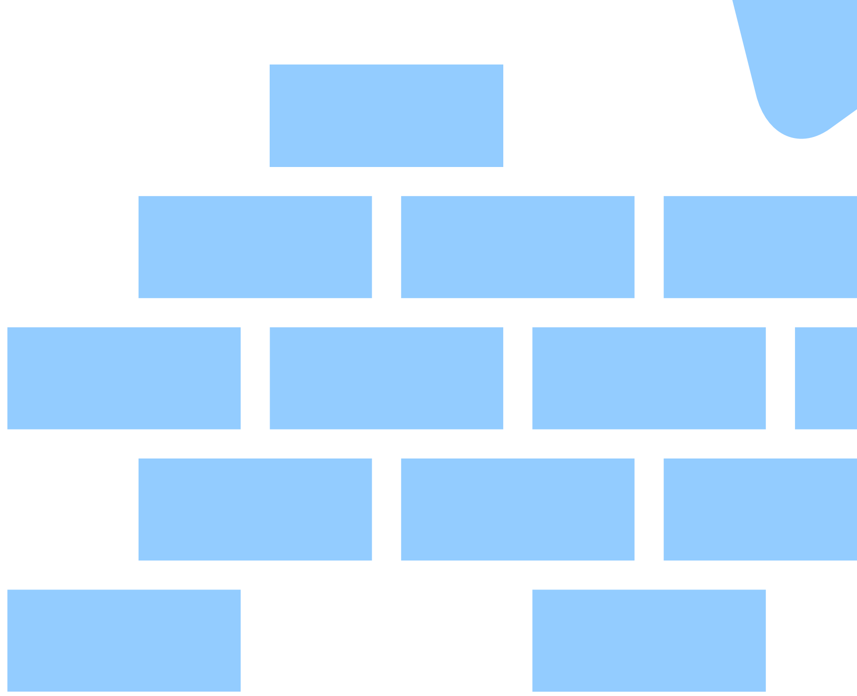
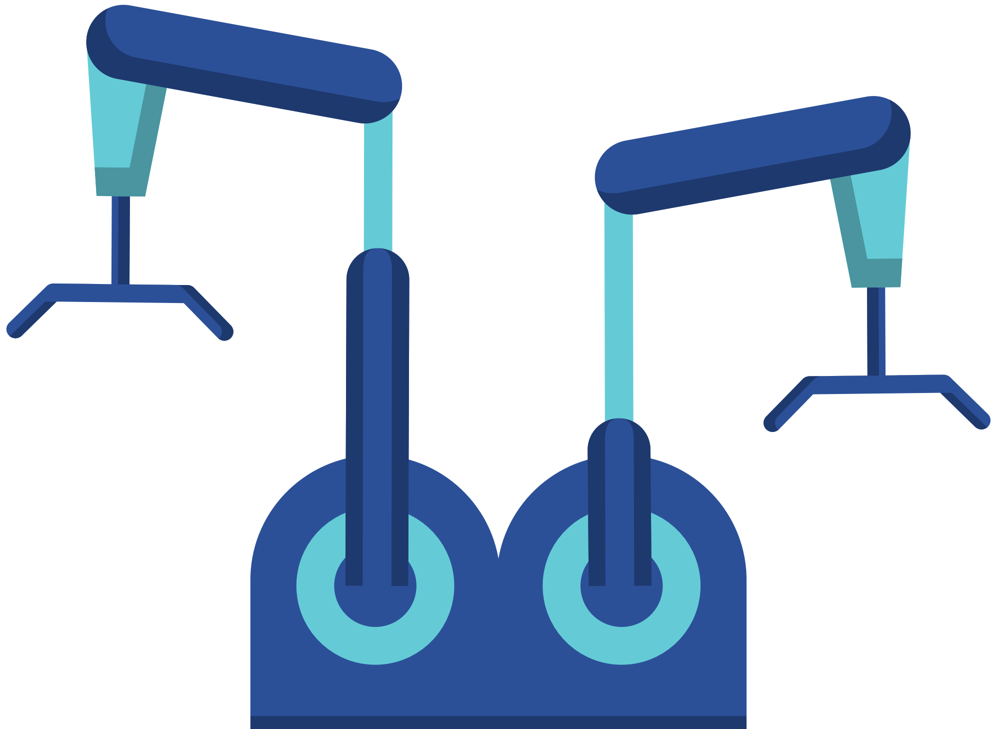


Dynamic ELT Jobstreet Data using Airflow and GCP

Muhammad Sidqy D

https://github.com/msidqyd/ELT-to-GBQ-Jobstreetscrape-Detailed



Project Overview

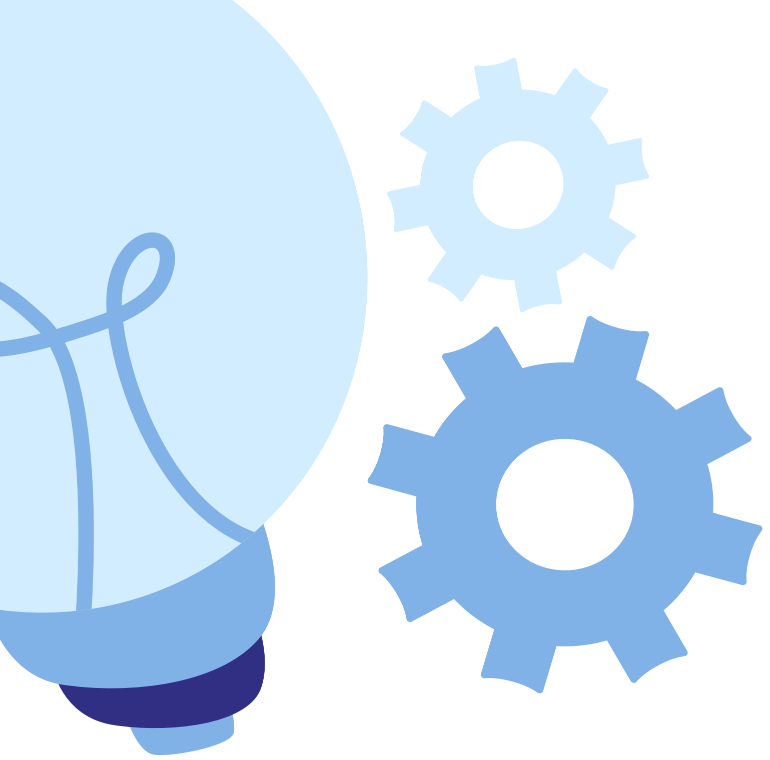
- Modular and scalable ELT pipeline
- Automate job data collection
- Multithreaded Web Scraping with Selenium
- Branching DAGs for dynamic input format handling
- Incremental & full load
- Stored intermediate files in local staging as Parquet
- transformations and load data to Big Query
- Data validation tasks after loading
- Data Modelling : Silver -> Bronze -> Gold

Education

Physics Engineering, institute Technology Sepuluh Nopember

Working

- Management Trainee – Korea Tomorrow & Global Indonesia
- Retail Sales Supervisor – Korea Tomorrow & Global Indonesia
- Productivity Specialist - Korea Tomorrow & Global Indonesia
- Field Operation Specialist - Korea Tomorrow & Global Indonesia



Background

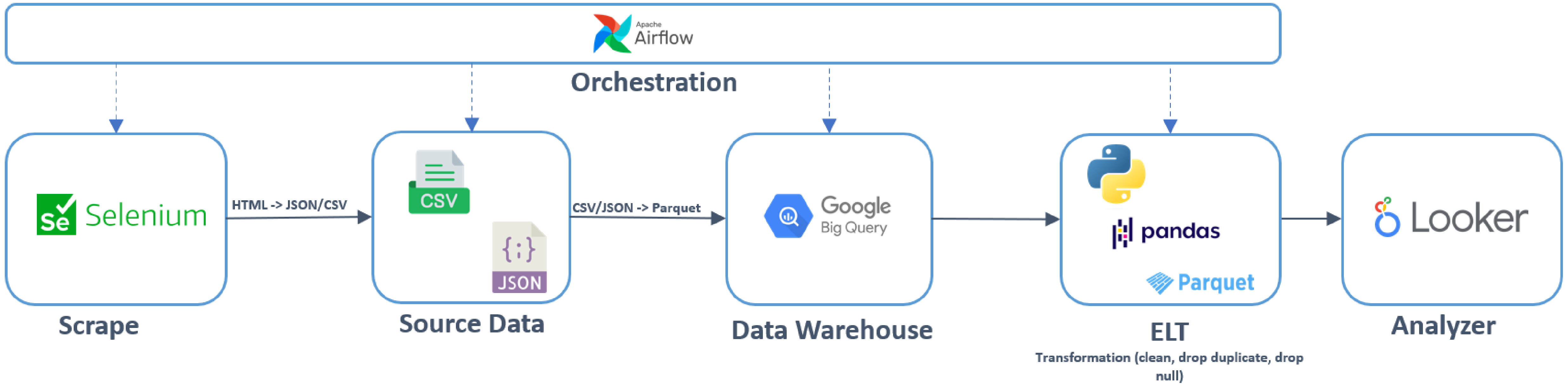
- Indonesia's unemployment rate in Feb 2025 reached 4.76%, above the healthy range of 3.5%–4.5%.
- The government needs to promote digital upskilling, especially in high-demand roles like Data Engineering, which is projected to be a top-growth job for more than 30%.
- remote job for data engineer in US reach 10% in 2024, but the number has plummet less than 2% in 2025.
- This project aims to collect and analyze Data Engineer job postings to identify the most in-demand skills and tools.

