Intro To Data Engineer Project

Intro To Data Engineer, Sekolah Engineering, Pacmann

**Outline**

[**Background Project 2**](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.6iqfjw70sxra)

[**Petunjuk Pengerjaan 4**](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.25dhvn3ake1m)

[Langkah #1 - Requirements Gathering & Proposed Solution 4](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.s8clu9kr177p)

[Langkah #2 - Designing ETL Pipeline 5](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.wa5ckzqcg7np)

[Langkah #3 - ETL Implementation & ETL Scheduling 6](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.vvs6il6b1n7w)

[ETL Implementation 6](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.9mpfhfh66hox)

[ETL Scheduling 7](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.3s9s8w5sf6gr)

[Langkah #4 - Testing Scenario 7](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.q8mxuk50pxbj)

[**Outcome Project 9**](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.nxtbsf7zp29l)

[Dokumentasi Project (Github) 9](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.b1f974eqqtec)

[Presentasi 9](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.2ty9vwx3lfoc)

[**Evaluasi 10**](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.i06luapc8118)

[**Tools 10**](https://docs.google.com/document/d/1uFJYmjokLf8Fl7dyRAXP_w4VrJxCxFourZZ7XUuc_t0/preview#heading=h.n17fxc2mp7ph)

Background Project



[*source*](https://www.freepik.com/free-vector/illustration-social-media-concept_2807766.htm)

Sebagai Data Engineer yang baik anda harus bisa membuat **ETL (Extract, Transform, Load)** Data Pipeline, karena itu merupakan salah satu task utama jika menjadi Data Engineer. Pada project kali ini, anda akan membuat ETL Pipeline tetapi tidak dari satu source saja, akan ada beberapa source data. Anda tidak hanya membuat ETL Pipeline saja, tetapi harus membuat prosesnya secara **end-to-end.**

Anda telah menjadi Data Engineer di Perusahaan XYZ, perusahaan masih terhitung baru dan masih belum memiliki infrastruktur data yang *robust.* Oleh karena itu, anda akan membuatkan **ETL Pipeline** dan **anda akan membuatkan satu database** yang digunakan untuk menyimpan seluruh data yang sudah melewati proses ETL. Ada beberapa data source yang dimiliki oleh perusahaan XYZ, oleh karena itu anda melakukan diskusi ke masing - masing Stakeholder untuk melakukan proses Requirements Gathering.

Ternyata, setelah berdiskusi dengan masing - masing Stakeholder anda menemukan temuan:

* Tim Sales sudah memiliki data penjualan barang di Database **PostgreSQL.** Tetapi masih banyak data yang **missing dan format data tidak benar**
* Tim Product juga sudah memiliki data tentang Electronic Product Pricing dalam bentuk **.csv** tetapi bentuk data masih berantakan dan banyak missing values
* Tim Data Scientist ingin melakukan research mengenai pemodel NLP (Natural Language Processing), **tetapi masih belum memiliki data sama sekali**. Akhirnya meminta bantuan ke Tim Data Engineer untuk melakukan proses Web Scraping dari sebuah website.

Ada beberapa hal yang harus dikerjakan dalam Final Project ini.

1. **Requirements Gathering & Proposed Solution**

Anda akan diminta untuk memahami problem apa yang diberikan oleh Stakeholder, setelah itu anda akan diminta untuk memberikan solusi berdasarkan problem tersebut.

1. **Designing ETL Pipeline**

Berdasarkan hasil Requirements Gathering & Proposed Solution, anda akan diminta untuk membuat ETL Design Pipeline secara overview tidak perlu detail.

1. **ETL Implementation**

Anda akan membuat ETL Pipeline dengan menggunakan Luigi berdasarkan design ETL yang sudah dibuat pada proses sebelumnya.

1. **ETL Scheduling**

Agar ETL Pipeline bisa berjalan sesuai dengan jadwal, anda diminta untuk membuat Scheduling untuk menjalankan Pipeline yang sudah dibuat menggunakan Crontab.

1. **Testing Scenario**

Agar Pipeline yang dibuat lebih *robust,* maka akan diberikan sebuah testing scenario untuk memastikan apakah Pipeline bisa berjalan sesuai dengan skenario yang diberikan.

