Proposal Title: Data Engineering a RAG system for Refugee Support.

Details: Working specifically on the Data Engineering of the data that will be fed into the RAG System. The RAG system will be specifically tailored to help Refugees with integration into the Netherlands/ understanding complex processes. This project would work with a Dutch Charity to provide the data and problem statement (*confirmation pending – otherwise synthetic data will be used*). This would enable the Spike Up project candidates to gain Data Engineering experience, as well as consulting experience by interviewing the Charity to understand the problem.

Overview of the main Data Engineering tasks for the project:

- Data Ingestion
- Data Cleaning, Preprocessing
- Chunking and Embedding Prep
- Vectorization and Indexing
- Data Versioning
- Monitoring, Logging
- Data Governance
- Integration with RAG System

Essential skills and tools (to be familiar with before starting):

- Python
- Git
- VS Code
    - *or another editor of your choice*
- PowerPoint
    - Presenting your findings and results in front of a group.

Complete tech-stack (not essential to know everything already beforehand):

- AWS services part of the stack:
    - AWS Lambda
    - AWS Bedrock
    - AWS S3
    - AWS ECS
    - AWS CloudFormation
    - AWS CDK
- Github + Github Actions
- Specific python tools

  - Pre-commit
  - ruff
  - pyenv
  - poetry
  - pytest
  - mypy
- Specific Python packages
  - AWS CDK
  - Streamlit
  - langchain
- Docker

Timeline:

Time	Topic	Details	Who?
Week 1	Onboarding to FF, the topic and the previous RAG project.	Predefined backlog. Daily standups. 1-week sprints.	2hr per week for FF employee
Week 2	Work on Project.		2hr per week for FF employee
Week 3	Work on Project.		2hr per week for FF employee
Week 4	Finished product.	Presentation in FF Friday Session as Preparation for Graduation ceremony	2hr per week for FF employee