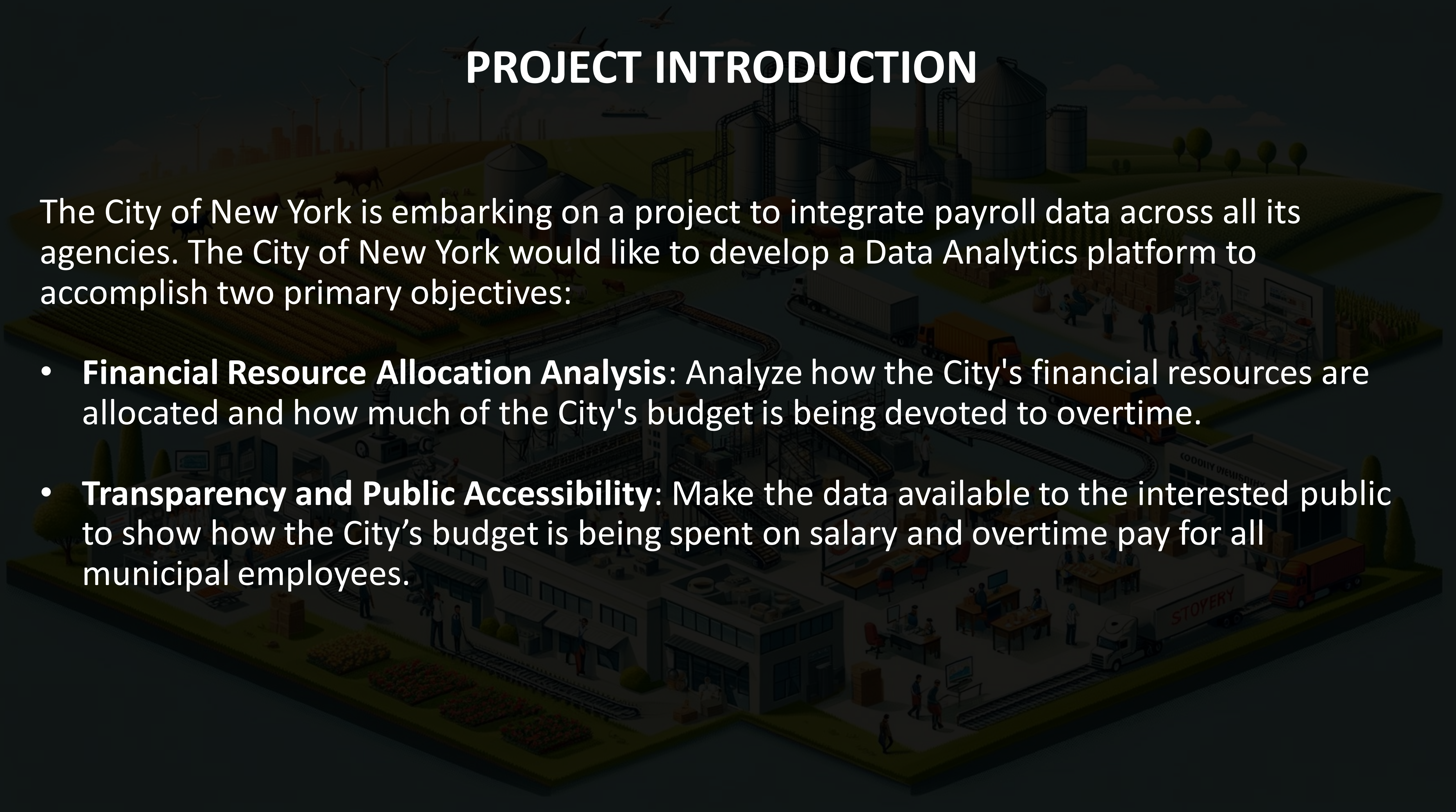
10Alytics

CAPSTONE PROJECT: NYC PAYROLL DATA INTEGRATION

CHU NGWOKE

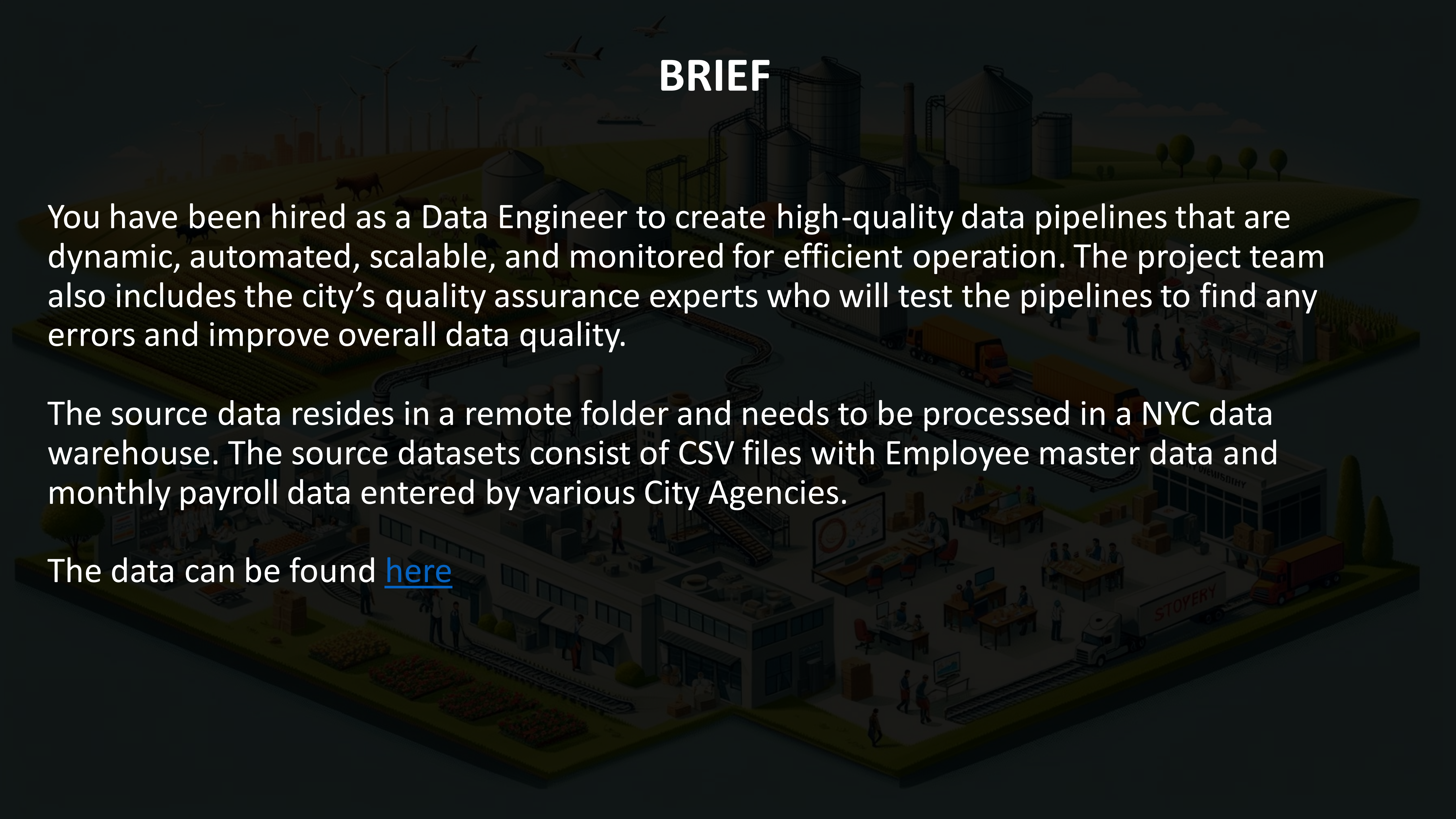
PROJECT INTRODUCTION



The City of New York is embarking on a project to integrate payroll data across all its agencies. The City of New York would like to develop a Data Analytics platform to accomplish two primary objectives:

- **Financial Resource Allocation Analysis**: Analyze how the City's financial resources are allocated and how much of the City's budget is being devoted to overtime.

- **Transparency and Public Accessibility**: Make the data available to the interested public to show how the City’s budget is being spent on salary and overtime pay for all municipal employees.