1. **Documentation**

Data Engineer yang baik adalah yang bisa membuat summary atau dokumentasi mengenai Pipeline yang telah dibuat, mulai dari apa solusi yang diberikan, arsitektur atau bagaimana design pipeline yang dibuat, bagaimana cara menjalankan pipeline, dsb.

Project ini memberikan kesempatan kepada Anda untuk mengaplikasikan pengetahuan dan keterampilan yang telah Anda pelajari pada course Intro to Data Engineer. Anda juga bisa memperdalam pemahaman tentang membuat ETL Pipeline, melakukan ETL Scheduling, dan membuat Design ETL Pipeline.

**Catatan:** Project dikerjakan secara **Individu**.

Petunjuk Pengerjaan

Untuk mempermudah pengerjaan, berikut adalah hal-hal yang dapat Anda lakukan.

Langkah #1 - Requirements Gathering & Proposed Solution

*Expected Time to Start: Week 4*

*Estimated Time of Completion: End of Week 5*

Pada tahapan ini, **anda harus memahami** mengenai data apa yang akan dipakai, source-nya darimana, bentuk dan konteks data seperti apa, dan solusi apa yang harus dipakai untuk menyelesaikan problem yang dihadapi dari Stakeholder. Solusi ini bisa seperti metode apa yang bisa kita pakai pada proses Transform, tools apa yang digunakan, dsb

Data source yang dipakai:

* **Sales data**

Untuk Sales data, anda dapat mengaksesnya pada Docker berikut <https://hub.docker.com/r/shandytp/amazon-sales-data-docker-db>

* **Marketing data**

Untuk Marketing data, anda dapat mengakses data pada link berikut [ElectronicsProductsPricingData.csv](https://drive.google.com/file/d/1J0Mv0TVPWv2L-So0g59GUiQJBhExPYl6/view?usp=sharing)

* **Web Scraping**

Anda diberi kebebasan untuk website apa yang ingin di scraping, boleh portal berita atau sejenisnya dan anda diberi kebebasan untuk melakukan scraping menggunakan struktur HTML atau menggunakan API. **Tetapi, pastikan website yang digunakan diperbolehkan untuk dilakukan proses scraping dan berikan disclaimer pada dokumentasi!**

Jangan lupa, seluruh data akan kita masukkan ke dalam satu database yang sama pada proses Load!

**Contoh Output**:

Problem:

* Tim A ingin melakukan proses analisa mengenai performa dari penjualan dari tahun 20xx sampai 20xx
* Tetapi data yang ingin dianalisa tersebar dimana - mana, sehingga itu mempersulit tim A untuk melakukan analisa
* …

Data Engineer Solutions:

* Data Engineer akan membuatkan Data Pipeline dan Infrastructure untuk mengambil data yang tersebar dan akan dijadikan ke dalam satu database
* Data Engineer juga akan membuat scheduling agar data yang di database terupdate terus
* …

