



INDONESIAN TECH JOB MARKET ANALYSIS

FINAL PROJECT

FITRAN DWI PRAMAKRISNA





FITRAN DWI PRAMAKRISNA

Education

- S1 Software Engineering - Institut Teknologi Telkom Purwokerto (2018 - 2022)

Work Experience

- IT Staff Support - Pelindo Solusi Digital (2022 - *present*)

Overview Projects

Web Scraping Sneakers Product

Scraping sneakers product using Selenium and Requests library

ETL with Airflow

Creating an ETL pipeline with dynamic DAGs in Airflow

Amazon Sales Report Data Modeling

Creating dimensional and fact table in Google BigQuery using DBT from Amazon Sales Report Dataset

Big Data Analytics using PySpark

Creating a customer retention analysis using Spark and Airflow for the ETL process with batch processing

Stream Processing using Spark Structured Streaming

Creating a real-time data streaming ingesting data from Apache Kafka that applies aggregation calculation for the output



PROJECT **BACKGROUND**



WHAT AND WHY THIS PROJECT?

Proyek ini bertujuan untuk membangun sebuah data pipeline ETL (Extract, Transform, dan Load) yang berfungsi dalam mengumpulkan, memproses, dan menyajikan data terkait pasar kerja di bidang teknologi di Indonesia. Dalam proses ini, data akan diekstraksi dari portal lowongan pekerjaan. Hasil akhir dari pipeline ini adalah data yang telah terstruktur dan dimuat ke dalam sistem penyimpanan yang memungkinkan eksplorasi lebih lanjut menggunakan data visualization tool.

Dengan adanya pipeline ini, proyek ini diharapkan dapat memberikan wawasan mendalam mengenai tren perekrutan, keterampilan yang paling dibutuhkan, serta perkembangan pasar kerja di industri teknologi di Indonesia.



PROBLEM **STATEMENT**



PROBLEMS

Masalah yang tercipta yang mendukung project ini dibuat:

1. **Sulitnya Memantau Perubahan dan Tren Pasar Kerja**, tren pasar kerja di industri teknologi berkembang dengan cepat. Namun, tidak ada sistem otomatis yang dapat mengumpulkan dan memproses data secara berkala, sehingga banyak insight yang seharusnya bisa didapatkan menjadi kurang relevan karena datanya tidak ter-update.
2. **Membantu Pengambilan Keputusan Berdasarkan Data bagi Job Seeker**, dapat melihat pekerjaan mana yang sedang *hot*, keterampilan dan jumlah pengalaman kerja yang paling banyak dibutuhkan. Hal ini dapat membuat job seeker bisa menyesuaikan pengalaman dan juga skill mereka
3. **Tidak Adanya Sumber Data Terpusat untuk Job Market Teknologi**, Saat ini, informasi tentang lowongan pekerjaan di industri teknologi tersebar di berbagai platform, seperti situs job portal dan website perusahaan. Namun, tidak ada satu sumber data yang mengintegrasikan informasi ini secara komprehensif untuk memudahkan analisis dan pengambilan keputusan bagi pencari kerja



DATA & PLATFORM **UNDERSTANDING**



WHAT KIND OF DATA??

Data diperoleh dari metode web scraping menggunakan Request pada python. Website yang akan di-scraping tentunya adalah website-website yang legal untuk di-scraping

Kalibrr

The screenshot displays the Kalibrr job portal interface. At the top, there is a navigation bar with the Kalibrr logo, a search bar labeled "Job Position", and buttons for "Masuk" (Login) and "Daftar" (Register). Below the navigation bar, there are several category tiles: "Featured Jobs", "Remote Jobs", "Hybrid", "Fresh Grads" (marked with a "NEW" badge), "All Jobs", and "New Jobs". The main content area features a grid of job listings. Each listing includes the company logo, job title, location, salary status, contract type, and the time since the recruiter was last active. The listings shown are:

- Security Engineer – Red Team** by FinAccel, located in South Jakarta, Indonesia. Salary is not disclosed. Full-time position. Recruiter was active a day ago. Application deadline: 1 Jul. Level: Mid-Senior Level Manager.
- Data Governance Analyst** by Kredit Pintar, located in Kota Jakarta Selatan, Indonesia. Salary is not disclosed. Full-time position. Recruiter was active 10 hours ago. Application deadline: 4 Apr. Level: Supervisor / Asisten Manager.
- IT Operations for Knowledge...** by PT Bumi Amarthha Teknologi Mandiri, located in South Jakarta, Indonesia. Salary is not disclosed. Contract position. Recruiter was active 10 hours ago. Application deadline: 8 Apr. Level: Supervisor / Asisten Manager.
- Head of AI** by PCS Group, located in Jakarta Selatan, Indonesia.
- IT Infrastructure Development...** by Indonesia Stock Exchange, located in Jakarta Selatan, Indonesia.
- Frontend Developer** by Magna Solusi Indonesia.

