



ETL Pipeline Project - Python

09.12.2023

Chinna Reddy Karri

+1 6025545409

chinnareddykarri8@gmail.com

<https://www.linkedin.com/in/chinnareddykarri/>

https://github.com/chinnareddykarri8/ETL_Pipeline_Project

Overview:

Hello, I'm excited to share a project I've recently completed, which I believe showcases my proficiency in data engineering and Python programming. This project, titled "Data ETL Pipeline using Python," allowed me to delve into the world of Extract, Transform, and Load (ETL) processes and solidify my skills in this critical area of data handling.

Goal:

In data engineering, ETL pipelines are fundamental tools for extracting, transforming, and loading data from diverse sources into databases for analysis. I embarked on this project with the goal of not only mastering these concepts but also applying them to a practical scenario.

Data Collection:

The project commenced with the collection of data from a source – in this case, I chose the Fashion-MNIST dataset, a well-known benchmark in the machine learning community, conveniently accessible through the Keras library. This dataset represents a diverse range of clothing items, making it an excellent choice for our purposes.

Data Cleaning and Transformation:

Next came the fascinating part: data cleaning and transformation. I implemented Python scripts to normalise pixel values, ensuring they fell within the 0 to 1 range. This preprocessing step was crucial to prepare the data for further analysis and modelling. Additionally, I reshaped the dataset into a 4D tensor, ready for feature extraction and model training.

Data Loading:

One of the most rewarding aspects of this project was the data-loading phase. I utilised SQLite, a lightweight and efficient database management system, to create a database named "images." Within this database, I carefully structured the tables to store the image data along with their corresponding labels.

Data Insertion:

The ETL pipeline concluded with the insertion of both training and test data into the database. To ensure data integrity and consistency, I used Python's SQLite library to execute SQL commands, storing each image as a binary blob alongside its label. This step was vital for enabling quick access to data for future analysis and model training.

Conclusion:

In summary, this project represents a significant milestone in my data engineering journey. It not only equipped me with a deeper understanding of ETL processes but also demonstrated my ability to apply these skills in a practical context. The resulting ETL pipeline successfully extracted, transformed, and loaded the Fashion MNIST dataset into an SQLite database, setting the stage for seamless data access and analysis.

Contact:

If you have any questions or would like to explore this project further, please feel free to reach out through mobile - **+1 6025545409** and E-mail - **chinnareddykarri8@gmail.com**. I'm excited to discuss how this experience has contributed to my proficiency in data engineering and how it aligns with the goals of your organization.

Thank you for considering my portfolio, and I look forward to the opportunity to discuss this project in more detail.