Source :
https://www.bps.go.id/en/pressrelease/2025/05/05/2432/unemployment-rate-was-4-76-percent--the-average-wage-of-employees-was-3-09-million-rupiah-.html
https://www.avatrade.com.au/education/economic-indicators/fundamental-indicators/unemploymentrate#:~:text=The%20optimum%20unemployment%20rate%20for,them%20to%20employ%20more%20people.
https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf
https://365datascience.com/career-advice/data-engineer-job-outlook-2025/

Problem Formulation

Gap	Opportunity	Solution	Impact	How to	Project Output
There's a gap reach 0.26% of unemployee rate need to reduce to reach the bare minimum of healthy economy.	The digital field has thrive, one of them is data engineer	Offering data for learning and coaching tools and skill needed as data engineer	The people who want to learn about data engineer could know the trend skill and demand for data engineer role.	1. Get the trend skills and tools from the job platform. 2. Build pipeline to ingest data and load it to database. 3. Transform data for collection and analyzing the data.	Show keywords with the frequencies for skill and tools data from job posting in indonesia or overseas platform.
	Data Engineer role projected to be a top growth job which reach more than 30% growth	Conduct training with curricullum alligned to data engineer job-market trend.	Private, public, and third sector, could have training with curriculum the most demanding and tren tools to accelerate the student learning process.		Dashboard for another information such as total jobs available, total jobs each location, and average wage. Also utilize the data for future research (forecast the job availability)
	There are 10% remote jobs in based on job posting from indeed in US for data engineer jobs	Generate indonesian talent to compete in indonesia or global job market.	Actively for contribution of reducing the unemployement rate for current and future condition		Collect and stay updated with job list data to domestic or global, so the talent couldn't miss the job posting or opportunity

Data Platform Understanding



Apache Airflow

Orchestration

Selenium

HTML -> JSON/CSV

Scrape

CSV

JSON

CSV/JSON -> Parquet

Source Data

Google Big Query

Data Warehouse

pandas

Parquet

ELT

Transformation (clean, drop duplicate, drop null)

Looker

Analyzer

Architecture

1. Scraping

JobStreet search pages and job detail pages (dynamic web content).

Targeting job postings with the keyword: “Data Engineer” in Indonesia and Singapore with multithreading regarding to scrape job page.

2. Ingestion

Scheduled with Apache Airflow to run regularly and support incremental scraping.

Data saved as raw JSON or CSV.

3. Storage (Bronze Layer)

Raw data stored locally or in Google Cloud Storage (GCS).

File formats: Parquet/CSV for efficient querying and processing.

4. Transformation (Silver Layer)

Parsed and cleaned text (job titles, descriptions, company names, locations).

Extracted structured features: skills, tools, experience levels, job types.

Transformation done using Pandas, SQL

5. Gold (Gold Layer)

Final dataset loaded into BigQuery for analytics and dashboarding.

Allows for aggregation, keyword frequency, job listing

6. Google Big Query

Word Frequencies : identified skills, tools, traits in-demand at the market.

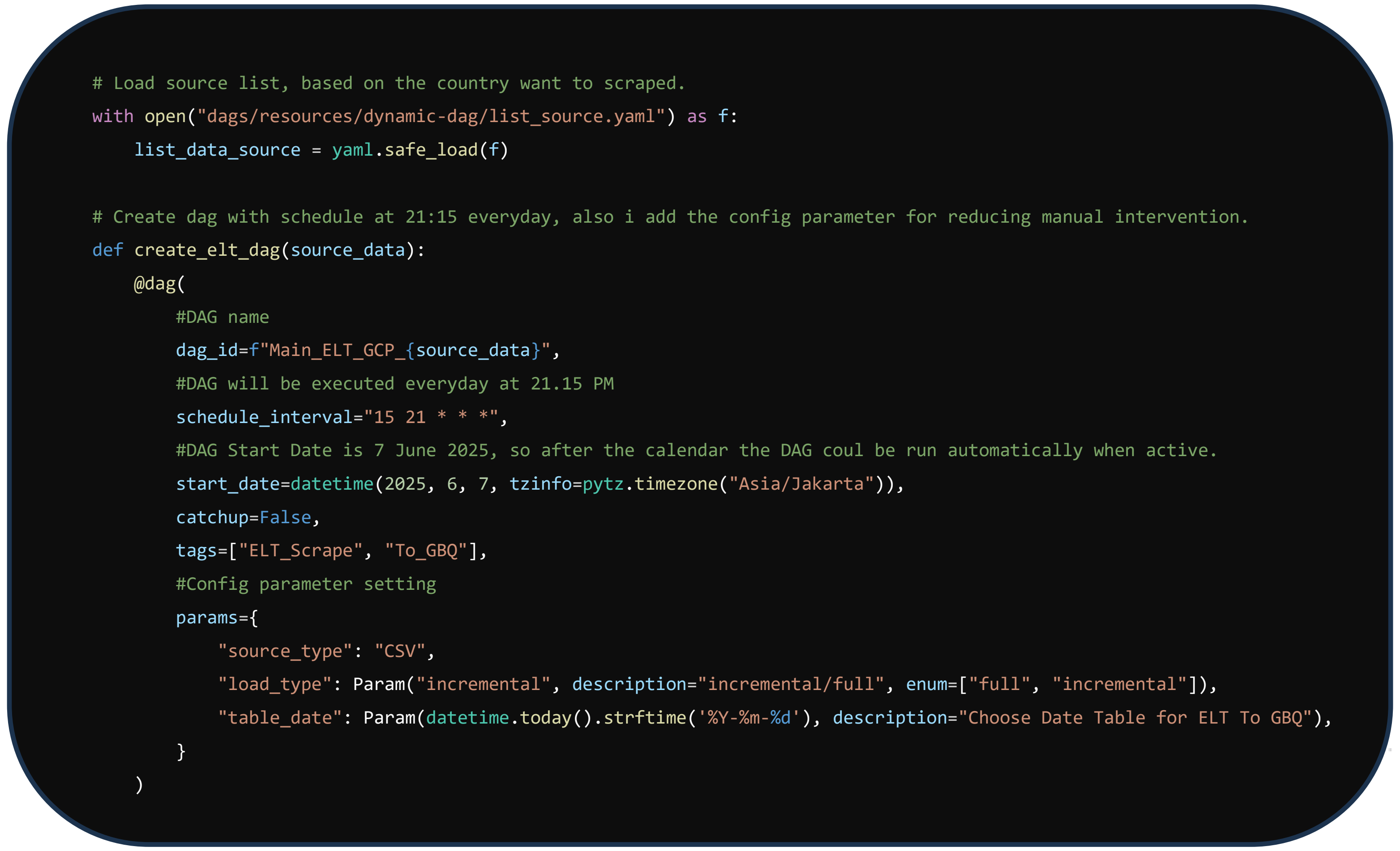
Dashboard: Further analysis current condition for data engineer job.

Job Market Search: Updated and sorted Data Engineer Job with all information.

Data Platform Understanding

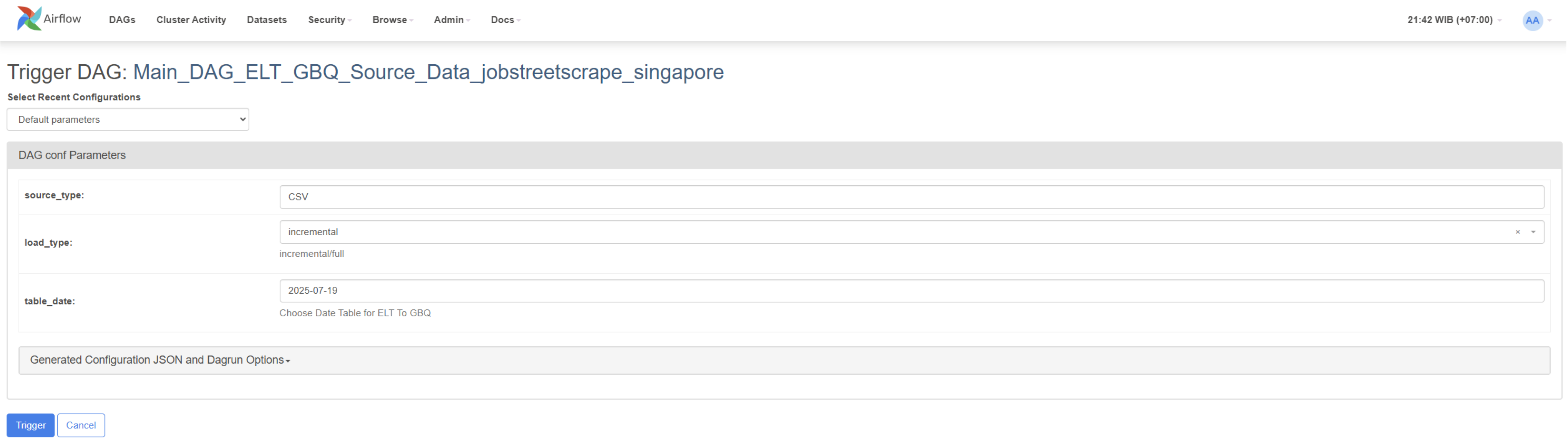
Features

1. Performed multithreading using Selenium (headless browser automation). Scheduled with airflow with config parameter
2. Dynamic configuration and parameterization
3. Modular & Scalable ELT architecture
4. Automate job everyday 21.15 O'clock



```python
# Load source list, based on the country want to scraped.
with open("dags/resources/dynamic-dag/list_source.yaml") as f:
    list_data_source = yaml.safe_load(f)

# Create dag with schedule at 21:15 everyday, also i add the config parameter for reducing manual intervention.
def create_elt_dag(source_data):
    @dag(
        #DAG name
        dag_id=f"Main_ELT_GCP_{source_data}",
        #DAG will be executed everyday at 21.15 PM
        schedule_interval="15 21 * * *",
        #DAG Start Date is 7 June 2025, so after the calendar the DAG coul be run automatically when active.
        start_date=datetime(2025, 6, 7, tzinfo=pytz.timezone("Asia/Jakarta")),
        catchup=False,
        tags=["ELT_Scrape", "To_GBQ"],
        #Config parameter setting
        params={
            "source_type": "CSV",
            "load_type": Param("incremental", description="incremental/full", enum=["full", "incremental"]),
            "table_date": Param(datetime.today().strftime('%Y-%m-%d'), description="Choose Date Table for ELT To GBQ"),
        }
    )
```