WHAT KIND OF DATA??

Data diperoleh dari metode web scraping menggunakan Request pada python. Website yang akan di-scraping tentunya adalah website-website yang legal untuk di-scraping

Dealls

Dealls! Loker Mentoring Perusahaan Events AI CV Analyzer IDN Masuk Daftar Untuk Perusahaan

Loker Butuh Cepat

IT Maintenance
Nuxatech

Penuh waktu

On-site • Lombok Barat Regency

Min. Freshgrad

Rp2.500.000 – 3.000.000

Rekruter aktif 3j lalu

Pelamar Masih Sedikit

Business System Intern
PT Darya-Varia Laboratoria Tbk

Magang

On-site • Jakarta Selatan

Min. 3rd Year College Student

Negotiable

Rekruter aktif 4j lalu

Pelamar Masih Sedikit

Associate AI Engineer (Jogja)
BitHealth

Penuh waktu

On-site • Yogyakarta

Freshgrad

Rp3.500.000 – 4.000.000

Rekruter aktif 5j lalu

Pelamar Masih Sedikit

Technical Lead
SG-EDTS

Penuh waktu

On-site • Jakarta Pusat

Min. 4 years of experience

Negotiable

Rekruter aktif 4j lalu

Pelamar Masih Sedikit

IT Internship - Jakata
PT Kalventis Sinergi Farma

Magang

On-site • Jakarta

Min. 3rd Year College Student

Negotiable

Rekruter aktif 5j lalu

Pelamar Masih Sedikit

Programmer (penempatan...
Universitas Ciputra

Penuh waktu

On-site • Surabaya

Min. 1 years of experience

Negotiable

Rekruter aktif 1j lalu

Pelamar Masih Sedikit

Front End Engineer - Marketpla...
Sea

Penuh waktu

On-site • Jakarta

Min. 1 years of experience

Pelamar Masih Sedikit

Senior Site Reliability Engineer
Reku

Penuh waktu

Remote

Min. 5+ years of experience

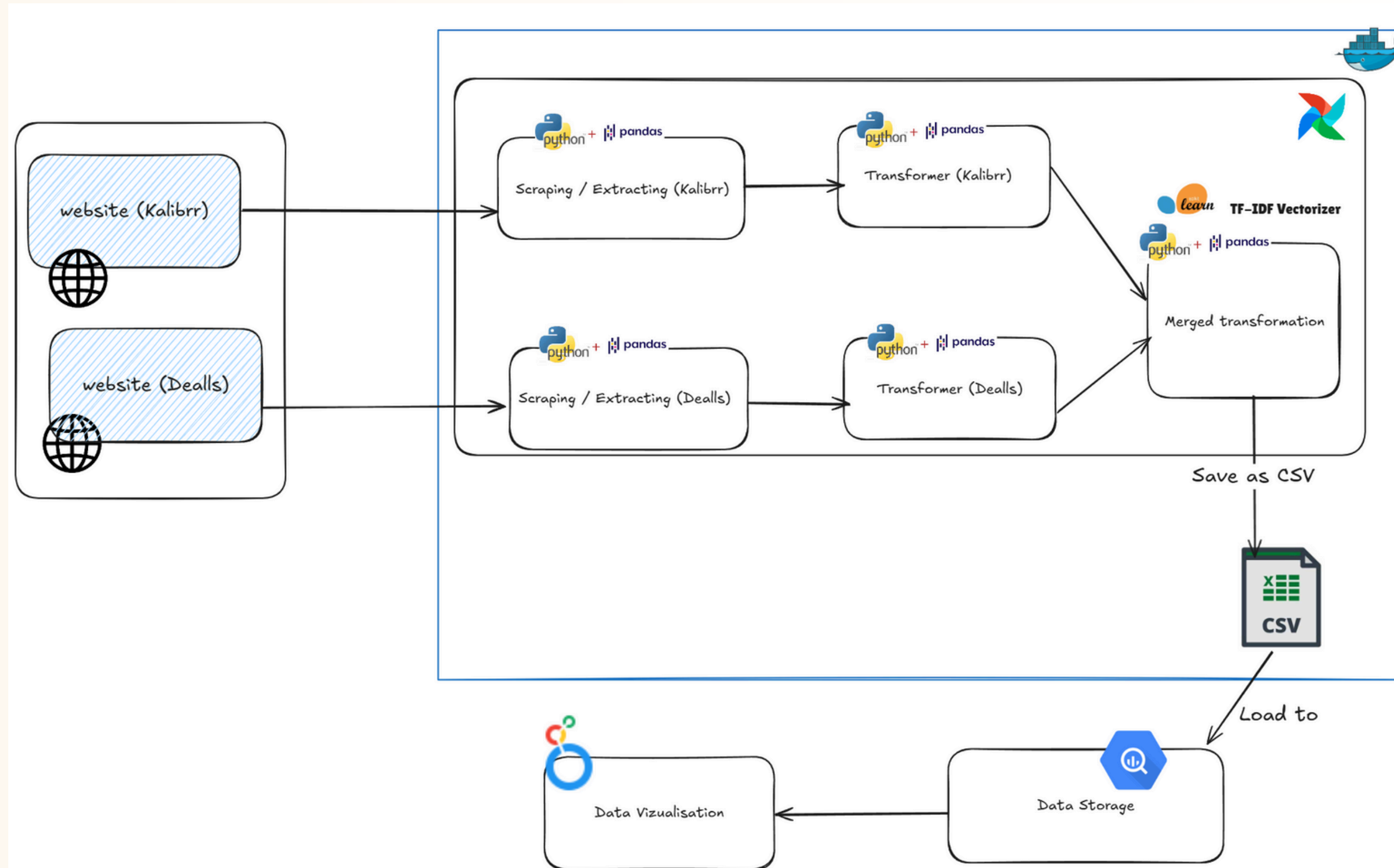
Documentation & Certification...
Frogs Indonesia