BRIEF

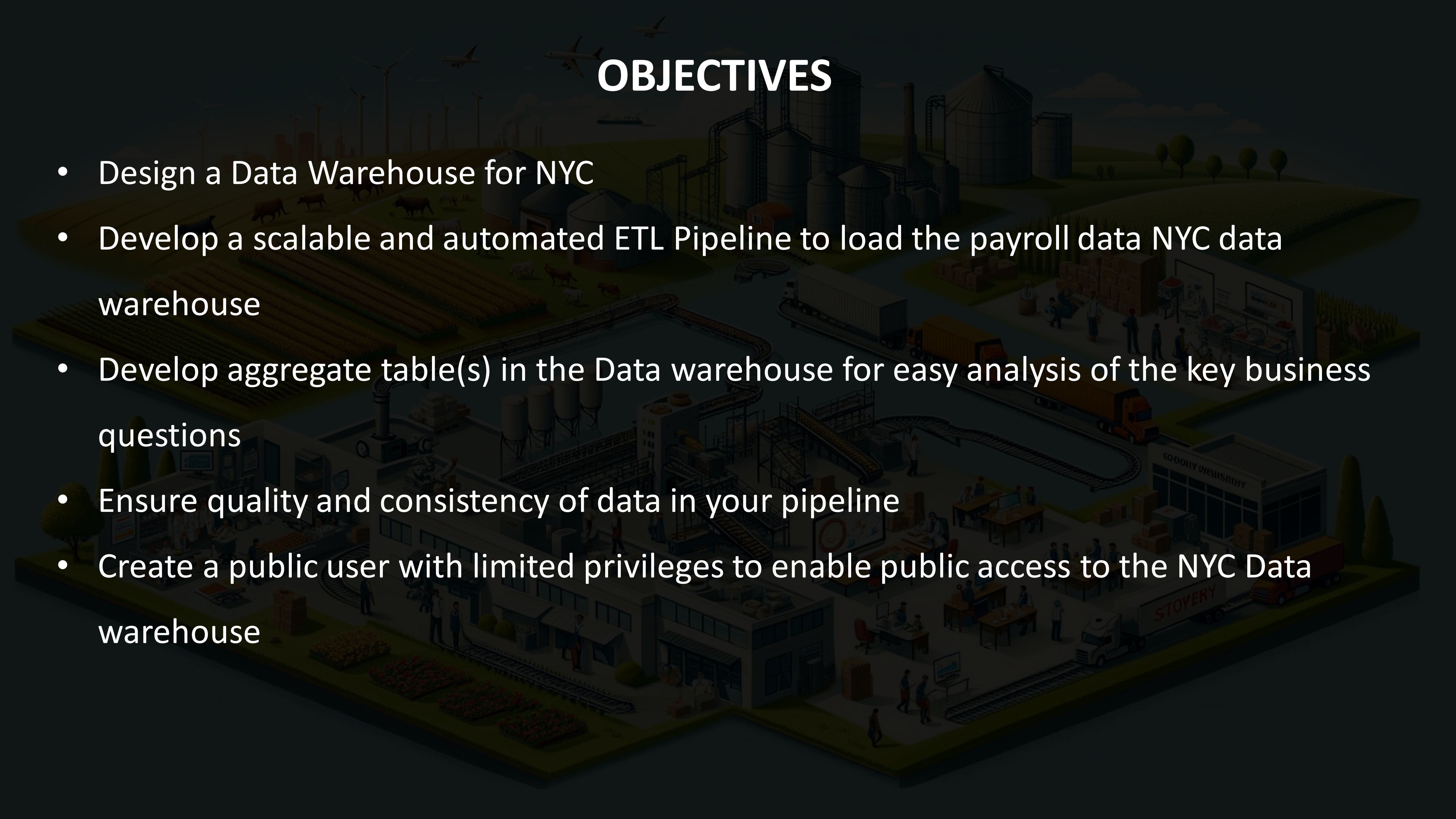
You have been hired as a Data Engineer to create high-quality data pipelines that are dynamic, automated, scalable, and monitored for efficient operation. The project team also includes the city’s quality assurance experts who will test the pipelines to find any errors and improve overall data quality.

The source data resides in a remote folder and needs to be processed in a NYC data warehouse. The source datasets consist of CSV files with Employee master data and monthly payroll data entered by various City Agencies.