Langkah #2 - Designing ETL Pipeline

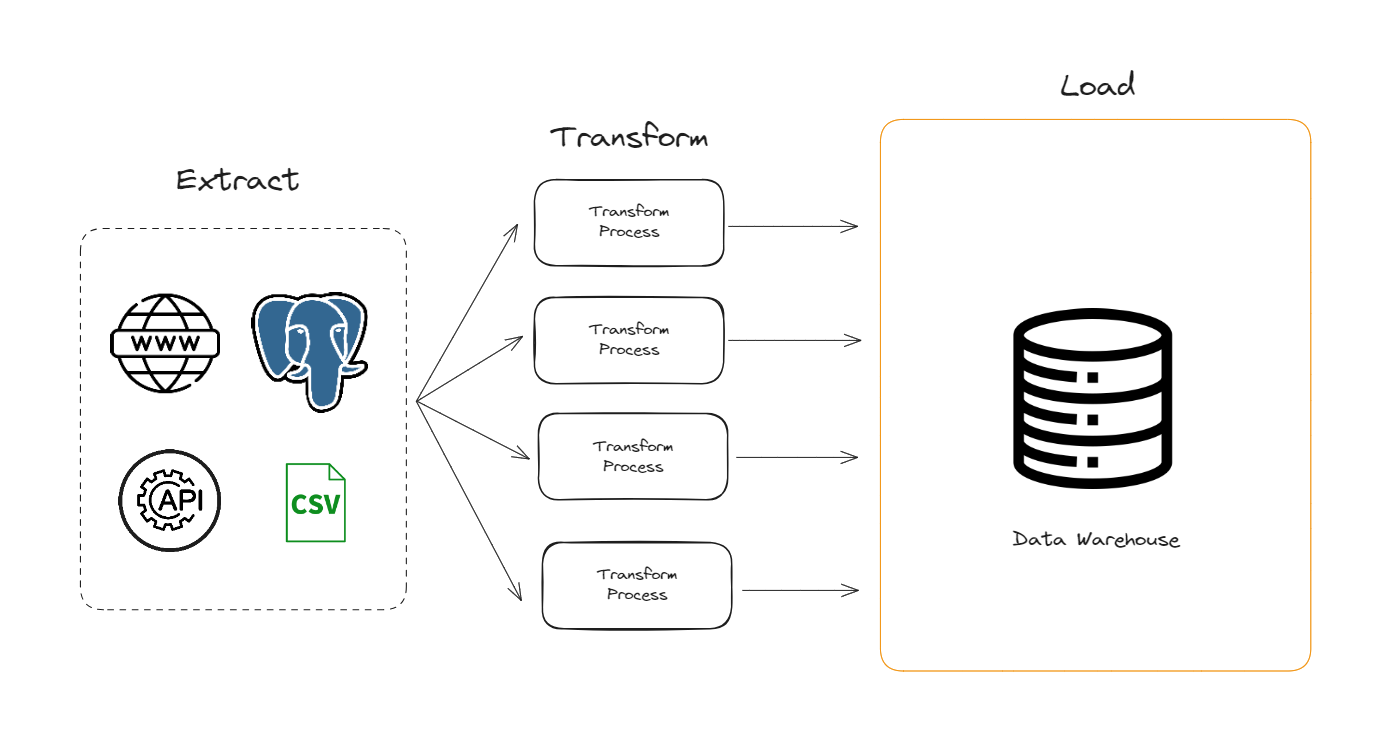
*Expected Time to Start: Week 4*

*Estimated Time of Completion: End of Week 5*

Setelah memikirkan solusi untuk ETL Pipeline, anda harus membuat Design ETL Pipeline terlebih dahulu. Untuk membuat Design ETL Pipeline, anda bisa menggunakan beberapa tools:

* [draw.io](https://app.diagrams.net/)
* [Excalidraw](https://excalidraw.com/)

**Contoh output dari ETL Pipeline Design**



Langkah #3 - ETL Implementation & ETL Scheduling

*Expected Time to Start: Week 5*

*Estimated Time of Completion: End of Week 7*

**ETL Implementation**

1. Setelah anda membuat Design ETL Pipeline, **anda akan implementasi** dengan menggunakan Python, Luigi, dan library / tools lain yang dibutuhkan. Pada proses ini anda bisa melakukan:

* Lakukan proses explore data yang akan digunakan
* Lakukan research mengenai metode apa yang dipakai
* Membuat table pada database yang digunakan untuk menampung semua data yang telah diproses
* Dsb

1. Pada proses Transform, **anda diberi kebebasan** untuk melakukan proses apa saja asalkan sesuai dengan requirements yang sudah anda tulisan.
2. Anda harus membuat satu database saja, karena akan digunakan untuk menampung seluruh data yang telah melewati proses ETL. Pastikan anda sudah menyiapkan schema dari masing - masing tablenya.
3. Expected output dari database ini adalah dalam bentuk Dockerfile atau Docker Image dari database yang digunakan pada proses **Load**

Anda diberi kebebasan untuk membuat schema table sesuai dengan kebutuhan, tetapi ada **satu kolom yang harus dimasukkan** di masing - masing table nya yaitu kolom **created\_at**