Trigger DAG: Main_DAG_ELT_GBQ_Source_Data_jobstreetscrape_singapore

Select Recent Configurations

Default parameters

DAG conf Parameters

source_type: CSV

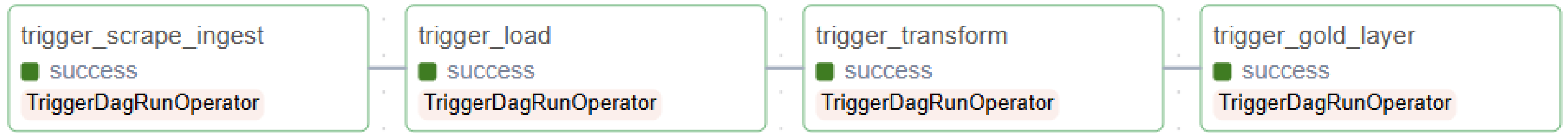
load_type: incremental

incremental/full

table_date: 2025-07-19

Choose Date Table for ELT To GBQ

Generated Configuration JSON and Dagrun Options



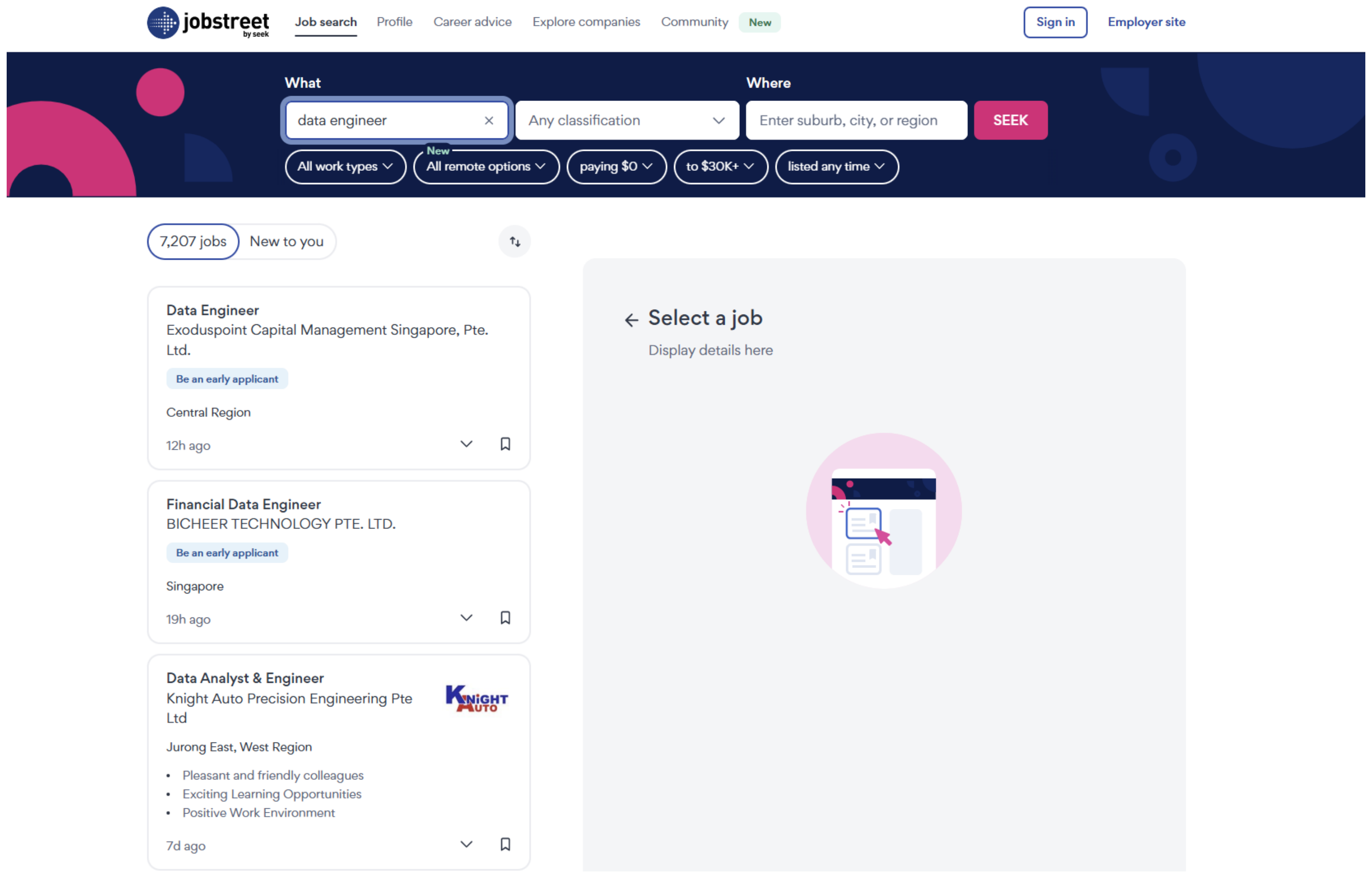
trigger_scrape_ingest
🟩 success
TriggerDagRunOperator

trigger_load
🟩 success
TriggerDagRunOperator

trigger_transform
🟩 success
TriggerDagRunOperator

trigger_gold_layer
🟩 success
TriggerDagRunOperator

Data Understanding



Data Engineer
Exoduspoint Capital Management Singapore, Pte. Ltd.
Be an early applicant
Central Region
12h ago

Financial Data Engineer
BICHEER TECHNOLOGY PTE. LTD.
Be an early applicant
Singapore
19h ago

Data Analyst & Engineer
Knight Auto Precision Engineering Pte Ltd
Jurong East, West Region
- Pleasant and friendly colleagues
- Exciting Learning Opportunities
- Positive Work Environment

7d ago

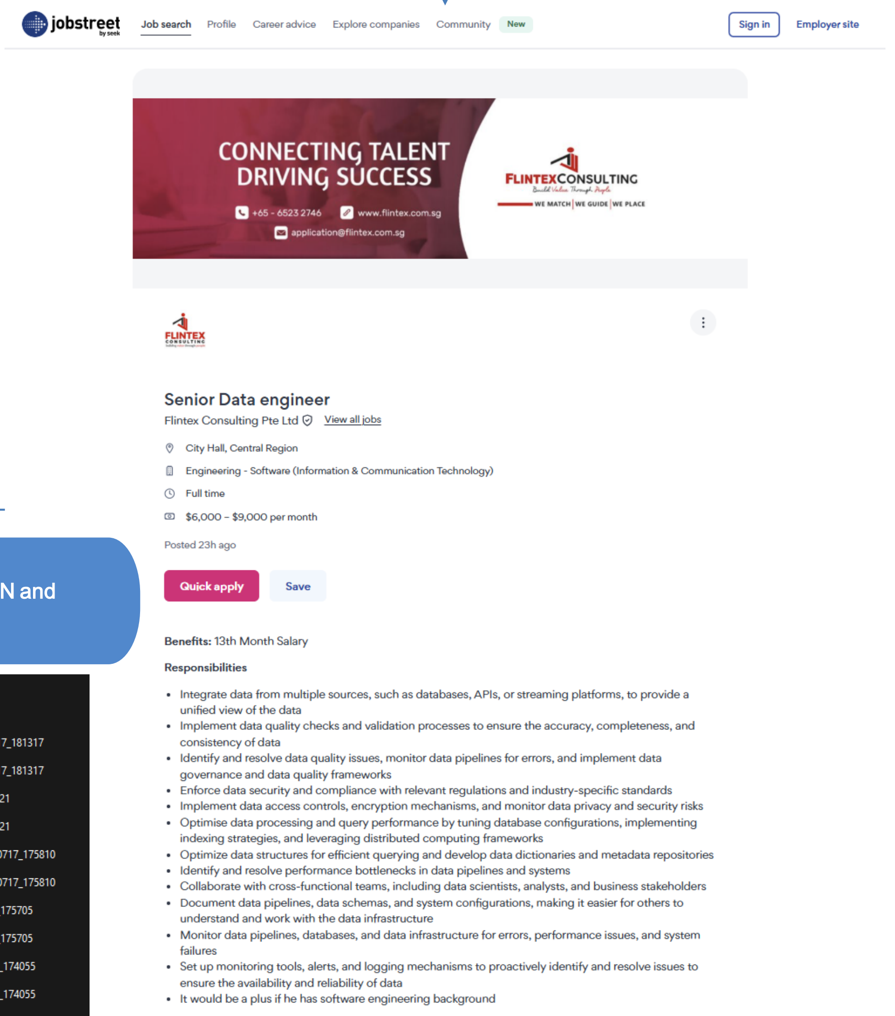
1. Search page

- job_id
- title — Job will selected if title match with more than 1 keywords
- url
- company_name — For incremental, filtered by max publish time from GBQ
- publish_time
- location

From URL use multithreading to open each job page

2. Job Page

- wage
- work_type
- job_desc



Senior Data engineer
Flintex Consulting Pte Ltd View all jobs

City Hall, Central Region
Engineering - Software (Information & Communication Technology)
Full time
$6,000 – $9,000 per month

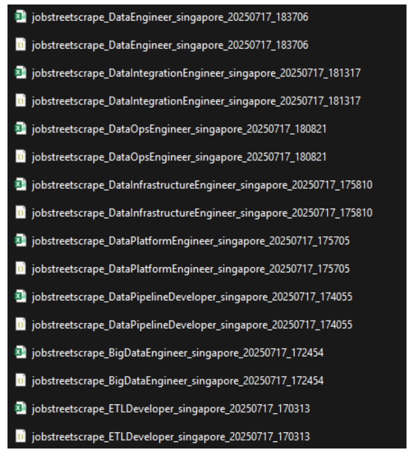
Posted 23h ago

Benefits: 13th Month Salary

Responsibilities

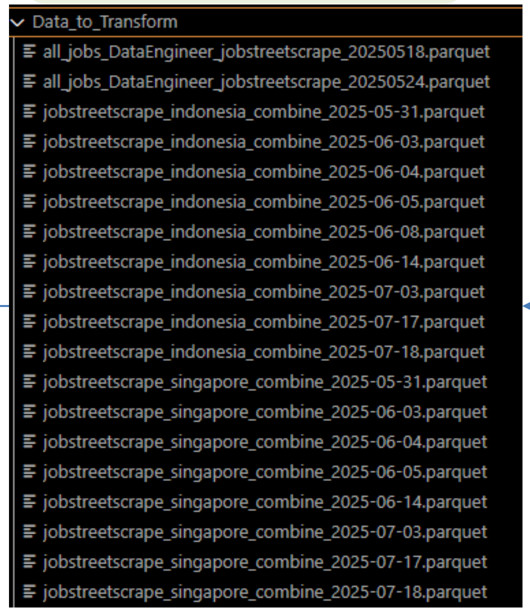
- Integrate data from multiple sources, such as databases, APIs, or streaming platforms, to provide a unified view of the data
- Implement data quality checks and validation processes to ensure the accuracy, completeness, and consistency of data
- Identify and resolve data quality issues, monitor data pipelines for errors, and implement data governance and data quality frameworks
- Enforce data security and compliance with relevant regulations and industry-specific standards
- Implement data access controls, encryption mechanisms, and monitor data privacy and security risks
- Optimise data processing and query performance by tuning database configurations, implementing indexing strategies, and leveraging distributed computing frameworks
- Optimize data structures for efficient querying and develop data dictionaries and metadata repositories
- Identify and resolve performance bottlenecks in data pipelines and systems
- Collaborate with cross-functional teams, including data scientists, analysts, and business stakeholders
- Document data pipelines, data schemas, and system configurations, making it easier for others to understand and work with the data infrastructure
- Monitor data pipelines, databases, and data infrastructure for errors, performance issues, and system failures
- Set up monitoring tools, alerts, and logging mechanisms to proactively identify and resolve issues to ensure the availability and reliability of data
- It would be a plus if he has software engineering background

