Siriraj Informatics and
Data Innovation Center

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
Thailand

OHDSI Thailand Meetup #1

OMOP ETL Pipeline by Siriraj

วันพฤหัสบดีที่ 26 มิถุนายน 2568
เวลา 12:30 – 13:30 น.

Join Zoom

https://bit.ly/omop-siriraj
Meeting ID: 991 2848 3445
Passcode: omop

Today's Agenda

26 June 2025 @ 12:30 – 13:30

12:30 – 12:40 ประชาสัมพันธ์ข่าวสาร OHDSI

12:40 – 13:10 OMOP ETL Pipeline โดย ณัฐวุฒิ อดุลยานุโกศล (Max)
รองหัวหน้าศูนย์นวัตกรรมข้อมูลศิริราช SiData+

13:10+ ถาม-ตอบ

มีการบันทึก Zoom

และจะอัพโหลดบน https://www.youtube.com/@OHDSIThailand

OHDSI Global Symposium 2025

Submission Deadline Tuesday 1 July 2025

- Oct. 7-9 at New Brunswick, New Jersey
- https://www.ohdsi.org/ohdsi2025/

- Past submissions:
  - https://www.ohdsi.org/2024-collaborator-showcase/
  - https://www.ohdsi.org/2023-collaborator-showcase/

OHDSI APAC

https://www.ohdsi.org/apac/

- **APAC Scientific Forums** are held monthly every first Thursday at 10 am Thailand time
  - July 3: Generative AI-based OMOP mapping with HDR UK's Lettuce tool (TBC)
- **APAC Community Calls** are held monthly every third Thursday at 10 am Thailand time
  - July 17: European Symposium Recap, RWD Studies by European Medicines Agency (EMA)
- **APAC Symposium 2025** December 6-7 in Shang Hai, China
  - Abstract submissions due September 7

OHDSI Thailand

Next meetup: Thu 24 July “Overview of OHDSI”

เรื่องที่ท่านสนใจ เรียงตามลำดับ สนใจมากที่สุดมาก่อน
18 of 18 answered

อันดับโดยเฉลี่ย

OMOP ETL Pipeline โดยศิริราช	3.7 <- อันดับ 1: 6 ท่าน
OMOP Mapping โดยศิริราช	4.7
OMOP Mapping โดยจุฬาฯ Tentative ครั้งที่ 3 ส.ค.	4.9
Research: Foundation Model โดยศิริราช	4.9
Vocab Mapping โดยศิริราช	5.2
Overview of OHDSI โดยศิริราช	5.7 <- อันดับ 1: 4 ท่าน

APAC Country Chapters

Activities | Community Support | Research | Funding

2025 ?

หากท่านสนใจเข้าร่วมเป็นผู้จัดกิจกรรมใน OHDSI Thailand Chapter กรุณาเลือกด้านที่ท่านถนัด
7 of 18 answered

OMOP ETL	57%	4 responses
Research	43%	3 responses
Vocabulary	29%	2 responses

natthawut.adu@mahidol.ac.th

Siriraj Informatics and
Data Innovation Center

OHDSI
Observational Health Data Sciences and Informatics

Mahidol University
Faculty of Medicine
Siriraj Hospital

OMOP ETL Pipeline

using on-premise SQLMesh

Mahidol University
Faculty of Medicine
Siriraj Hospital

Siriraj's Data and Challenges

pre-OMOP

EHR data over 20 years in a centralized data warehouse

- 2.5M total patients
- 49M clinic visits
- 1.9M admissions
- 140M diagnoses
- 235M lab
- 104M drug events

Can't let researchers access raw EHR data (privacy, security, messiness)

2º Data Usage
- Research
- Data Analytics
(mainly Tableau)

Less reproducible research, Less trust in results

Can't effectively share analytics scripts/ dashboards

Challenging network research studies

Takes time to prepare data (inclusion/exclusion criteria + formatting)

Takes time to analyze data

Most of the time researchers want quick feasibility assessment

Siriraj believes OMOP CDM can help

~~Can't let researchers access raw EHR data (privacy, security, messiness)~~

Structured data structure with proper de-identification

OMOP CDM

ATLAS helps cohort generation
i2b2 helps feasibility assessment
GenAI helps SQL code generation