The data can be found here

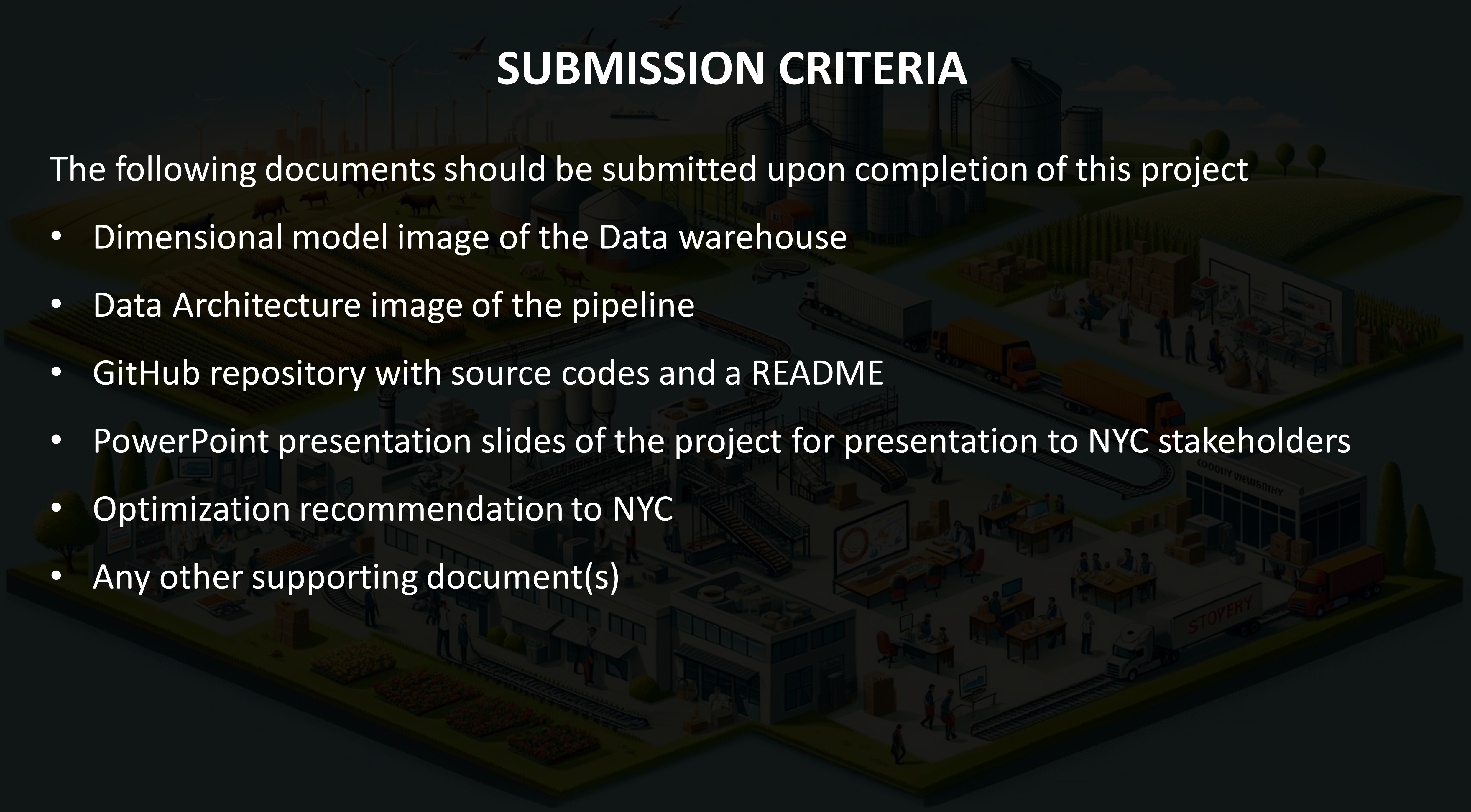
OBJECTIVES

- Design a Data Warehouse for NYC
- Develop a scalable and automated ETL Pipeline to load the payroll data NYC data warehouse
- Develop aggregate table(s) in the Data warehouse for easy analysis of the key business questions
- Ensure quality and consistency of data in your pipeline
- Create a public user with limited privileges to enable public access to the NYC Data warehouse



OBJECTIVES

- Properly document all your processes for reproducibility – version control
- Maintain a cloud hosted repository of your codebase to facilitate collaboration with team members