Use Branch Operator CSV or JSON and combine all keywords data



```
jobstreetscrape_DataEngineer_singapore_20250717_183706
jobstreetscrape_DataEngineer_singapore_20250717_183706
jobstreetscrape_DataIntegrationEngineer_singapore_20250717_181317
jobstreetscrape_DataIntegrationEngineer_singapore_20250717_181317
jobstreetscrape_DataOpsEngineer_singapore_20250717_180821
jobstreetscrape_DataOpsEngineer_singapore_20250717_180821
jobstreetscrape_DataInfrastructureEngineer_singapore_20250717_175810
jobstreetscrape_DataInfrastructureEngineer_singapore_20250717_175810
jobstreetscrape_DataPlatformEngineer_singapore_20250717_175705
jobstreetscrape_DataPlatformEngineer_singapore_20250717_175705
jobstreetscrape_DataPipelineDeveloper_singapore_20250717_174055
jobstreetscrape_DataPipelineDeveloper_singapore_20250717_174055
jobstreetscrape_BigDataEngineer_singapore_20250717_172454
jobstreetscrape_BigDataEngineer_singapore_20250717_172454
jobstreetscrape_ETLDeveloper_singapore_20250717_170313
jobstreetscrape_ETLDeveloper_singapore_20250717_170313
```

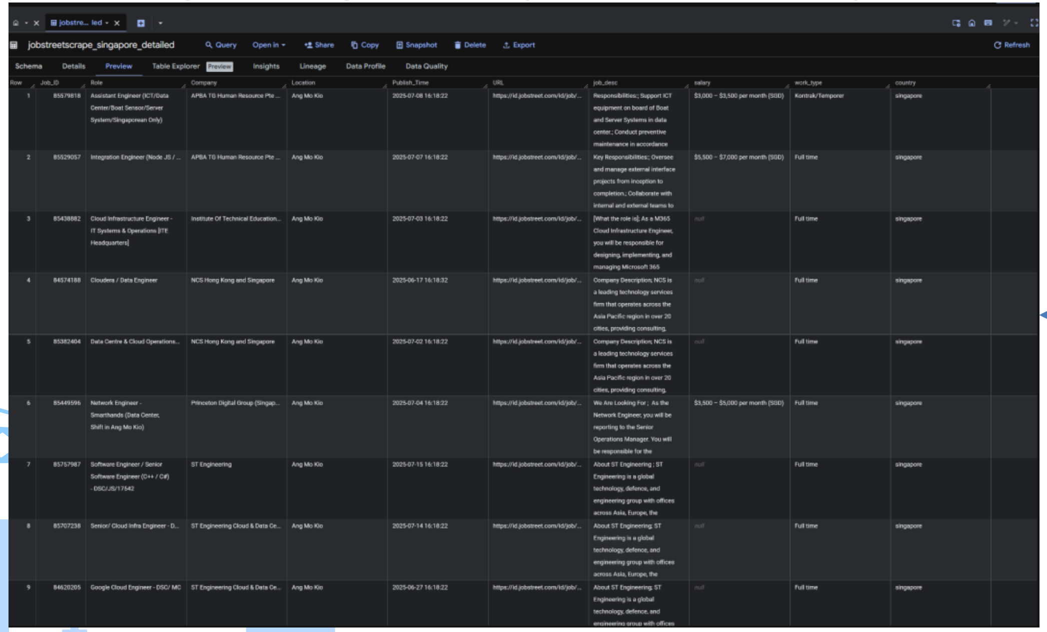
Data Staging



```
Data_to_Transform
  all_jobs_DataEngineer_jobstreetscrape_20250518.parquet
  all_jobs_DataEngineer_jobstreetscrape_20250524.parquet
  jobstreetscrape_indonesia_combine_2025-05-31.parquet
  jobstreetscrape_indonesia_combine_2025-06-03.parquet
  jobstreetscrape_indonesia_combine_2025-06-04.parquet
  jobstreetscrape_indonesia_combine_2025-06-05.parquet
  jobstreetscrape_indonesia_combine_2025-06-08.parquet
  jobstreetscrape_indonesia_combine_2025-06-14.parquet
  jobstreetscrape_indonesia_combine_2025-07-03.parquet
  jobstreetscrape_indonesia_combine_2025-07-17.parquet
  jobstreetscrape_indonesia_combine_2025-07-18.parquet
  jobstreetscrape_singapore_combine_2025-05-31.parquet
  jobstreetscrape_singapore_combine_2025-06-03.parquet
  jobstreetscrape_singapore_combine_2025-06-04.parquet
  jobstreetscrape_singapore_combine_2025-06-05.parquet
  jobstreetscrape_singapore_combine_2025-06-14.parquet
  jobstreetscrape_singapore_combine_2025-07-03.parquet
  jobstreetscrape_singapore_combine_2025-07-17.parquet
  jobstreetscrape_singapore_combine_2025-07-18.parquet
```

Use Branch Operator CSV or JSON and combine all keywords data

Google Big Query (Bronze layer)



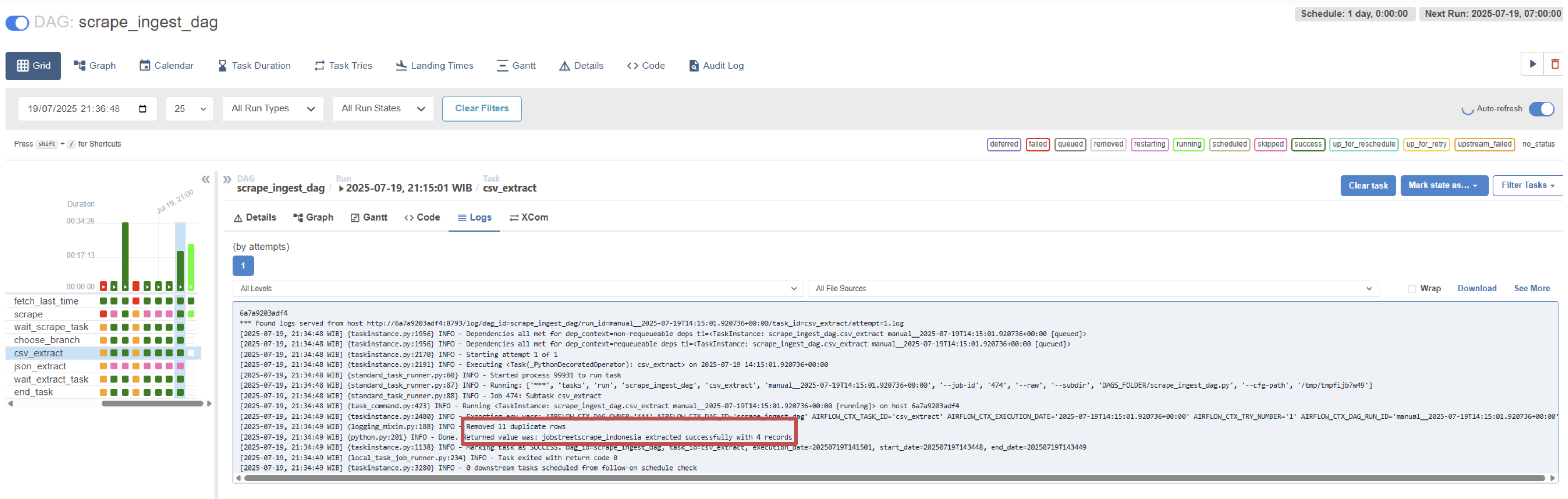
jobstreetscrape_singapore_detailed

Row	Job_ID	Role	Company	Location	Publish_Time	URL	job_desc	salary	work_type	country
1	85579818	Assistant Engineer (ICT/Data Center/Boat Sensor/Server System/Singaporean Only)	APBA TG Human Resource Pte ...	Ang Mo Kio	2025-07-08 16:18:22	https://id.jobstreet.com/id/job/...	Responsibilities:; Support ICT equipment on board of Boat and Server Systems in data center.; Conduct preventive maintenance in accordance	$3,000 – $3,500 per month (SGD)	Kontrak/Temporer	singapore
2	85529057	Integration Engineer (Node JS / ...	APBA TG Human Resource Pte ...	Ang Mo Kio	2025-07-07 16:18:22	https://id.jobstreet.com/id/job/...	Key Responsibilities:; Oversee and manage external interface projects from inception to completion.; Collaborate with internal and external teams to	$5,500 – $7,000 per month (SGD)	Full time	singapore
3	85438882	Cloud Infrastructure Engineer - IT Systems & Operations [ITE Headquarters]	Institute Of Technical Education...	Ang Mo Kio	2025-07-03 16:18:22	https://id.jobstreet.com/id/job/...	[What the role is]; As a M365 Cloud Infrastructure Engineer, you will be responsible for designing, implementing, and managing Microsoft 365	null	Full time	singapore
4	84574188	Cloudera / Data Engineer	NCS Hong Kong and Singapore	Ang Mo Kio	2025-06-17 16:18:32	https://id.jobstreet.com/id/job/...	Company Description; NCS is a leading technology services firm that operates across the Asia Pacific region in over 20 cities, providing consulting,	null	Full time	singapore
5	85382404	Data Centre & Cloud Operations...	NCS Hong Kong and Singapore	Ang Mo Kio	2025-07-02 16:18:22	https://id.jobstreet.com/id/job/...	Company Description; NCS is a leading technology services firm that operates across the Asia Pacific region in over 20 cities, providing consulting,	null	Full time	singapore
6	85449596	Network Engineer - Smarthands (Data Center, Shift in Ang Mo Kio)	Princeton Digital Group (Singap...	Ang Mo Kio	2025-07-04 16:18:22	https://id.jobstreet.com/id/job/...	We Are Looking For ; As the Network Engineer, you will be reporting to the Senior Operations Manager. You will be responsible for the	$3,500 – $5,000 per month (SGD)	Full time	singapore
7	85757987	Software Engineer / Senior Software Engineer (C++ / C#) - DSC/JS/17542	ST Engineering	Ang Mo Kio	2025-07-15 16:18:22	https://id.jobstreet.com/id/job/...	About ST Engineering ; ST Engineering is a global technology, defence, and engineering group with offices across Asia, Europe, the	null	Full time	singapore
8	85707238	Senior/ Cloud Infra Engineer - D...	ST Engineering Cloud & Data Ce...	Ang Mo Kio	2025-07-14 16:18:22	https://id.jobstreet.com/id/job/...	About ST Engineering; ST Engineering is a global technology, defence, and engineering group with offices across Asia, Europe, the	null	Full time	singapore
9	84620205	Google Cloud Engineer - DSC/ MC	ST Engineering Cloud & Data Ce...	Ang Mo Kio	2025-06-27 16:18:22	https://id.jobstreet.com/id/job/...	About ST Engineering; ST Engineering is a global technology, defence, and engineering group with offices	null	Full time	singapore