2° Data Usage
- Research
- Data Analytics
- AI

Use publicly available codes:
Official HADES R libraries,
Community contribution, AI libs

Sharable analytics scripts/dashboards

Reproducible research, More trust in results

Easier network research studies

~~Takes time to prepare data (inclusion/exclusion criteria + formatting)~~

~~Takes time to analyze data~~

~~Most of the time researchers want quick feasibility assessment~~

OMOP Data Conversion

ETL = Extract Transform Load

Data experts and CDM experts together design the ETL

Data Analysis

People with medical knowledge create the code mappings

Vocabulary Mapping

Quality Assurance

All are involved in quality control

ETL Documentation

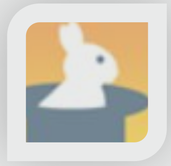
ETL

A technical person implements the ETL

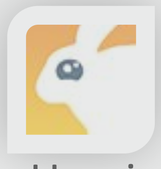
"เนื่องจากการแปลงข้อมูลมีความซับซ้อนเชิงเทคนิคสูง (ประเภท DB, logics, security, ฯลฯ) OHDSI จึงไม่มีเครื่องมือมาตรฐานมาให้ครับ"

OHDSI Tools

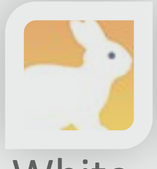
White Rabbit



Rabbit In a Hat



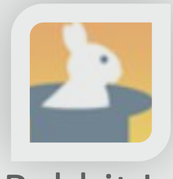
Usagi



White Rabbit

ACHILLES

DQD



Rabbit In a Hat

Siriraj ELT Pipeline

Extract - Load - Transform

Mahidol University
Faculty of Medicine
Siriraj Hospital

Data Lakehouse

DuckLake

Migrating

Data Warehouse

MS SQL SERVER

IPD

OPD

Lab

etc

MS SSIS (nightly)

Migrating

Dagster + Sling

SQLMesh

Transformation and Modeling Framework

Impact Analysis | Unit Tests | Transform | Audit | Blue/Green Deployment

https://sqlmesh.readthedocs.io/

SQLMesh as Seeds

R

python

+ableau

OMOP Vocab from Athena (.csv)

Vocab Mapping from Usagi (.csv)

Who are you?

1. Data/IT/Dev
2. Researcher
3. Clinician
4. Student
5. Other

Does your org have OMOP dataset?

