Data Warehouse & Business Intelligence Project Report.

Collab Notebook Link: **DWH.ipynb**

Assignment Requirements

- Connect to at least two academic data sources (e.g., Semantic Scholar and Google Scholar).
- Write Python functions to fetch research papers using APIs.
- Design a data warehouse schema inside **Supabase** containing tables such as `papers`, `authors`, `paper_authors`, and `ingest_log`.
- Handle errors, API limits, and data quality issues.
- Implement deduplication logic so that the same paper appearing in both sources is stored only once.
- Log each ingestion with topic, counts, and timestamp.
- Submit a well-documented Jupyter Notebook and Report.

Proposed Solution and Architecture

The implemented solution was a modular, step-by-step ETL pipeline built in **Python (Jupyter Notebook)**. It follows the three classic stages of data warehousing:The overall flow is visualized as:

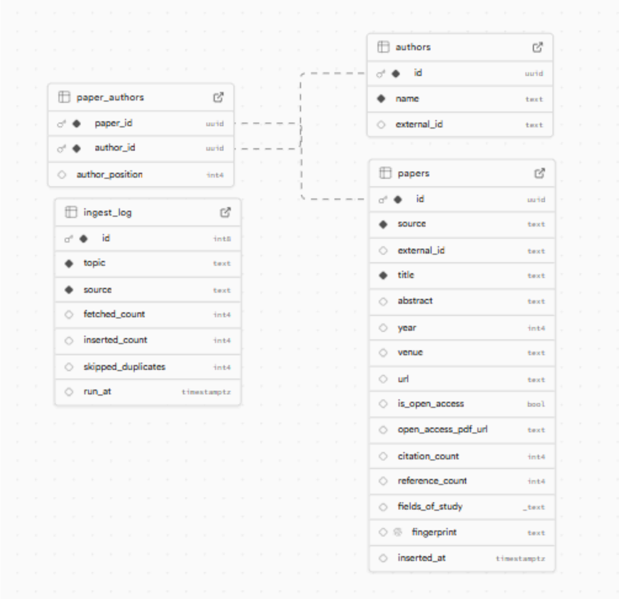
```
Semantic Scholar + Google Scholar
             ↓
       Fetch & Parse JSON
             ↓
  Normalize (title, authors, year)
             ↓
     Generate Fingerprints
             ↓
         Deduplicate
             ↓
        Insert → Supabase
             ↓
         Log Summary
```

Detailed Step-by-Step Implementation

Step 1 — Connecting to Supabase

A Supabase project was created and configured with the URL and Anon Key.
Tables were defined using SQL DDL:

The Python client was tested with a simple `.select("*")` query to confirm connection.



paper_authors	
paper_id	uuid
author_id	uuid
author_position	int4

ingest_log	
id	int8
topic	text
source	text
fetched_count	int4
inserted_count	int4
skipped_duplicates	int4
run_at	timestamptz

authors	
id	uuid
name	text
external_id	text

papers	
id	uuid
source	text
external_id	text
title	text
abstract	text
year	int4
venue	text
url	text
is_open_access	bool
open_access_pdf_url	text
citation_count	int4
reference_count	int4
fields_of_study	_text
fingerprint	text
inserted_at	timestamptz

Step 2 — Fetching from Semantic Scholar

A function `fetch_all_from_semantic_scholar(api_key, topic, max_pages, per_page)` used the Semantic Scholar REST API.

Step 3 — Fetching from Google Scholar via SerpApi

Because Google Scholar has no public API, the **SerpApi** REST endpoint was used.
The function `fetch_from_google_scholar_serpapi(topic, api_key, limit)` fetched 10 papers per page until the limit was reached.
It extracted titles, snippets, authors, links, and citation counts.
This ensured compliance with API rate limits and avoided blocking errors (`429 Too Many Requests`).

Step 4 — Data Merging

Instead of inserting data separately (as in earlier versions), both lists were concatenated:

```
all_papers = semantic_papers + google_papers
```

This produced one combined dataset representing all research results for the given topic.

Step 5 — Data Cleaning and Normalization

- Converted all titles to lowercase.
- Removed punctuation and extra spaces.
- Replaced missing values with empty strings.
- Ensured consistent author list format (e.g., `['A. Smith', 'B. Jones']`).

Step 6 — Fingerprint Deduplication

A custom function `make_fingerprint()` was defined:

```
def make_fingerprint(title, first_author, year):
```

By combining normalized title, first author, and year, each record became uniquely identifiable.
The helper `deduplicate_papers()` removed any repeated fingerprints, ensuring one unique record per paper across both sources.

