoretic point of view) very little information was transmitted back. Proteineer can verify who is running queries in a KYC process and can limit the volume of queries processed to further reduce the amount of information returned to a querier and further protect supplier 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.

We follow a two-week agile iterative development process and practice continuous integration, user feedback sessions and unit testing to ensure high-quality software deliverables. We are addressing risks in product design via talking with domain experts in the field, including leading industry biologists and professors in Germany and abroad. We address development risks via careful documentation and code review as well as leveraging open-source tools where relevant. We address testing risks via rapid testing of new features and plan to build out further environments (staging, UAT, etc) as our development team grows.

The major risks we see for this solution are that users may be wary about uploading digital sequence information that they have laboriously obtained and that can hold great industrial value. We will mitigate these risks by either acting as a third party (for those willing to trust our software's redaction/MTA workflows) or via providing template solutions such that they can easily join a federated search network.

Furthermore we intend as we grow to conduct a security audit for our software via a trustworthy third-party firm as well as obtain industry-standard certifications similar to ISO 9001 for our software development processes.

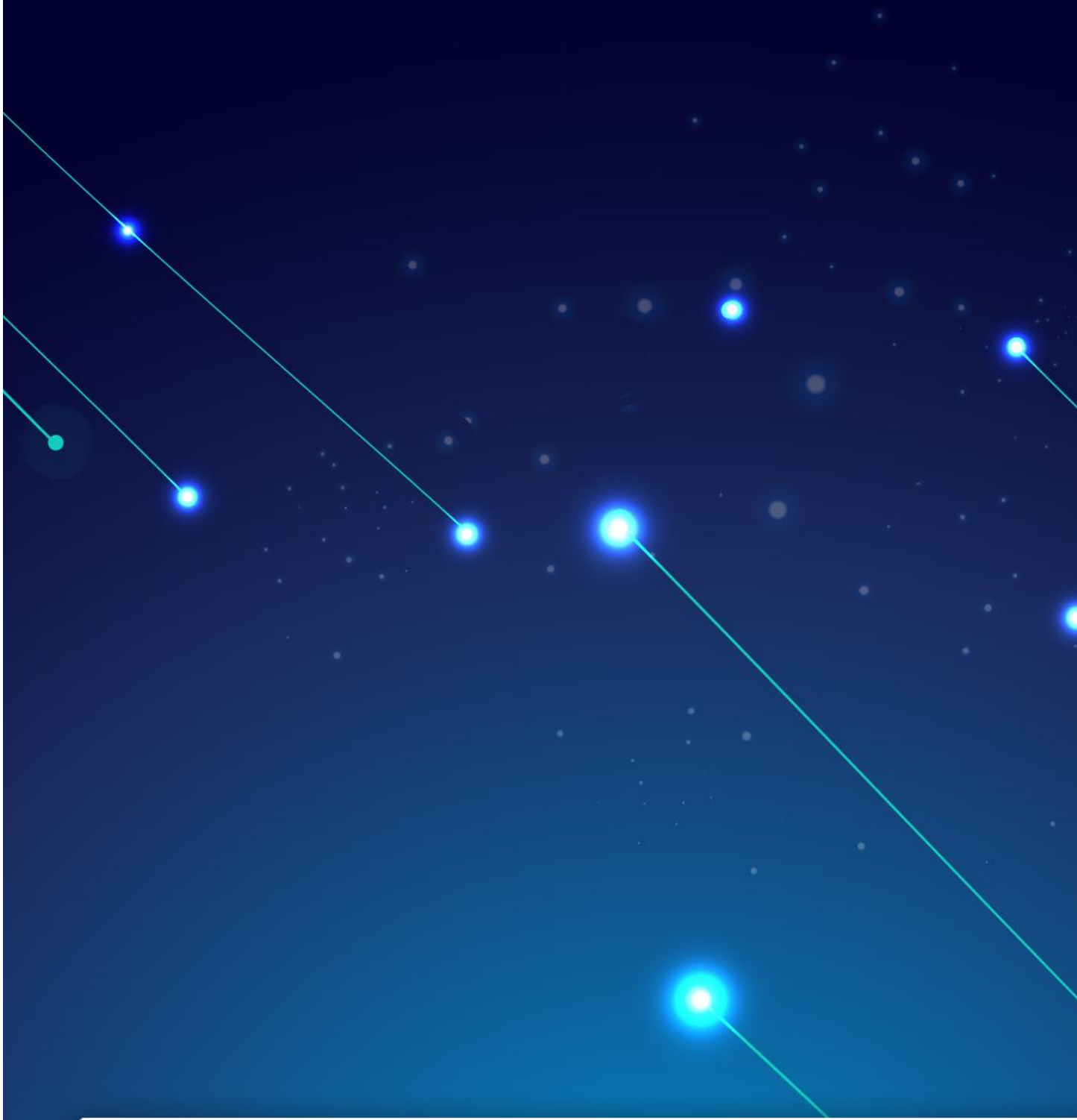
This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

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