1. Yes, on premise
2. Yes, on cloud
3. Developing
4. No
5. I don’t know

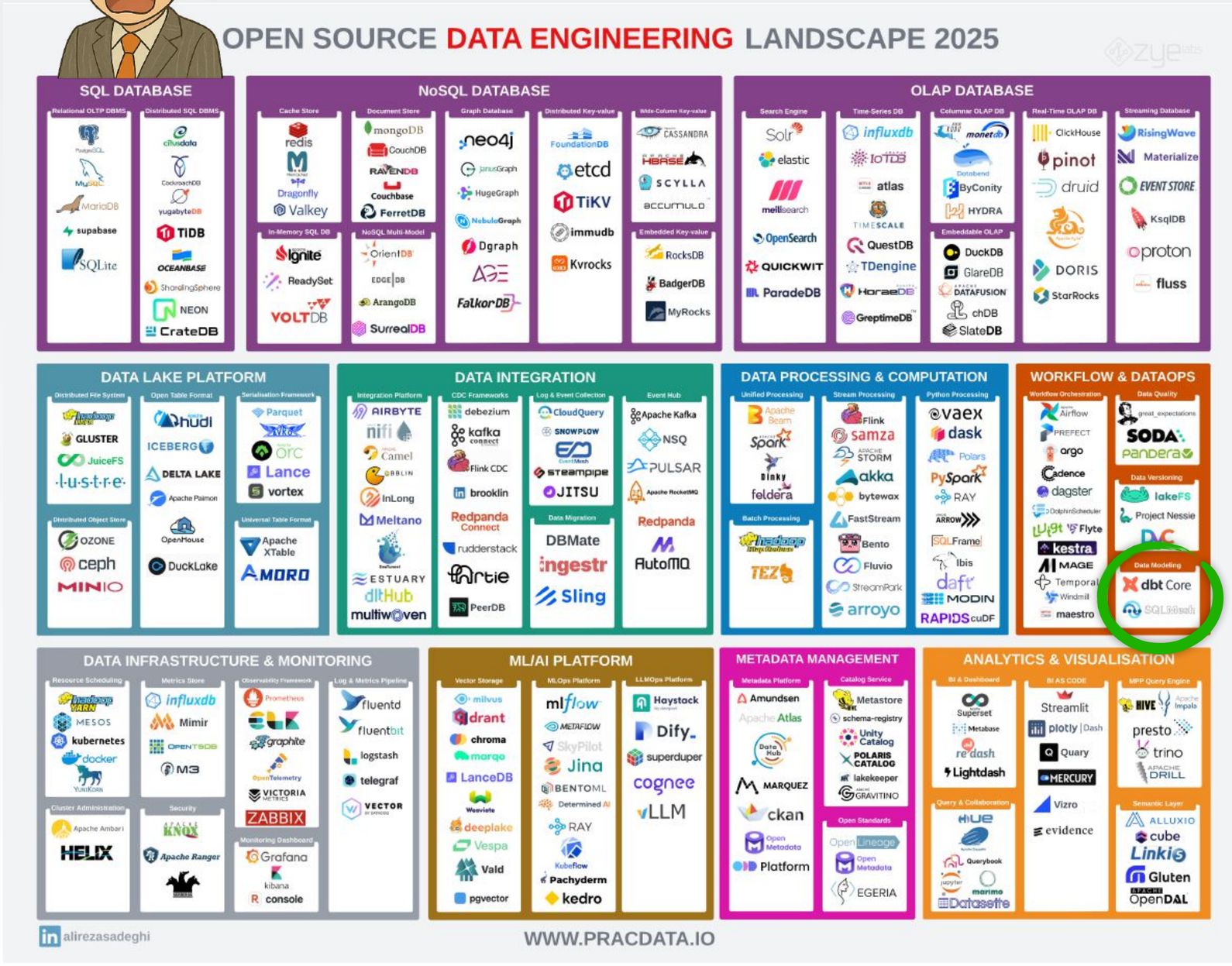
Data Tool Landscape

Mahidol University
Faculty of Medicine
Siriraj Hospital

“ไม่มีเครื่องมือที่ดีที่สุด
ต้องพิจารณาเลือก
ตามความเหมาะสม
และข้อจำกัดครับ”

Commercial

https://portable.io/learn/best-etl-tools



Siriraj: from dbt to SQLMesh

Mahidol University
Faculty of Medicine
Siriraj Hospital

Using dbt—a free and open-source software framework—to transform data into OMOP CDM in the ETL process

PRESENTER: Thanapat 'Thane' Pitchayarat
thanapat.pit@mahidol.edu

INTRO:

- The conversion of medical data into the OMOP CDM format requires a managed data engineering pipeline commonly referred to as the extract, transform, and load (ETL) process.
- The main transformation tasks in a typical OMOP CDM conversion include combining data from multiple sources, changing the original data models to match the OMOP CDM, retrieving the concept IDs of source values, and mapping the source concept IDs to the standard IDs.
- The complexity of the data transformation SQL scripts may grow rapidly beyond manageable. To keep the ETL pipeline maintainable, Siriraj Hospital uses dbt™ to transfrom its data to the OMOP CDM.
- dbt™ (shortened from data build tool) is a free and open-source software (FOSS) framework available at https://www.getdbt.com. It could be applied to data transformation at other institutions.

METHODS:

The data conversion pipeline at Siriraj Hospital can be summarized as:

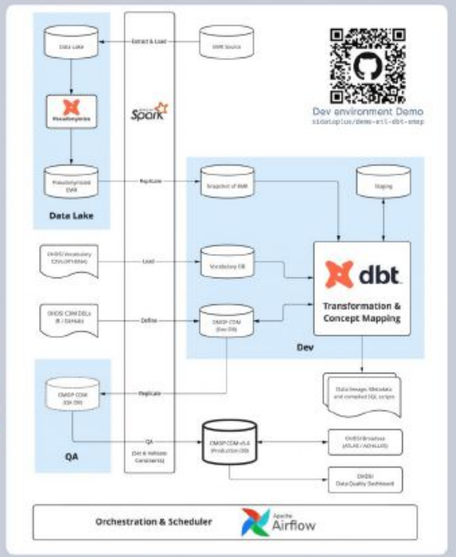
1. Extraction of data from hospital sources with Apache Spark
2. Load the data into data lake and Development environment with Apache Spark
3. Transform the data to match the OMOP CDM specifications with dbt
4. Load the OMOP CDM-ed data into QA and Production environments