Step 7 — Loading into Supabase

The function `populate_all_tables(topic, cleaned_papers, batch_size=50)` performed batch insertions:

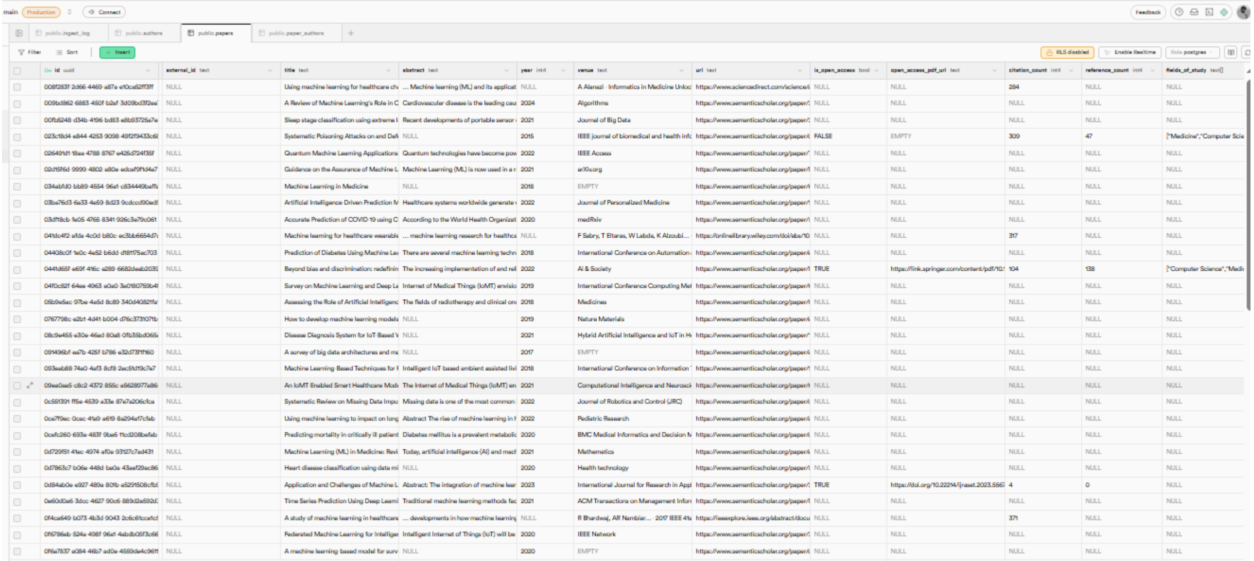
1. Inserted into `papers`.