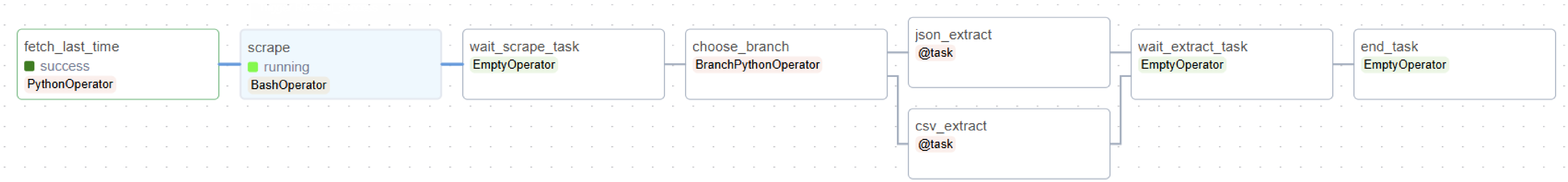
Use Branch Operator CSV or JSON and combine all keywords data

Data Understanding

DAG for Scrape & Ingestion



```
6a7a9203adf4
*** Found logs served from host http://6a7a9203adf4:8793/log/dag_id=scrape_ingest_dag/run_id=manual__2025-07-19T14:15:01.920736+00:00/task_id=csv_extract/attempt=1.log
[2025-07-19, 21:34:48 WIB] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: scrape_ingest_dag.csv_extract manual__2025-07-19T14:15:01.920736+00:00 [queued]>
[2025-07-19, 21:34:48 WIB] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: scrape_ingest_dag.csv_extract manual__2025-07-19T14:15:01.920736+00:00 [queued]>
[2025-07-19, 21:34:48 WIB] {taskinstance.py:2170} INFO - Starting attempt 1 of 1
[2025-07-19, 21:34:48 WIB] {taskinstance.py:2191} INFO - Executing <Task(_PythonDecoratedOperator): csv_extract> on 2025-07-19 14:15:01.920736+00:00
[2025-07-19, 21:34:48 WIB] {standard_task_runner.py:60} INFO - Started process 99931 to run task
[2025-07-19, 21:34:48 WIB] {standard_task_runner.py:87} INFO - Running: ['***', 'tasks', 'run', 'scrape_ingest_dag', 'csv_extract', 'manual__2025-07-19T14:15:01.920736+00:00', '--job-id', '474', '--raw', '--subdir', 'DAGS_FOLDER/scrape_ingest_dag.py', '--cfg-path', '/tmp/tmpfijb7w49']
[2025-07-19, 21:34:48 WIB] {standard_task_runner.py:88} INFO - Job 474: Subtask csv_extract
[2025-07-19, 21:34:48 WIB] {task_command.py:423} INFO - Running <TaskInstance: scrape_ingest_dag.csv_extract manual__2025-07-19T14:15:01.920736+00:00 [running]> on host 6a7a9203adf4
[2025-07-19, 21:34:49 WIB] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='***' AIRFLOW_CTX_DAG_ID='scrape_ingest_dag' AIRFLOW_CTX_TASK_ID='csv_extract' AIRFLOW_CTX_EXECUTION_DATE='2025-07-19T14:15:01.920736+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIRFLOW_CTX_DAG_RUN_ID='manual__2025-07-19T14:15:01.920736+00:00'
[2025-07-19, 21:34:49 WIB] {logging_mixin.py:188} INFO - Removed 11 duplicate rows
[2025-07-19, 21:34:49 WIB] {python.py:201} INFO - Done. Returned value was: jobstreetscrape_indonesia extracted successfully with 4 records
[2025-07-19, 21:34:49 WIB] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=scrape_ingest_dag, task_id=csv_extract, execution_date=20250719T141501, start_date=20250719T143448, end_date=20250719T143449
[2025-07-19, 21:34:49 WIB] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-07-19, 21:34:49 WIB] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check
```



fetch_last_time
success
PythonOperator

scrape
running
BashOperator

wait_scrape_task
EmptyOperator

choose_branch
BranchPythonOperator

json_extract
@task

csv_extract
@task

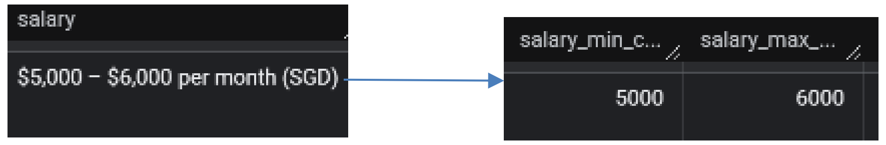
wait_extract_task
EmptyOperator

end_task
EmptyOperator

Transformation & consideration

Transformation (Silver Layer)

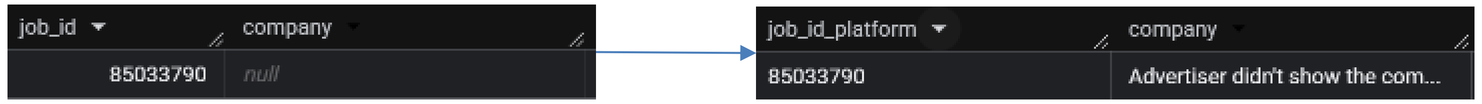
1. Separate salary to be absolut number, to be salary_min_cleaned and salary_max_cleaned .



salary
$5,000 – $6,000 per month (SGD)

salary_min_c...	salary_max_...
5000	6000

2. Fill column for blank value of company.

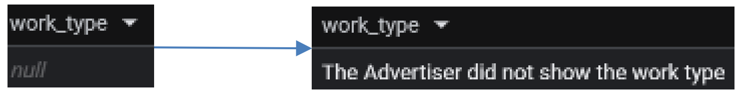


job_id	company
85033790	null

job_id_platform	company
85033790	Advertiser didn't show the com...

4. Parsed and cleaned text (title, job_desc, company, location).

4. Fill the work_type with value NULL



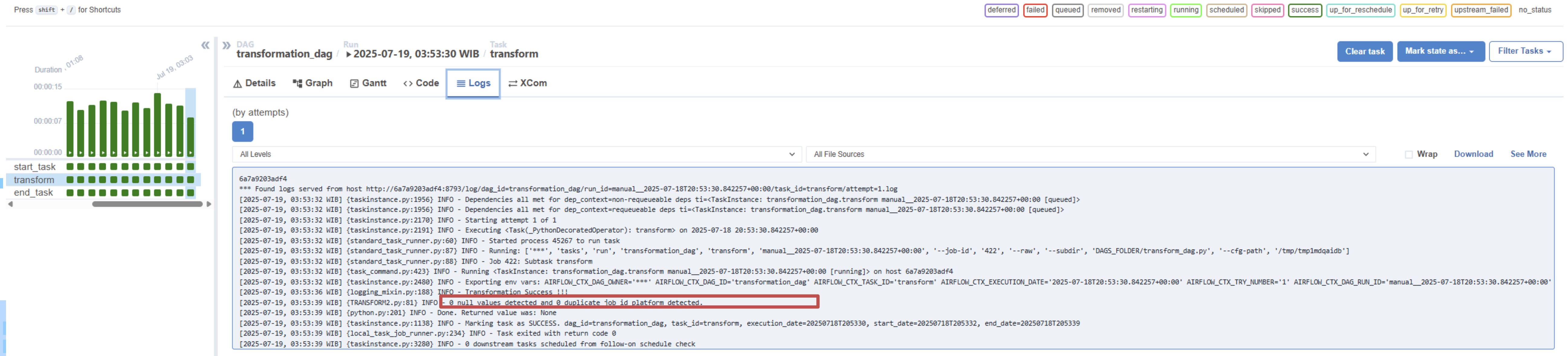
work_type
null

work_type
The Advertiser did not show the work type

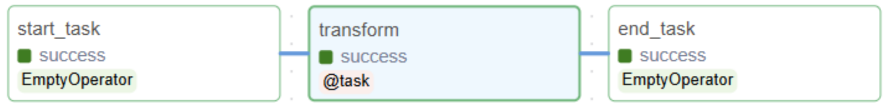
5. Drop duplicate regarding to job_id_platform.

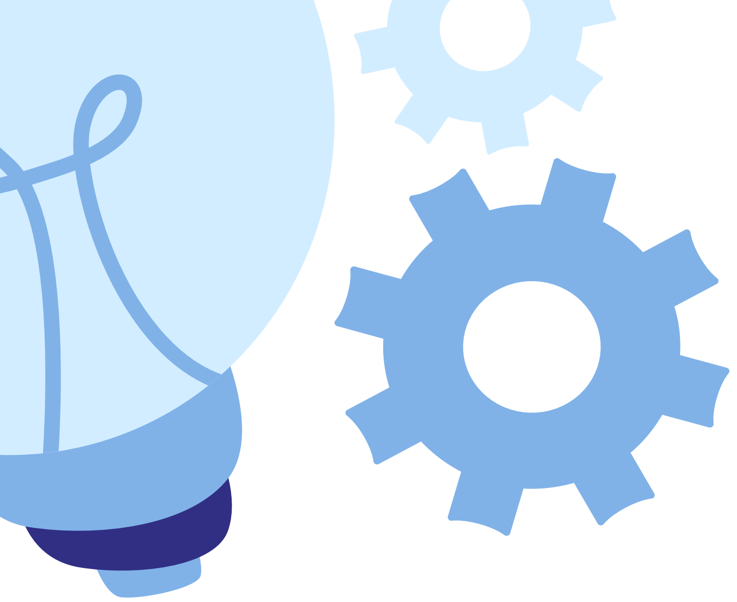
6. Null posted time filled by '1900-01-01 00:00:00'