Penuh waktu

On-site • Yogyakarta

Min. Freshgrad

DATA PLATFORM





TRANSFORMATION & **CONSIDERATIONS**



TRANSFORMATION & CONSIDERATION

1. Processing Method

- Multithreading to scrap different pages
- Batch Processing (daily)

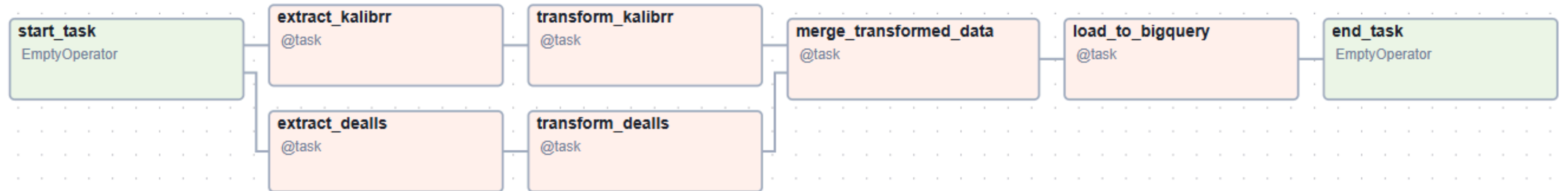
2. Tools Use

- Requests: To scrap the websites
- Pandas: To process and transform a medium-level-sized dataset

3. Key Transformation

- Handling Duplicate Values
- Standarization string colum job_title (using Tf-Idf vectorizer), and industry
- Replacing 'kabupaten', 'kota', 'regency' and grouping Area of Jakarta to be only 'Jakarta'

BATCH PROCESSING





DATA MODELING (**BUSINESS**)



DATA MODELING

jobs_detail

job_title	varchar
company_name	varchar
city	varchar
job_posted	datetime
job_type	varchar
level_experience	varchar
education_min_lvl	varchar
industry	varchar
min_salary	int
max_salary	int
source_job	varchar



dbdiagram.io



CONCLUSION & RECOMENDATION (**BUSINESS**)



| CONCLUSION

The data pipeline extracts data from two different web sources, **Kalibrr** and **Dealls**, using the Python library Requests. After the data is successfully extracted in each task, it undergoes a transformation process to filter out the words "kabupaten", "regency", and "kota", and to group all Jakarta areas into a single "Jakarta" category in the city column. Additionally, transformations are applied to the education_min_lvl and level_experience columns to convert their values into more meaningful ones.

Next, the data is merged and further transformed using scikit-learn (TF-IDF Vectorizer), particularly for the job_title and industry columns. The final transformation step involves removing duplicate values. The cleaned data is then stored in BigQuery and subsequently visualized using Looker Studio.

DATA VIZ

Looker Studio (link access)



RECOMMENDATION

- To perform web scraping where the data source can be directly retrieved from its API, the easiest way is to use the Requests library.
- Using multi-threading allows executing the scraping script for multiple pages simultaneously in a single iteration.
- With Looker Studio, the data source can be easily connected to Google BigQuery. If there are any changes in the data, we can simply click "Refresh dataset" on the Looker Studio dashboard.
- When transforming the job_title column, TF-IDF Vectorizer is needed to normalize strings if the job_title does not exist in the mapping_job_title.py file.



THANK YOU |

<https://www.datascienceportfol.io/fitranpramakrisna>



fitranpramakrisna@gmail.com



[Fitran Dwi Pramakrisna](#)



Jakarta

