**Wikipedia Data Scraping: Efficient Extraction and Database Integration**

**Project Overview**

The Wikipedia data harvest project revolves around the extraction of data from Wikipedia, specifically focusing on universities, and its systematic integration into various databases. The primary objective is to automate the retrieval and storage of university-related information for diverse database platforms, initially encompassing MySQL, Postgres, and SQL Server.

**Project Learning Opportunities**

* Implement web scraping using Python libraries such as Beautiful Soup
* Set up and configure different databases using Docker
* Understand HTML structure to navigate and extract relevant information from Wikipedia pages
* Implement code to efficiently load scraped data into various databases
* Explore techniques for managing database connections and transactions

**Learning Skills**

* Web Scraping skills
* Python Programming
* Database Management Skills
* SQL
* Diverse Database Knowledge

**Executive Summary**

The Wikipedia data harvest project revolves around the extraction of data from Wikipedia, specifically focusing on universities, and its systematic integration into various databases. The primary objective is to automate the retrieval and storage of university-related information for diverse database platforms, initially encompassing MySQL, Postgres, and SQL Server.

**Problem Statement**

In the ever-evolving landscape of educational data management, the absence of an automated and versatile solution for acquiring comprehensive university information presents a significant challenge. Manually collecting and updating data from Wikipedia is time-consuming, error-prone, and lacks scalability. The need for a streamlined process to systematically harvest university data from Wikipedia and integrate it into various databases has become increasingly apparent.

The Wikipedia data harvest project revolves around the extraction of data from Wikipedia, specifically focusing on universities, and its systematic integration into various databases. The primary objective is to automate the retrieval and storage of university-related information for diverse database platforms, initially encompassing MySQL, Postgres, and SQL Server.

**Objectives**

**Implement Web Scraping:** Develop a robust web scraping module using Python libraries to extract relevant university data from Wikipedia pages

**Database Setup and Configuration:** Set up and configure databases with Docker to establish a foundation for data storage

**Database Schema Design:** Design and implement appropriate database schemas to accommodate the scraped university data

**Database Integration:** Develop scripts to load scraped data seamlessly into the designated databases

**SQL Query Development:** Craft SQL queries for creating, reading, updating, and deleting data in the databases

**Cost Efficiency:** Optimize resource utilization and costs by leveraging Distributed SQL Queries.

**Data Pipeline**

A diagram of a computer code

Description automatically generated

**Tech Stack**

1. A. SQL Server
2. B. PostgreSQL Database
3. C. MySQL

**Development**

A. Python

B. Docker

**Project Scope**

* Understanding the project Overview and Architecture.
* The systematic extraction of relevant university information from Wikipedia page.
* Utilizing Docker containers for different database systems
* Loading data into different databases from python environment