7. Data validation function to ensure the table has no duplicate and null value regarding to job_id_platform.



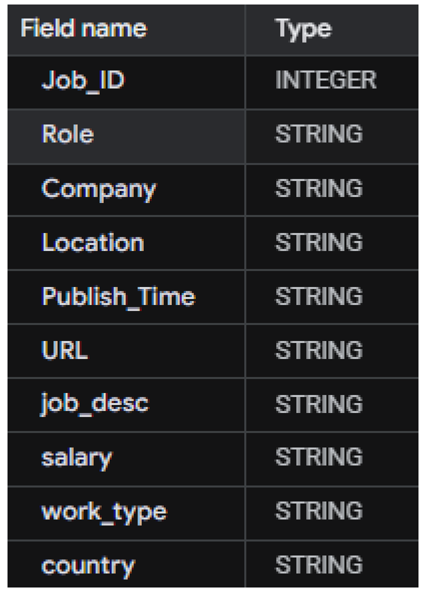
```
6a7a9203adf4
*** Found logs served from host http://6a7a9203adf4:8793/log/dag_id=transformation_dag/run_id=manual__2025-07-18T20:53:30.842257+00:00/task_id=transform/attempt=1.log
[2025-07-19, 03:53:32 WIB] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: transformation_dag.transform manual__2025-07-18T20:53:30.842257+00:00 [queued]>
[2025-07-19, 03:53:32 WIB] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: transformation_dag.transform manual__2025-07-18T20:53:30.842257+00:00 [queued]>
[2025-07-19, 03:53:32 WIB] {taskinstance.py:2170} INFO - Starting attempt 1 of 1
[2025-07-19, 03:53:32 WIB] {taskinstance.py:2191} INFO - Executing <Task(_PythonDecoratedOperator): transform> on 2025-07-18 20:53:30.842257+00:00
[2025-07-19, 03:53:32 WIB] {standard_task_runner.py:60} INFO - Started process 45267 to run task
[2025-07-19, 03:53:32 WIB] {standard_task_runner.py:87} INFO - Running: ['***', 'tasks', 'run', 'transformation_dag', 'transform', 'manual__2025-07-18T20:53:30.842257+00:00', '--job-id', '422', '--raw', '--subdir', 'DAGS_FOLDER/transform_dag.py', '--cfg-path', '/tmp/tmp1mdqaidb']
[2025-07-19, 03:53:32 WIB] {standard_task_runner.py:88} INFO - Job 422: Subtask transform
[2025-07-19, 03:53:32 WIB] {task_command.py:423} INFO - Running <TaskInstance: transformation_dag.transform manual__2025-07-18T20:53:30.842257+00:00 [running]> on host 6a7a9203adf4
[2025-07-19, 03:53:32 WIB] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='***' AIRFLOW_CTX_DAG_ID='transformation_dag' AIRFLOW_CTX_TASK_ID='transform' AIRFLOW_CTX_EXECUTION_DATE='2025-07-18T20:53:30.842257+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIRFLOW_CTX_DAG_RUN_ID='manual__2025-07-18T20:53:30.842257+00:00'
[2025-07-19, 03:53:36 WIB] {logging_mixin.py:188} INFO - Transformation Success !!!
[2025-07-19, 03:53:39 WIB] {TRANSFORM2.py:81} INFO - 0 null values detected and 0 duplicate job id platform detected.
[2025-07-19, 03:53:39 WIB] {python.py:201} INFO - Done. Returned value was: None
[2025-07-19, 03:53:39 WIB] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=transformation_dag, task_id=transform, execution_date=20250718T205330, start_date=20250718T205332, end_date=20250718T205339
[2025-07-19, 03:53:39 WIB] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-07-19, 03:53:39 WIB] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check
```





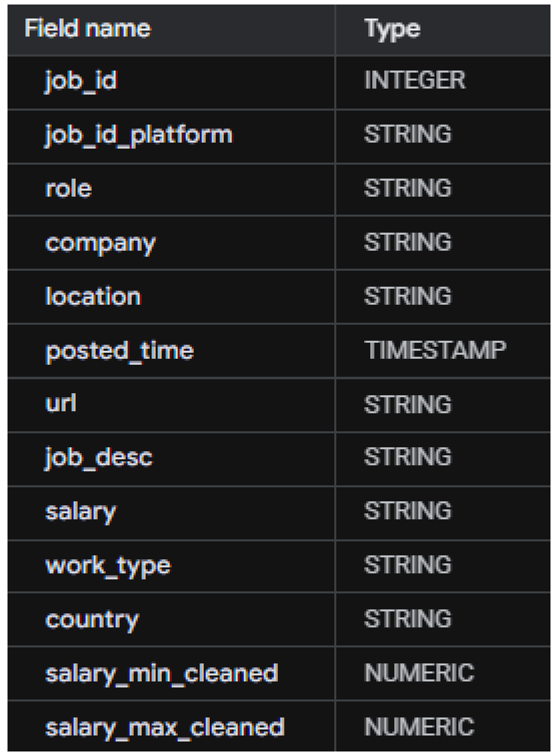
Data Modelling

Bronze Schema



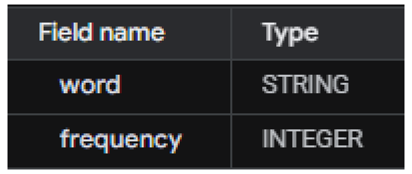
Field name	Type
Job_ID	INTEGER
Role	STRING
Company	STRING
Location	STRING
Publish_Time	STRING
URL	STRING
job_desc	STRING
salary	STRING
work_type	STRING
country	STRING

Silver Schema



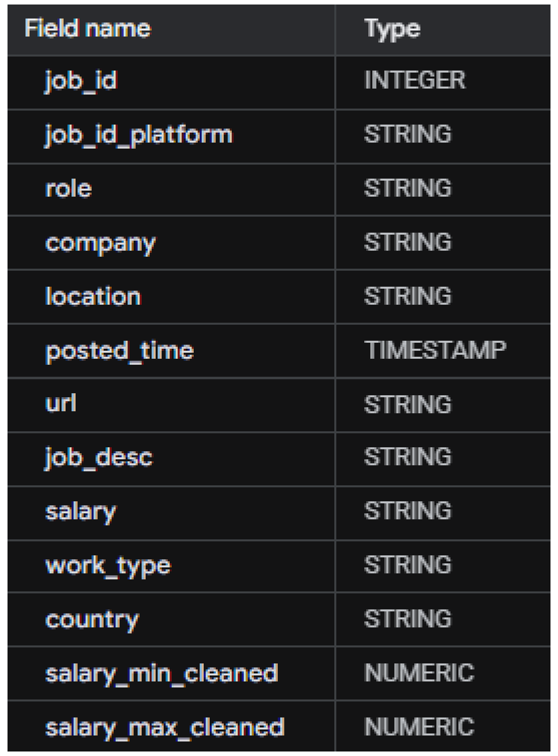
Field name	Type
job_id	INTEGER
job_id_platform	STRING
role	STRING
company	STRING
location	STRING
posted_time	TIMESTAMP
url	STRING
job_desc	STRING
salary	STRING
work_type	STRING
country	STRING
salary_min_cleaned	NUMERIC
salary_max_cleaned	NUMERIC

Gold Schema



Field name	Type
word	STRING
frequency	INTEGER

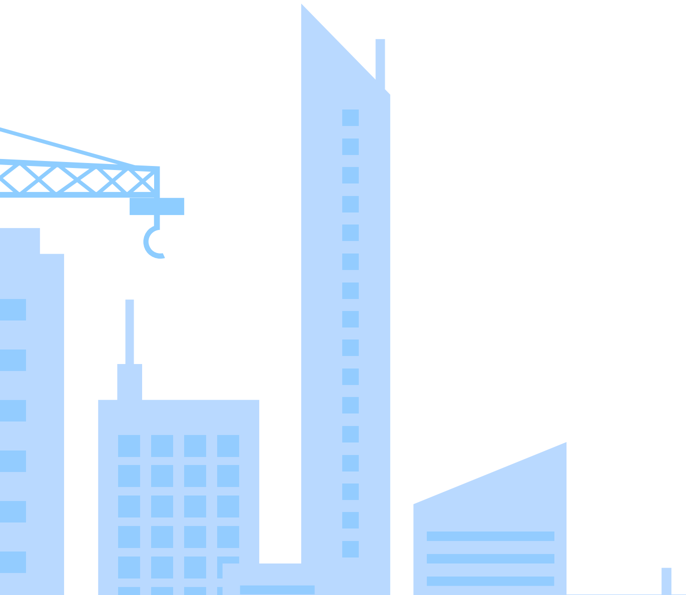
word_frequency table utilized for identify skill and tools in demand

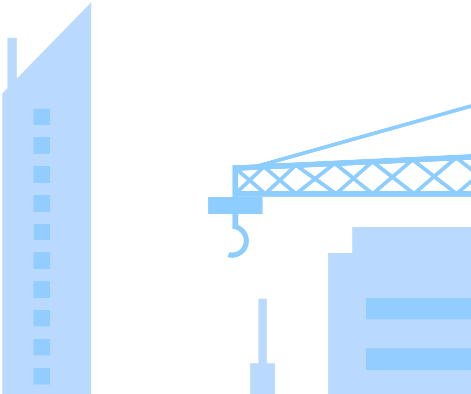


Field name	Type
job_id	INTEGER
job_id_platform	STRING
role	STRING
company	STRING
location	STRING
posted_time	TIMESTAMP
url	STRING
job_desc	STRING
salary	STRING
work_type	STRING
country	STRING
salary_min_cleaned	NUMERIC
salary_max_cleaned	NUMERIC

jobstreetscrape_singapore_detailed table utilized for :

