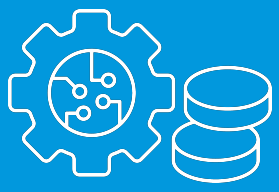

» CUSTOMER CASE STUDY

Cloud-Native Extract, Transform, Load (ETL) Pipeline

Working with a cross-functional team Nebulaworks built, operationalized, and maintained a bioinformatics ETL pipeline

NEBULAWORKS

Nebulaworks engineered a FAIR Compliant Cloud-Native ETL solution with a cross-functional team

Challenge

The challenge faced by the client was the need for a robust and scalable computing solution that could handle vast amounts of bioinformatics data while ensuring compliance with FAIR (Findable, Accessible, Interoperable, Reusable) principles. This compliance was crucial for the data to be traceable and reusable, enhancing confidence in the data across various stages of research and development and into product launch. The client required a system that could bridge the gap between bioinformatics and engineering domains, requiring a solution that was not only technically sound but also scalable to accommodate data from both private and third-party sources.

Solution

Nebulaworks engineered a cloud-native ETL (Extract, Transform, Load) solution using a cross-functional team. The client required the ability to import various types of data from third-parties, public, and private data. This requirement influenced the architecture leveraged for this solution. The custom system was built exclusively on AWS, utilizing a technology stack that included the MERN Stack, DocumentDB, GraphQL, R, Python, and other industry-standard tools like Terraform and GitHub. The front-end compontent was served with CloudFront. The system adhered to a code-first approach, ensuring everything was deployable, reproducible, and environment agnostic. The architecture included load-balanced container resources to ensure scalability and reliability, supporting heavy usage across development, testing, and production environments. Since this was a FAIR compliant system, the cross-functional teams enabled the ability to attach metadata to all data during the Transform step in order to have a consistent data form when executing queries across various sources. This capability enabled data rollups and the ability to aggregate data from diverse collections into a single corhort collection. In addition, the system was able to create unified data where needed and provided pre-formed queries. Once the data was acquired, cleaned, tagged with metadata and imported, it was ready to be visualized. Front end mechanisms

NEBULAWORKS

were created to be highly interactive tables, and avoid the "untracked spreadsheet" that breaks FAIR principles, ultimately resulting in an improved UX and knowledge preservation for users.

Why Nebulaworks

Nebulaworks was chosen for their extensive expertise in software and cloud engineering, full stack ownership, and deep understanding of the software development lifecycle (SDLC). They brought operational reliability, robust architecture, and KPI-driven iteration processes to the project. Their approach to work directly with line-of-business stakeholders and their ability to embed their team within the client's operations allowed for a deeply integrated solution tailored to the client's specific needs.

Outcomes

The implementation of the scalable ETL solution by Nebulaworks resulted in several significant outcomes:

FAIR Compliance: The platform fully adhered to FAIR principles, providing a system where data is findable, accessible, interoperable, and reusable. This compliance ensures that data and analysis results are traceable and can be confidently used and reused over time.

Scalability and Flexibility: The solution was built to scale, capable of ingesting and processing data not only from private sources but also from third-party data providers. This scalability ensured that the platform could handle increasing data loads without performance degradation.

Enhancec Data Management: With the FAIR-compliant platform, updates to data, ontologies, and algorithms are managed more safely and effectively, preserving institutional memory around data and data analysis.

Operational Efficiency: The system's design to be environment agnostic and its use of industrial-grade platforms for all operational environments minimized downtime and maximized performance.

In conclusion, the cloud-native ETL solution designed by Nebulaworks not only met the complex requirements of the client but also set a new standard for how bioinformatics data can be managed and utilized in a FAIR-compliant manner.