**Contoh schema:**

|  |
| --- |
| CREATE TABLE IF NOT EXISTS table\_name\_1(     ...,     ...,     created\_at timestamp NOT NULL DEFAULT now()  )  CREATE TABLE IF NOT EXISTS table\_name\_2(     id uuid not null primary key,     product\_name varchar not null,     created\_at timestamp NOT NULL DEFAULT now()  ) |

**ETL Scheduling**

1. Setelah membuat ETL Pipeline, anda harus membuat schedule untuk Pipeline yang dibuat
2. Anda bisa menggunakan **Crontab** untuk menjalankan Pipeline pada periode tertentu
3. Jika anda membuat ETL Pipeline ini menggunakan Virtual Environment, kami sarankan untuk membuat **Shell Script** agar dibungkus menjadi satu
4. Berikan **screenshot** terhadap crontab yang sudah dibuat di dokumentasi nanti

**Contoh Shell Script:**

|  |
| --- |
| #!/bin/bash  # Virtual Environment Path  VENV\_PATH="/path/to/venv/bin/activate"  # Activate venv  source "$VENV\_PATH"  # set python script  PYTHON\_SCRIPT="/path/to/script/script.py"  # run python script and logging  python "$PYTHON\_SCRIPT" >> /path/to/log/logfile.log 2>&1 |

**Contoh Screenshot Crontab:**



Langkah #4 - Testing Scenario

*Expected Time to Start: Week 7*

*Estimated Time of Completion: End of Week 8*

Ada beberapa Testing Scenario yang harus yang harus dilakukan dan ETL Pipeline harus melewati proses ini terlebih dahulu untuk dikatan ETL Pipeline yang anda buat sudah benar.

1. **Scenario 1**

Pada scenario ini cukup simple, pastikan ETL Pipeline sudah berjalan dan data sudah berpindah ke satu database yang sama

1. **Scenario 2**

Pada scenario 2 ini, pastikan ETL Pipeline anda sudah dilakukan proses scheduling. Karena, pada scenario ini anda akan menambahkan data baru pada database Marketing. Scenario ini dilakukan untuk memastikan apakah ETL Pipeline yang anda buat benar - benar berjalan dengan schedule dan apakah benar bisa mengambil data yang terbaru.

Tetapi harus diperhatikan, pastikan data yang masuk atau record yang masuk hanyalah data yang terbaru saja. Semisal, ada satu record baru yang masuk maka data yang terupdate hanya satu record yang baru saja atau istilah ini disebut dengan **UPSERT**

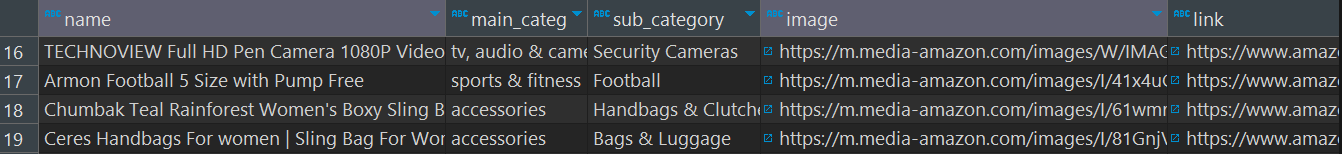
Data yang dimasukkan untuk melakukan scenario 2 adalah sebagai berikut

|  |
| --- |
| INSERT INTO amazon\_sales\_data ("name",main\_category,sub\_category,image,link,ratings,no\_of\_ratings,discount\_price,actual\_price)  values ('Testing Product', 'Testing Category', 'Testing Sub Category', 'https://sekolahdata-assets.s3.ap-southeast-1.amazonaws.com/notebook-images/mde-intro-to-data-eng/testing\_image.png', 'https://pacmann.io/', 5, 30, 450, 1000) |

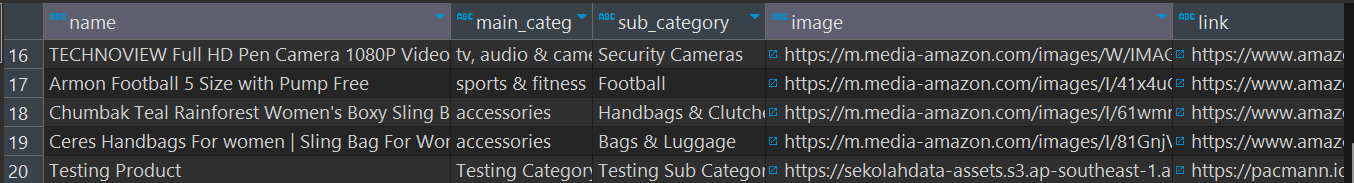
Setelah anda melakukan scenario 2 dan berhasil menambahkan data ke database Load, anda bisa screenshot hasilnya dan masukkan ke dalam dokumentasi.