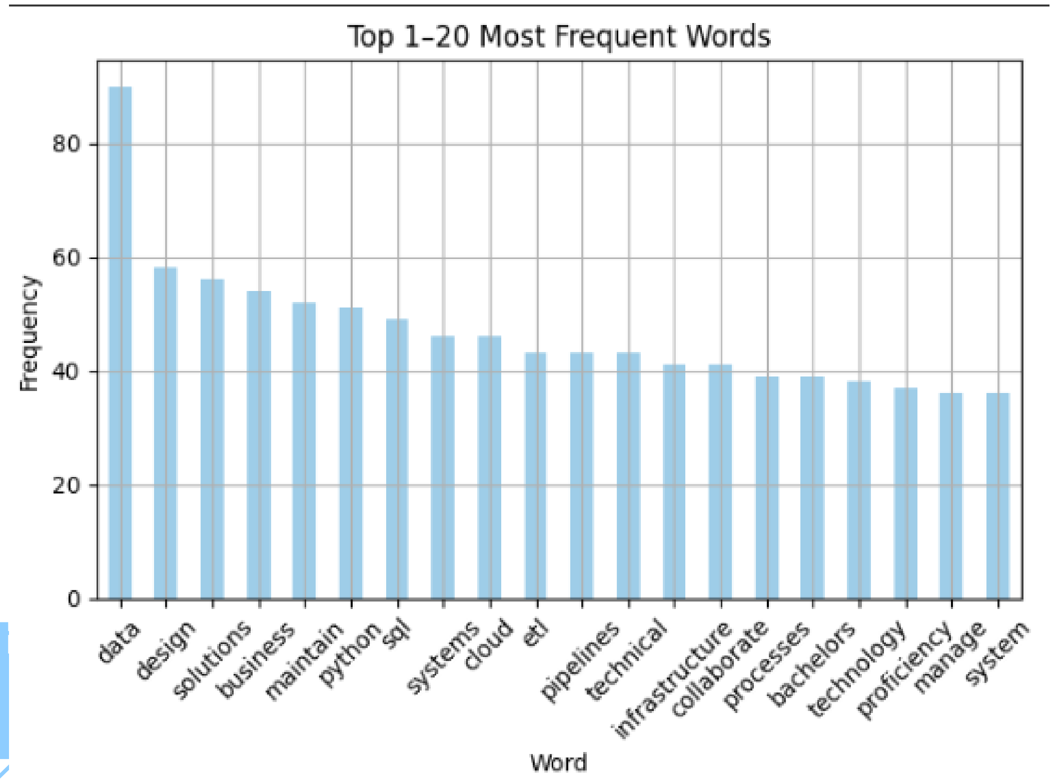
1. Daily dashboard such as count job posting each day, count location, etc
2. Job updated, thus the work searcher could prepare more for the job



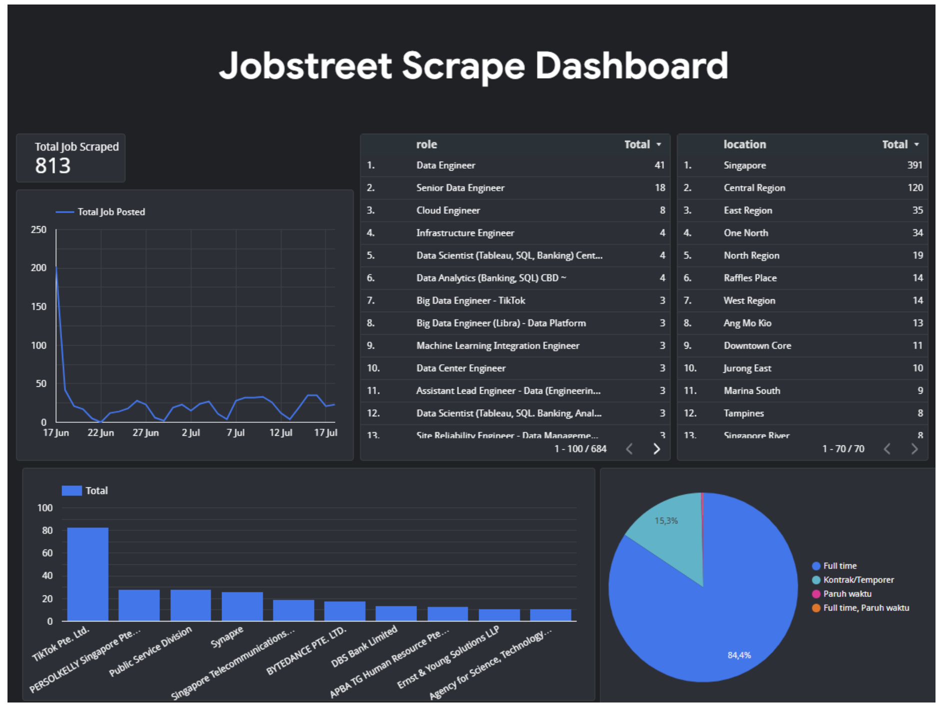


Dashboard

1. Word Frequency Count



2. Dashboard


Jobstreet Scrape Dashboard

Total Job Scraped
813

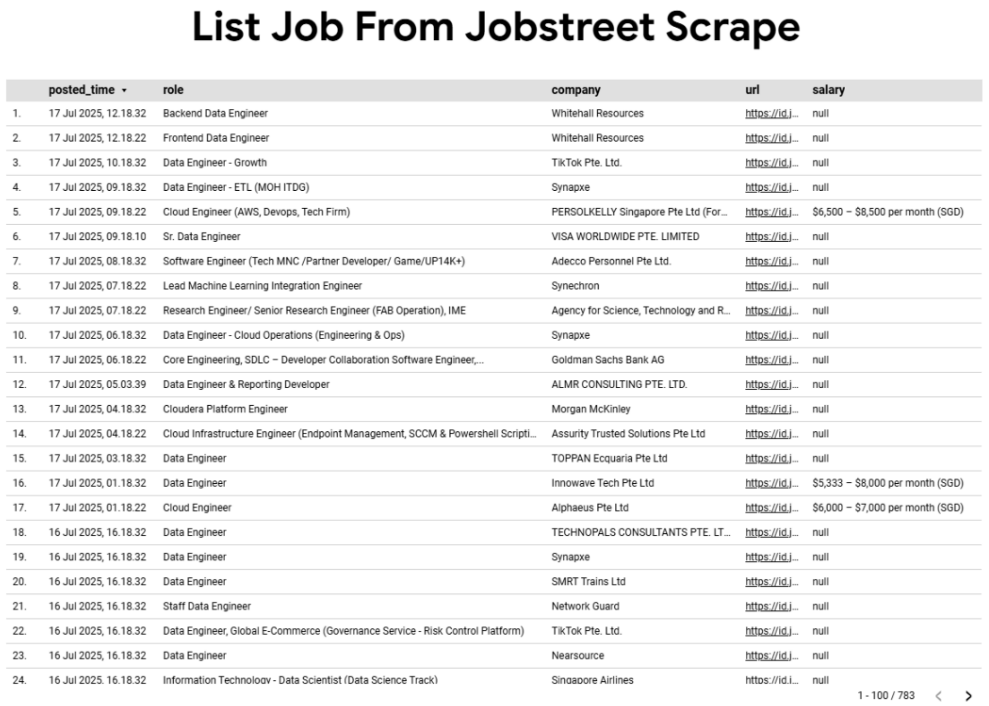
	role	Total ▾
1.	Data Engineer	41
2.	Senior Data Engineer	18
3.	Cloud Engineer	8
4.	Infrastructure Engineer	4
5.	Data Scientist (Tableau, SQL, Banking) Cent...	4
6.	Data Analytics (Banking, SQL) CBD ~	4
7.	Big Data Engineer - TikTok	3
8.	Big Data Engineer (Libra) - Data Platform	3
9.	Machine Learning Integration Engineer	3
10.	Data Center Engineer	3
11.	Assistant Lead Engineer - Data (Engineerin...	3
12.	Data Scientist (Tableau, SQL. Banking, Anal...	3
13.	Site Reliability Engineer - Data Manageme...	3

1 - 100 / 684

	location	Total ▾
1.	Singapore	391
2.	Central Region	120
3.	East Region	35
4.	One North	34
5.	North Region	19
6.	Raffles Place	14
7.	West Region	14
8.	Ang Mo Kio	13
9.	Downtown Core	11
10.	Jurong East	10
11.	Marina South	9
12.	Tampines	8
13.	Singapore River	8

1 - 70 / 70

3. Job List


List Job From Jobstreet Scrape

	posted_time ▾	role	company	url	salary
1.	17 Jul 2025, 12.18.32	Backend Data Engineer	Whitehall Resources	https://id.j...	null
2.	17 Jul 2025, 12.18.22	Frontend Data Engineer	Whitehall Resources	https://id.j...	null
3.	17 Jul 2025, 10.18.32	Data Engineer - Growth	TikTok Pte. Ltd.	https://id.j...	null
4.	17 Jul 2025, 09.18.32	Data Engineer - ETL (MOH ITDG)	Synapxe	https://id.j...	null
5.	17 Jul 2025, 09.18.22	Cloud Engineer (AWS, Devops, Tech Firm)	PERSOLKELLY Singapore Pte Ltd (For...	https://id.j...	$6,500 – $8,500 per month (SGD)
6.	17 Jul 2025, 09.18.10	Sr. Data Engineer	VISA WORLDWIDE PTE. LIMITED	https://id.j...	null
7.	17 Jul 2025, 08.18.32	Software Engineer (Tech MNC /Partner Developer/ Game/UP14K+)	Adecco Personnel Pte Ltd.	https://id.j...	null
8.	17 Jul 2025, 07.18.22	Lead Machine Learning Integration Engineer	Synechron	https://id.j...	null
9.	17 Jul 2025, 07.18.22	Research Engineer/ Senior Research Engineer (FAB Operation), IME	Agency for Science, Technology and R...	https://id.j...	null
10.	17 Jul 2025, 06.18.32	Data Engineer - Cloud Operations (Engineering & Ops)	Synapxe	https://id.j...	null
11.	17 Jul 2025, 06.18.22	Core Engineering, SDLC – Developer Collaboration Software Engineer,...	Goldman Sachs Bank AG	https://id.j...	null
12.	17 Jul 2025, 05.03.39	Data Engineer & Reporting Developer	ALMR CONSULTING PTE. LTD.	https://id.j...	null
13.	17 Jul 2025, 04.18.32	Cloudera Platform Engineer	Morgan McKinley	https://id.j...	null
14.	17 Jul 2025, 04.18.22	Cloud Infrastructure Engineer (Endpoint Management, SCCM & Powershell Scripti...	Assurity Trusted Solutions Pte Ltd	https://id.j...	null
15.	17 Jul 2025, 03.18.32	Data Engineer	TOPPAN Ecquaria Pte Ltd	https://id.j...	null
16.	17 Jul 2025, 01.18.32	Data Engineer	Innowave Tech Pte Ltd	https://id.j...	$5,333 – $8,000 per month (SGD)
17.	17 Jul 2025, 01.18.22	Cloud Engineer	Alphaeus Pte Ltd	https://id.j...	$6,000 – $7,000 per month (SGD)
18.	16 Jul 2025, 16.18.32	Data Engineer	TECHNOPALS CONSULTANTS PTE. LT...	https://id.j...	null
19.	16 Jul 2025, 16.18.32	Data Engineer	Synapxe	https://id.j...	null
20.	16 Jul 2025, 16.18.32	Data Engineer	SMRT Trains Ltd	https://id.j...	null
21.	16 Jul 2025, 16.18.32	Staff Data Engineer	Network Guard	https://id.j...	null
22.	16 Jul 2025, 16.18.32	Data Engineer, Global E-Commerce (Governance Service - Risk Control Platform)	TikTok Pte. Ltd.	https://id.j...	null
23.	16 Jul 2025, 16.18.32	Data Engineer	Nearsource	https://id.j...	null
24.	16 Jul 2025, 16.18.32	Information Technology - Data Scientist (Data Science Track)	Singapore Airlines	https://id.j...	null