2. Inserted unique authors into `authors`.

3. Created links in `paper_authors`.

4. Logged counts in `ingest_log`.

5. All database operations used Supabase's Python client for transactional safety.



id uuid	external_id text	title text	abstract text	year int4	venue text	url text	is_open_access bool	open_access_pdf_url text	citation_count int4	reference_count int4	fields_of_study text[]
008f283f 2d66 4469 a87a e10ca52ff3ff	NULL	Using machine learning for healthcare ch	... Machine learning (ML) and its applicat	NULL	A Alanazi · Informatics in Medicine Unloc	https://www.sciencedirect.com/science/	NULL	NULL	284	NULL	NULL
009bd862 6883 450f b2af 3d09bd3f2ea	NULL	A Review of Machine Learning's Role in C	Cardiovascular disease is the leading cau	2024	Algorithms	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
00fb6248 d34b 4196 bd83 e8b93725a7e	NULL	Sleep stage classification using extreme l	Recent developments of portable sensor	2021	Journal of Big Data	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
023c18d4 e844 4253 9098 49f2f9433c6	NULL	Systematic Poisoning Attacks on and Def	NULL	2015	IEEE journal of biomedical and health info	https://www.semanticscholar.org/paper/	FALSE	EMPTY	309	47	["Medicine","Computer Scie
026491d1 18ae 4788 8767 e425d724f35f	NULL	Quantum Machine Learning Applications	Quantum technologies have become pow	2022	IEEE Access	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
02d15f6d 9999 4802 a80e edcef9f1d4a7	NULL	Guidance on the Assurance of Machine L	Machine Learning (ML) is now used in a r	2021	arXiv.org	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
034abfd0 bb89 4554 96a1 c834449baffa	NULL	Machine Learning in Medicine	NULL	2018	EMPTY	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
03ba76d3 6a33 4e59 8d23 9cdccd90ed	NULL	Artificial Intelligence Driven Prediction M	Healthcare systems worldwide generate	2022	Journal of Personalized Medicine	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
03df18cb fe05 4765 8341 926c3e79c061	NULL	Accurate Prediction of COVID 19 using C	According to the World Health Organizat	2020	medRxiv	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
041dc4f2 afda 4c0d b80c ec3bb6654d7	NULL	Machine learning for healthcare wearable	... machine learning research for healthca	NULL	F Sabry, T Eltaras, W Labda, K Alzoubi...	https://onlinelibrary.wiley.com/doi/abs/10	NULL	NULL	317	NULL	NULL
04408c0f 1e0c 4e52 b6dd d181175ac703	NULL	Prediction of Diabetes Using Machine Le	There are several machine learning techn	2018	International Conference on Automation	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0441d65f e69f 416c a289 6682deab203	NULL	Beyond bias and discrimination: redefinin	The increasing implementation of and rel	2022	AI & Society	https://www.semanticscholar.org/paper/	TRUE	https://link.springer.com/content/pdf/10.	104	138	["Computer Science","Medi
04f0c82f 64ee 4963 a0a0 3e0180759b4	NULL	Survey on Machine Learning and Deep Le	Internet of Medical Things (IoMT) envisio	2019	International Conference Computing Met	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
05b9e5ac 97be 4e5d 8c89 340d40821fe	NULL	Assessing the Role of Artificial Intelligenc	The fields of radiotherapy and clinical onc	2018	Medicines	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0767798c e2b1 4d41 b004 d76c3731071b	NULL	How to develop machine learning models	NULL	2019	Nature Materials	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
08c9e455 e30e 46ad 80a8 0fb35bd065	NULL	Disease Diagnosis System for IoT Based W	NULL	2021	Hybrid Artificial Intelligence and IoT in He	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
091496bf ea7b 425f b786 e32d73f1f160	NULL	A survey of big data architectures and ma	NULL	2017	EMPTY	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
093eeb88 74e0 4af3 8cf8 2ec51d19c7e7	NULL	Machine Learning Based Techniques for F	Intelligent IoT based ambient assisted livi	2018	International Conference on Information	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
09ea0ea5 c8c2 4372 855c a5628977a86	NULL	An IoMT Enabled Smart Healthcare Mod	The Internet of Medical Things (IoMT) en	2021	Computational Intelligence and Neurosci	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0c561391 ff5e 4539 a33e 87e7a206cfca	NULL	Systematic Review on Missing Data Impu	Missing data is one of the most common	2022	Journal of Robotics and Control (JRC)	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0ce7f9ec 0cec 41a9 a619 8a294a17cfab	NULL	Using machine learning to impact on long	Abstract The rise of machine learning in h	2022	Pediatric Research	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0cefc260 693e 483f 9ba6 11cd208befeb	NULL	Predicting mortality in critically ill patient	Diabetes mellitus is a prevalent metabolic	2020	BMC Medical Informatics and Decision M	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0d729f51 41ec 4974 af0e 93127c7ad431	NULL	Machine Learning (ML) in Medicine: Revi	Today, artificial intelligence (AI) and mach	2021	Mathematics	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0d7863c7 b06e 448d ba0e 43eef29ec86	NULL	Heart disease classification using data mi	NULL	2020	Health technology	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0d84ab0e e927 489a 801b a5291508cfb	NULL	Application and Challenges of Machine L	Abstract: The integration of machine lear	2023	International Journal for Research in Appl	https://www.semanticscholar.org/paper/	TRUE	https://doi.org/10.22214/ijraset.2023.5567	4	0	NULL
0e60d0e6 3dcc 4627 90c6 889d2e592d	NULL	Time Series Prediction Using Deep Learni	Traditional machine learning methods fac	2021	ACM Transactions on Management Infor	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0f4ca649 b073 4b3d 9043 2c6c61cce1c	NULL	A study of machine learning in healthcare	... developments in how machine learning	NULL	R Bhardwaj, AR Nambiar... · 2017 IEEE 4th	https://ieeexplore.ieee.org/abstract/docu	NULL	NULL	371	NULL	NULL
0f6786eb 524e 498f 96a1 4abdb05f3c66	NULL	Federated Machine Learning for Intellige	Intelligent Internet of Things (IoT) will be	2020	IEEE Network	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL
0f6a7837 a084 46b7 ad0e 4559de4c961	NULL	A machine learning based model for surv	NULL	2020	EMPTY	https://www.semanticscholar.org/paper/	NULL	NULL	NULL	NULL	NULL