**Output**:

Contoh Database Load sebelum ditambahkan record baru di source pertama



Contoh Database Load setelah ditambahkan record baru dari source pertama



Untuk penggunaan UPSERT dengan menggunakan Pandas, bisa menggunakan library berikut <https://github.com/ThibTrip/pangres>

Outcome Project

Setelah Anda mengerjakan itu semua, kami ingin Anda dapat merangkum hasilnya dalam sebuah *Dokumentasi Project (Github)* & *presentasi*.

Dokumentasi Project (Github)

* Merangkum hasil pengerjaan project tersebut dalam bentuk Dokumentasi github.
* Dokumentasi dalam bentuk **README.md** **bukan pdf**
* Dokumentasi tersebut berisi :
  + Requirements Gathering & Solution
    - Source data berasal
    - Solusi apa yang di propose
    - Metode apa saja yang dilakukan pada tahapan ETL
    - dsb
  + Design ETL Pipeline, pada tahapan ini anda bisa memasukkan gambar design dari ETL Pipeline yang dibuat
  + Cara kerja dari ETL Pipeline yang dibuat dan bagaimana cara menjalankannya
    - Stack atau tools atau library yang digunakan apa
    - Snippet Code dari masing - masing task
  + Hasil Output dari masing - masing scenario
* Push ke Repository Github :
  + Project
  + Dokumentasi
* Referensi Github:
  + <https://github.com/shandytepe/simple-luigi-pipeline>
  + <https://github.com/san089/goodreads_etl_pipeline>
* Referensi menulis Dokumentasi yang Baik dan Benar :
  + [Design Docs at Google](https://www.industrialempathy.com/posts/design-docs-at-google/)
  + [GitHub - jamiebuilds/documentation-handbook: How to write high-quality friendly documentation that people want to read.](https://github.com/jamiebuilds/documentation-handbook)

Presentasi

* Penjelasan project yang Anda lakukan secara ringkas.
* Berdurasi maksimal 10 menit.
* Berisi:
  + Pengenalan diri
  + Background Project
  + Penjelasan langkah-langkah penyelesaian project.
  + Demonstrasi dari project yang telah dibuat
* Di upload ke Youtube.
  + Judul  
    *ETL Pipeline* Pacmann - Intro to Data Engineer
  + Permission  
    Set sebagai publik agar dapat tim Pacmann periksa.

Evaluasi

Kami akan mengevaluasi beberapa komponen berikut.

Fitur penilaian                   Bobot [%]

1. Requirements Gathering & Solution 10
2. ETL Pipeline Design:
   1. Membuat Design ETL Pipeline 10
   2. Penjelasan mengenai Design ETL Pipeline 10
3. ETL Implementation:
   1. Membuat ETL Pipeline 20
   2. Extract seluruh source data 10
   3. ETL Scheduling 5
   4. Membuat Dockerfile / Docker Image 5
4. Testing Scenario 10
5. Publikasi Project di Github 10
6. Video Presentasi 10

Tools

Anda dapat menggunakan tools berikut untuk menunjang proses pengerjaan proyek yang dilakukan

* Python
* Luigi
* BeautifulSoup
* Docker
* PostgreSQL
* DBeaver

**Link Form Submission:** [**Submission**](https://forms.gle/ZGdVumo4C1Ajs4zm7)

Need Assistance?

Jika anda memiliki pertanyaan atau kesulitan dalam mengerjakan project ini, Anda bisa memanfaatkan fasilitas Whatsapp Help Desk/Student Support : <https://bit.ly/PacmannStudentSupport>