Each step is containerized with Docker. All steps are ochestrated and scheduled by Apache Airflow. Codes are version controlled with GitHub.

dbt:

- dbt comes with a command-line interface with commands that compile SQL scripts and execute the code on the connected database engines, as well as a graphical user interface.
- The core library of dbt is a Python package that supplements traditional SQL scripts with Pythonic Jinja templating.
- With the Jinja templating,
  - any frequently used SQL command can be packaged as a modular macro that can take parameters similar to a Python function, and;
  - the Jinja tags enable data lineage tracking that can be visualized on an interactive web application generated by dbt command. The web application referred to as dbt documentation also presents metadata, such as table & field descriptions, data testing conditions, upstream and downstream tables. The metadata are partly generated automatically and can be added manually as YAML files.
- To verify data quality, dbt can run automated tests during transformation execution or on demand.
- Given the popularity of dbt in the enterprise analytics space, there are many tools that can be integrated with dbt, namely Airflow for data pipeline orchestration, GreatExpectations for data quality, and DataHub for data catalog.

"An organized approach to build a maintainable ETL pipeline for the OMOP CDM with minimal cost while keeping our data engineers sane 😍"



Simplified architectural diagram of the OMOP CDM conversion pipeline at Siriraj Hospital.

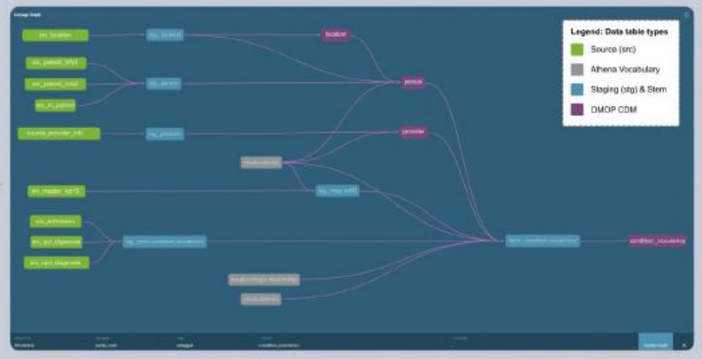


Table data lineage automatically generated by dbt. Each node represents a table or a view of data. Each linking edge represents a data flow from the source(s) to its destination(s), with data transformation in between. Each of the data transformation step is programmed as an SQL SELECT script.

Full paper

(a)

```sql
SELECT
  person_patient_id AS person_id,
  gender_concept.concept_id AS gender_concept_id,
  ...,
  race_concept.concept_id AS race_concept_id,
  ...,
  -- the rest of SELECT statement omitted for brevity
  -- please refer to OMOP CDM PERSON table for CDM fields
FROM {{ ref('stg__person') }} AS person
{{ map_concept(cdm_table='person', concept_code_field='gender_concept_code',
               vocabulary_id='gender') }}
{{ map_concept(cdm_table='person', concept_code_field='race_concept_code',
               vocabulary_id='race') }}
```

(b)

```sql
{%- macro map_concept(cdm_table="", concept_code_field="", vocabulary_id="") -%}

LEFT JOIN {{ source('vocab', 'concept') }} AS {{vocabulary_id}}_concept
ON {{cdm_table}}.{{concept_code_field}} = {{vocabulary_id}}_concept.concept_code
  AND {{vocabulary_id}}_concept.vocabulary_id = '{{vocabulary_id}}'
  AND {{vocabulary_id}}_concept.standard_concept = 'S'

{%- endmacro -%}
```

Simplified SQL snippets (a) to create the CDM PERSON table with data from a staging table joined with the vocabulary concept tables via macros (b) to set a macro template for concept mapping. These SQL snippets with Jinja tags are to be compiled and submitted to the database engine by dbt.