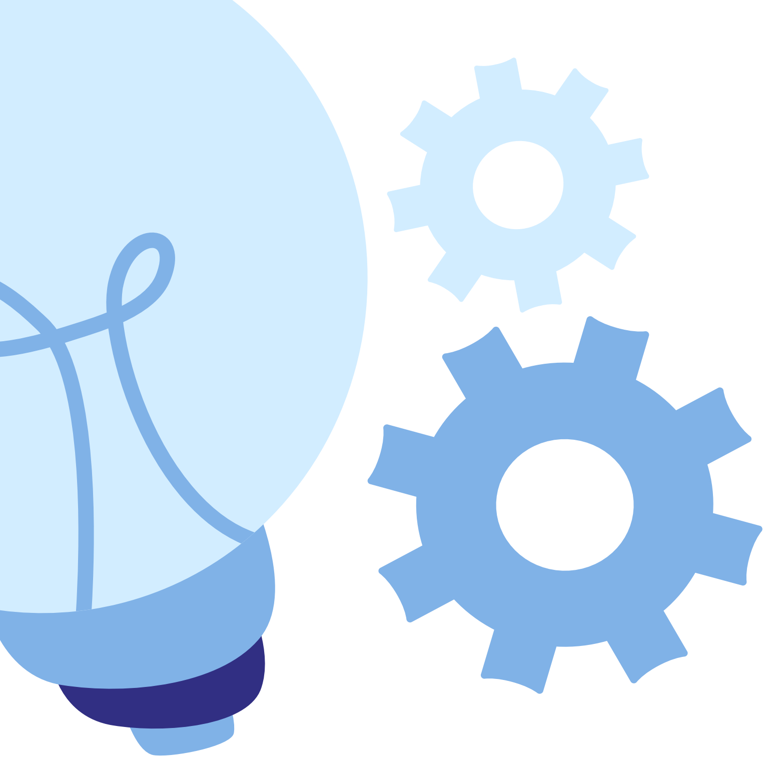
1 - 100 / 783

Conclusion

This project could fulfill the background problem to enhance the public sector, private sector, or third sector to have curriculum aligned to the market demand, moreover the dashboard result or the data could be utilize by job seeker to prepare of their best for every job posted.

The showed pipeline run successfully which consist:

1. Built a scalable, modular ELT pipeline using Airflow with main DAG triggering child DAGs.
2. Enabled dynamic configuration and support for full & incremental loads via parameters.
3. Performed data transformation, validation, and structured into bronze, silver, and gold layers.
4. Implemented multithreaded web scraping from JobStreet for efficient data extraction.

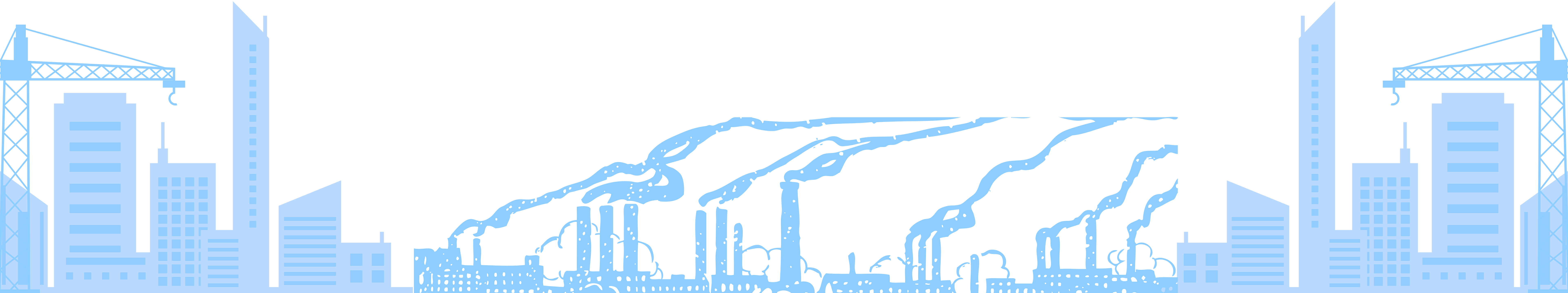


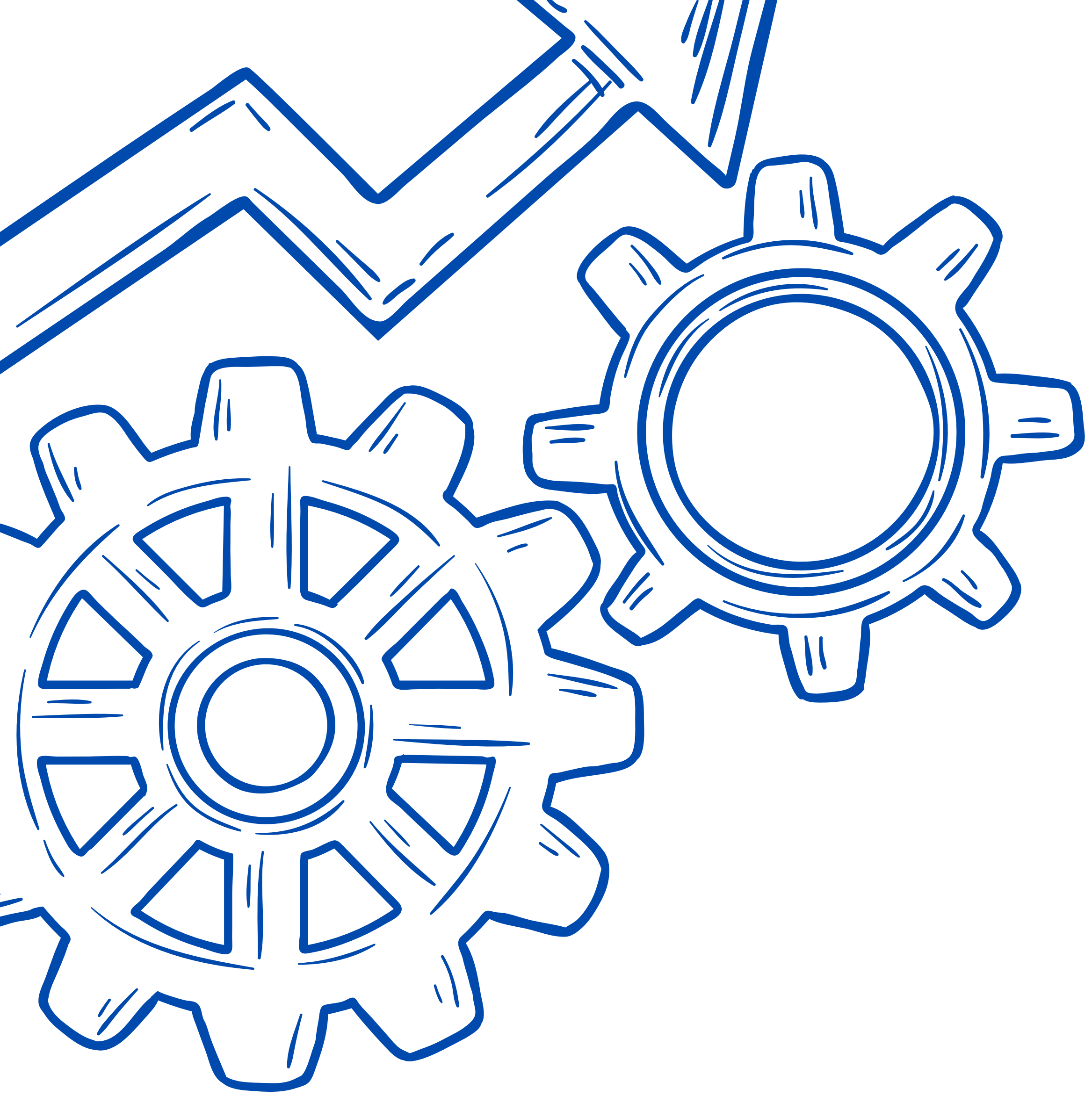
Recommendation

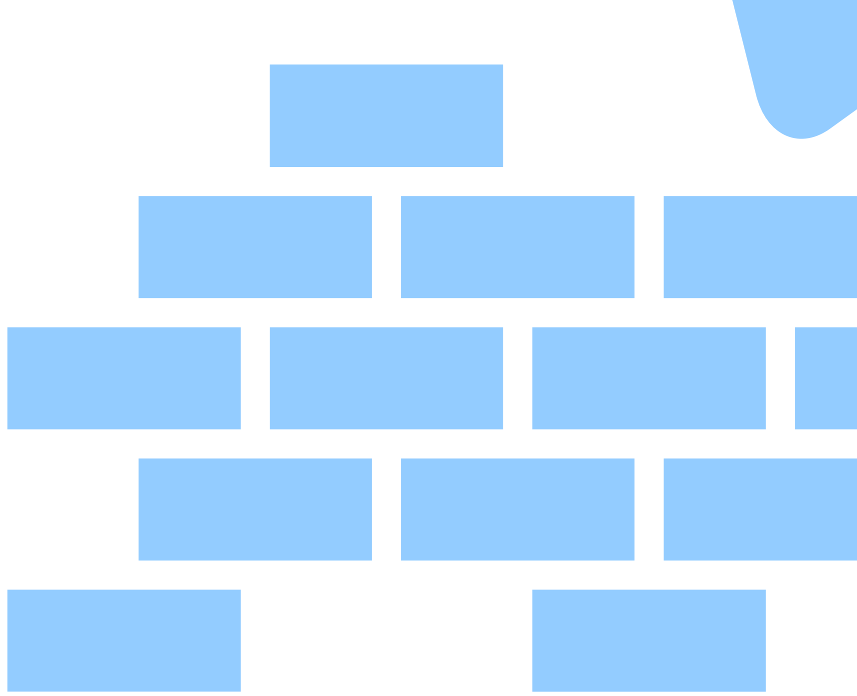
The future improvement for this project could be add more the scrape data source not only jobstreet, but could be glints, dealls, etc.

For error handling, could append retry at the DAG or task also notification if the dag is successful run or not.

At report section, could append forecast job based on posted time, moreover at the word frequency could be added a word filter more.







Thank You
谢谢
Terima Kasih
감사합니다

https://github.com/msidqyd

https://www.linkedin.com/in/muhammad-sidqy-dhiaulhaq-3b2000153/