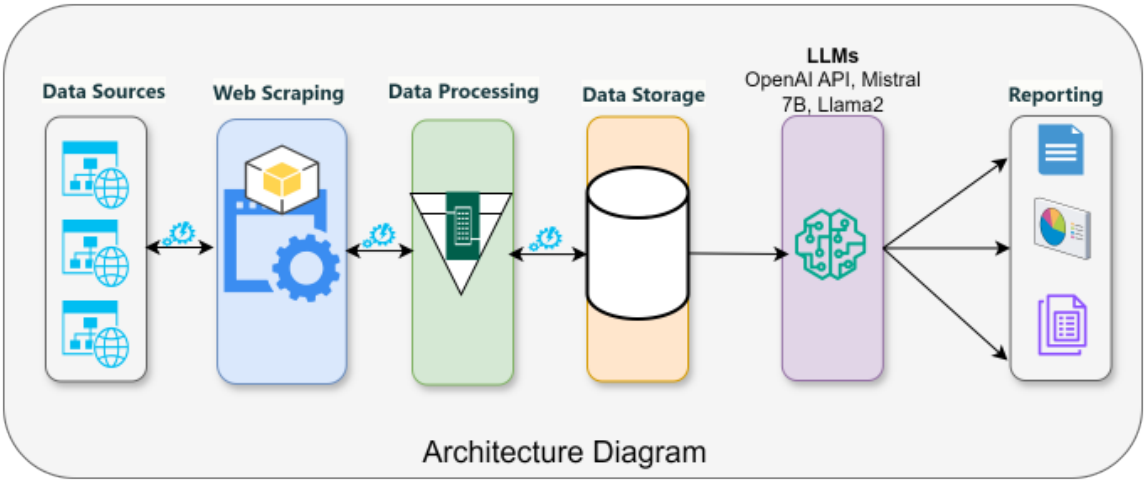
Data Engineering - AI/ML Task

The task is about a trial assignment on **AI** for a data engineering role, focused on Automating the process using AI/ML design & Web scraping.

The task involves creating a scalable and automated data pipeline to continuously extract, process, and standardize data from various online sources related to construction and infrastructure projects in California. The key challenges include identifying reliable data sources, developing effective web scraping methods, standardizing the diverse data collected, and automating the entire process to ensure it runs regularly without manual intervention. The solution must be robust enough to handle large volumes of data, ensure continuity, and be ready for deployment in a production environment.

The main objectives are

1. **Research and Data Sourcing**:
    - Identify 5-10 reliable online sources of data related to construction and infrastructure projects and tenders in California. This involves conducting online research, possibly with the assistance of language models like OpenAI's GPT.
2. **Data Extraction and Standardization**:
    - Develop methods to scrape the identified data sources using web scraping tools and language models.
    - Standardize the scraped data according to specific guidelines, ensuring consistency and usability.
3. **Automation and Continuous Updating**:
    - Design a system to automate the data scraping and standardization processes.
    - Implement scheduling tools like cron jobs to ensure the system continuously updates the data.
4. **Architecture diagram**:



Key Deliverables:

- Python scripts for web scraping and data processing.
- Documentation explaining the approach and the logic used.
- Sample datasets to demonstrate the results.
- A plan for deploying the solution in a production environment.

Evaluation Criteria:

- The scalability of the solution.
- The quality of the data standardization.
- The effectiveness of the automation process.

Additional Notes:

- The task emphasizes creating a scalable and production-ready solution.
- Candidates should clearly articulate how AI models are used in the process.

The overall goal is to build a robust and automated data pipeline that can continuously extract, process, and update data from multiple online sources related to infrastructure projects

After completing this project, students will gain practical experience in several key areas:

1. **Web Scraping Techniques**: Students will learn how to extract data from websites using tools like Python, including handling challenges such as navigating dynamic content and dealing with anti-scraping measures.
2. **Data Standardization**: They will understand the importance of data consistency and how to clean and standardize data from diverse sources to make it usable for analysis or other applications.
3. **Automation and Scheduling**: Students will gain experience in automating repetitive tasks using scheduling tools like cron jobs, ensuring that data pipelines run efficiently and consistently over time.
4. **Scalable Data Engineering Solutions**: The project will teach students how to design and implement scalable data pipelines that can handle large volumes of data, with a focus on reliability and maintainability.
5. **Integration of AI Tools**: By potentially incorporating AI language models into the data extraction and processing tasks, students will learn how to leverage AI to enhance traditional data engineering workflows.
6. **Deployment in a Production Environment**: They will also gain insights into deploying data engineering solutions in real-world production environments, considering factors like data storage, access, and monitoring.

Overall, this project will equip engineers with a comprehensive understanding of building, automating, and deploying data pipelines in a professional setting, preparing them for real-world data engineering challenges.