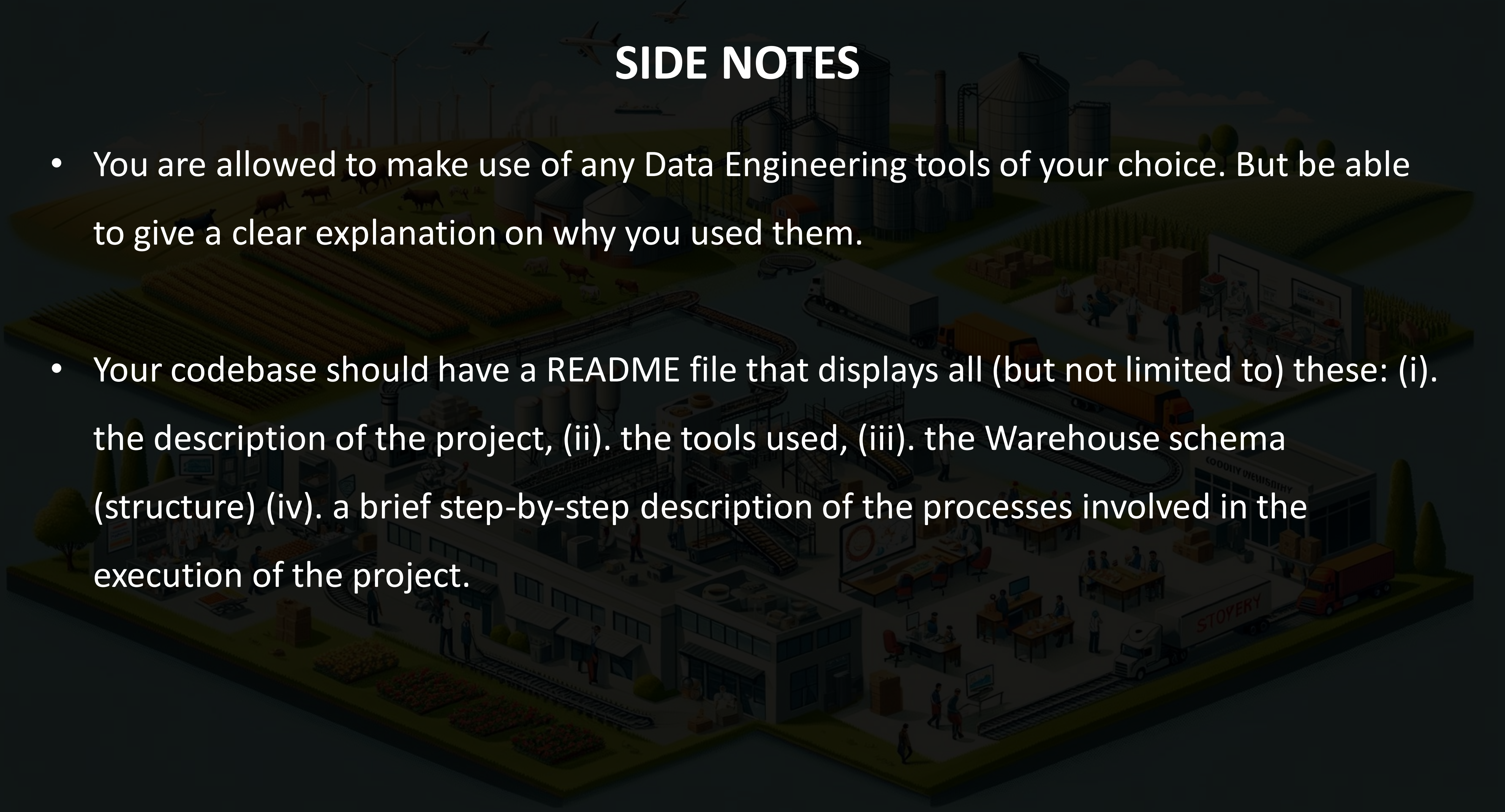


SUBMISSION CRITERIA

The following documents should be submitted upon completion of this project

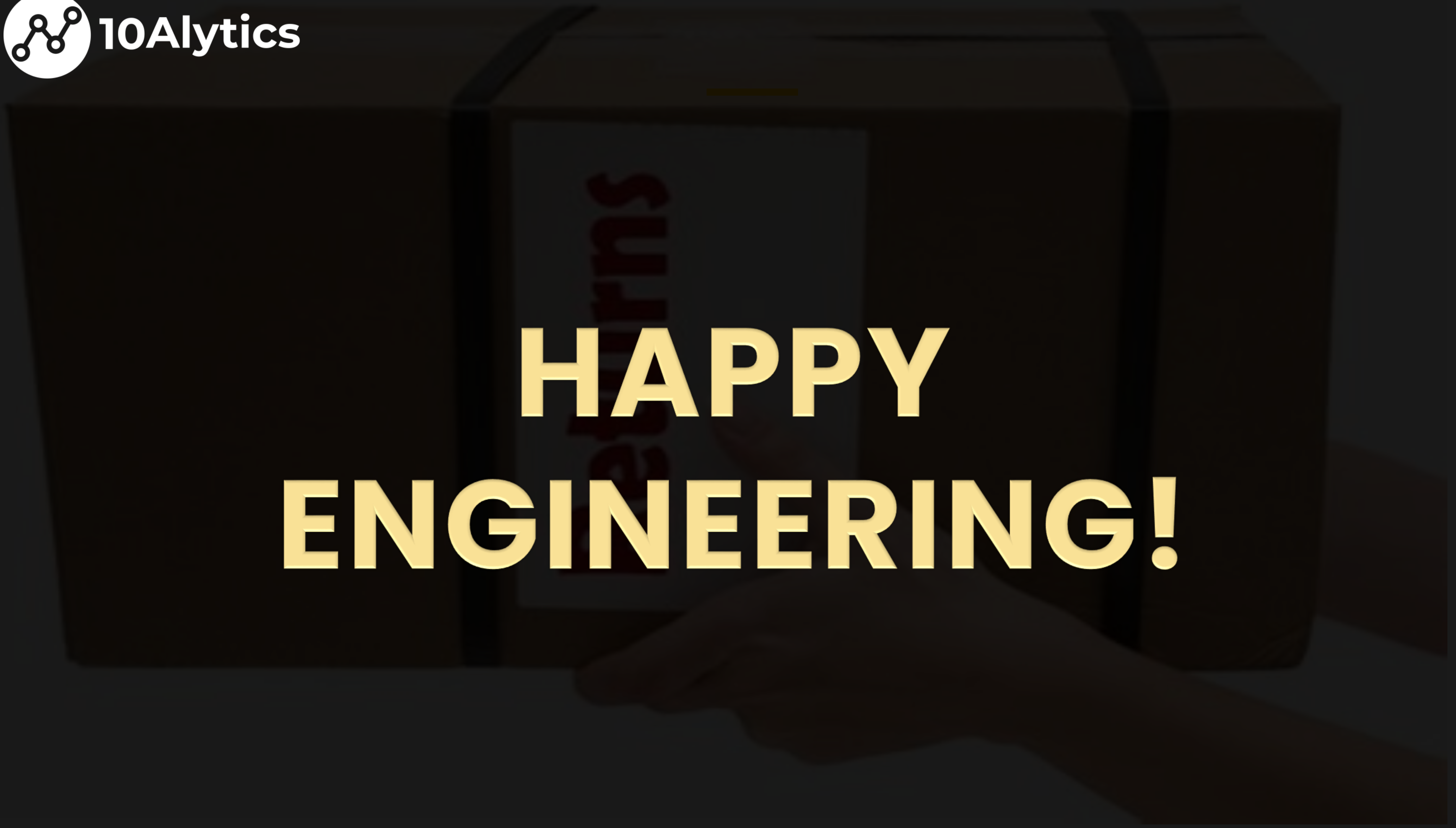
- Dimensional model image of the Data warehouse
- Data Architecture image of the pipeline
- GitHub repository with source codes and a README
- PowerPoint presentation slides of the project for presentation to NYC stakeholders
- Optimization recommendation to NYC
- Any other supporting document(s)

SIDE NOTES



- You are allowed to make use of any Data Engineering tools of your choice. But be able to give a clear explanation on why you used them.

- Your codebase should have a README file that displays all (but not limited to) these: (i). the description of the project, (ii). the tools used, (iii). the Warehouse schema (structure) (iv). a brief step-by-step description of the processes involved in the execution of the project.



HAPPY ENGINEERING!