CONCLUSION:

- dbt is a promising free and open-source software framework that massively facilitates the data conversion process into OMOP CDM.
- dbt programmatically manages the SQL transformation scripts in the ETL process, and consequently enhances the maintainability of the data pipeline.
- Data engineers with proficiency in SQL and Python could learn dbt in a few days and probably take a few weeks to implement dbt in the pipeline.

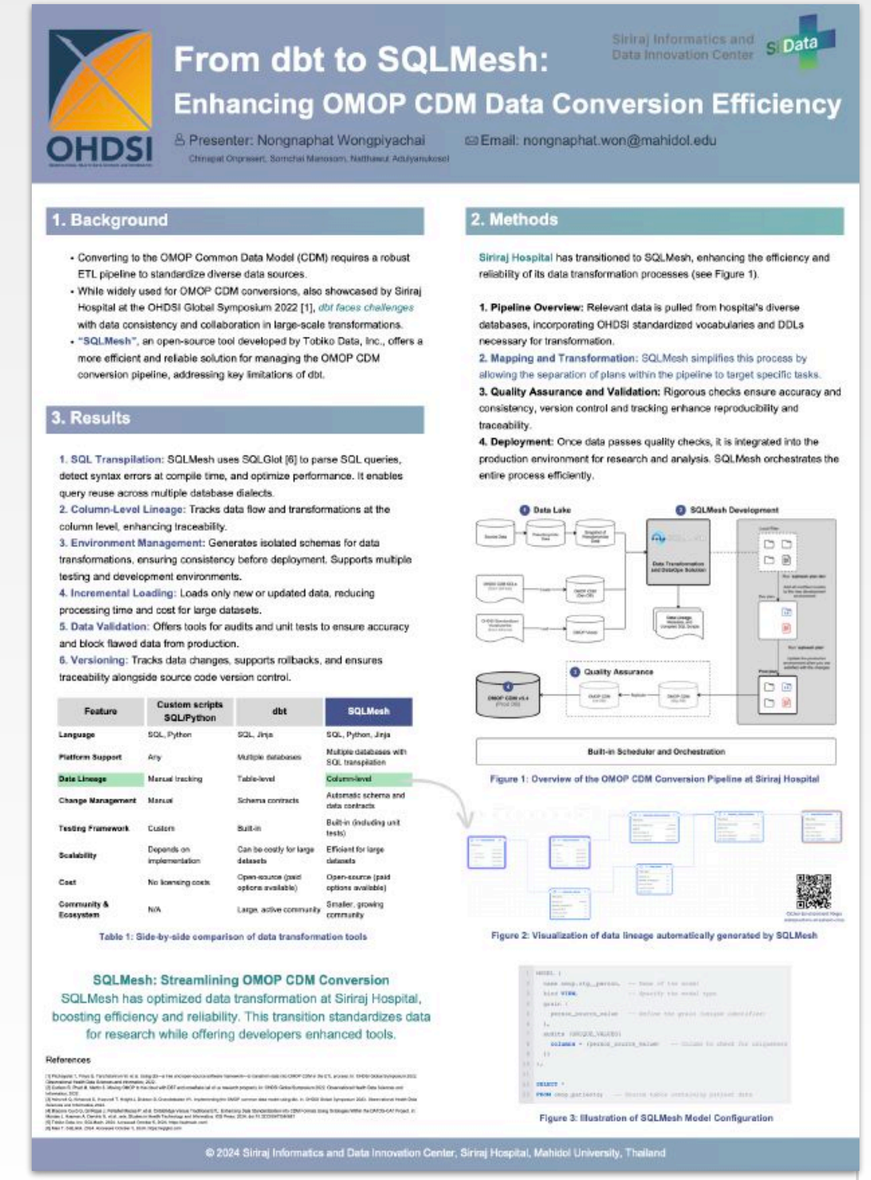
REFERENCES:

