Title: Building data pipeline to collect data FROM THE COURSE SALE WEBSITE

Data source: https://unica.vn/

Industrial university of Ho Chi Minh city

for Data Engineer

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student ID | Fullname | Task | Contribution rate (%) | Score (lecturer) |
| 22695761 | Nguyễn Thành Đạt | * Crawl data (technology, construction design) * Control and debug source code * Process and clean data | 40% |  |
| 22665311 | Thái Văn Sơn | * Write a data crawling program (unica\_crawler) * Crawl data (marketing) | 13% |  |
| 22691511 | Đỗ Viết Thanh Bằng | * Crawl data (foreign languages) * Merge crawled data (remove duplicates) * Push full data to MongoDB | 11% |  |
| 22693501 | Trần Tuấn Sang | * Crawl data (Finance and Accounting) * Design entity model | 13% |  |
| 22696961 | Mai Hoàng Tuấn | * Crawl data (Health and Beauty) * Create tables and insert cleaned data into PostgreSQL | 12% |  |
| 22690271 | Đỗ Tiến Đạt | * Crawl data (Office Applications) * Deploy the system on a virtual machine running Ubuntu Server based on Docker Test | 11% |  |

Trello URL: https://trello.com/invite/b/6707e985a3e2c825354f3699/ATTI6405a2570e6ed98c25ad2b7d87a123e9BB4A1DAA/unicaproject Established Date: 09/09/2024

Github URL: https://github.com/3002tad/Unica\_Project Established Date: 09/09/2024

Data Pipeline System Architecture Design