1. dbt Labs, Inc.. dbt-core [Internet]. 2022. Available from: https://github.com/dbt-labs/dbt-core
2. OHDSI. WhiteRabbit [Internet]. 2022. Available from: https://github.com/OHDSI/WhiteRabbit
3. OHDSI. Rabbit-in-a-Hat [Internet]. 2022. Available from: http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html
4. OHDSI. Usagi [Internet]. 2022. Available from: https://github.com/OHDSI/Usagi
5. The Pallets Projects. Jinja [Internet]. 2022. Available from: https://palletsprojects.com/p/jinja/
6. Superconductive Health, Inc.. Welcome to great expectations [Internet]. 2022. Available from: https://greatexpectations.io/
7. Calogica. dbt_expectations [Internet]. 2022. Available from: https://hub.getdbt.com/calogica/dbt_expectations/0.1.2/
8. dbt Labs, Inc.. Success Stories [Internet]. 2022. Available from: https://www.getdbt.com/success-stories/
9. Apache Software Foundation. dbt Cloud Operators [Internet]. 2022. Available from: https://airflow.apache.org/docs/apache-airflow-providers-dbt-cloud/stable/operators.html
10. DataHub Project. dbt [Internet]. 2022. Available from: https://datahubproject.io/docs/generated/ingestion/sources/dbt/

Thanapat Pitchayarat, Gun Pinyo, Watcharaporn Tanchotsrinon, Somkid Khamsrimuang, Chalita Issarasittiphap, Chaiyanun Bootnumpech, Noppon Siangchin, Kanphitcha Promma, Nattachai Bovornmongkolsak, Prapat Suriyaphol, Natthawut Adulyanukosol
Siriraj Informatics and Data Innovation Center (SIData+), Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Mahidol University Faculty of Medicine Siriraj Hospital
Siriraj Informatics and Data Innovation Center
OHDSI

https://www.ohdsi.org/wp-content/uploads/2022/10/2-Pitchayarat-OHDSI2022Poster-Adulyanukosol-scaled.jpg



From dbt to SQLMesh:
Enhancing OMOP CDM Data Conversion Efficiency

OHDSI

Siriraj Informatics and Data Innovation Center SIData

Presenter: Nongnaphat Wongpiyachai
Email: nongnaphat.won@mahidol.edu
Chinapat Onpraseert, Somchai Manasorn, Natthawut Adulyanukosol

1. Background

- Converting to the OMOP Common Data Model (CDM) requires a robust ETL pipeline to standardize diverse data sources.
- While widely used for OMOP CDM conversions, also showcased by Siriraj Hospital at the OHDSI Global Symposium 2022 [1], *dbt faces challenges* with data consistency and collaboration in large-scale transformations.
- **"SQLMesh"**, an open-source tool developed by Tobiko Data, Inc., offers a more efficient and reliable solution for managing the OMOP CDM conversion pipeline, addressing key limitations of dbt.

2. Methods

Siriraj Hospital has transitioned to SQLMesh, enhancing the efficiency and reliability of its data transformation processes (see Figure 1).

1. **Pipeline Overview:** Relevant data is pulled from hospital's diverse databases, incorporating OHDSI standardized vocabularies and DDLs necessary for transformation.
2. **Mapping and Transformation:** SQLMesh simplifies this process by allowing the separation of plans within the pipeline to target specific tasks.
3. **Quality Assurance and Validation:** Rigorous checks ensure accuracy and consistency, version control and tracking enhance reproducibility and traceability.
4. **Deployment:** Once data passes quality checks, it is integrated into the production environment for research and analysis. SQLMesh orchestrates the entire process efficiently.

3. Results

1. **SQL Transpilation:** SQLMesh uses SQLGlot [6] to parse SQL queries, detect syntax errors at compile time, and optimize performance. It enables query reuse across multiple database dialects.
2. **Column-Level Lineage:** Tracks data flow and transformations at the column level, enhancing traceability.
3. **Environment Management:** Generates isolated schemas for data transformations, ensuring consistency before deployment. Supports multiple testing and development environments.
4. **Incremental Loading:** Loads only new or updated data, reducing processing time and cost for large datasets.
5. **Data Validation:** Offers tools for audits and unit tests to ensure accuracy and block flawed data from production.
6. **Versioning:** Tracks data changes, supports rollbacks, and ensures traceability alongside source code version control.

Feature	Custom scripts SQL/Python	dbt	SQLMesh
Language	SQL, Python	SQL, Jinja	SQL, Python, Jinja
Platform Support	Any	Multiple databases	Multiple databases with SQL transpilation
Data Lineage	Manual tracking	Table-level	Column-level
Change Management	Manual	Schema contracts	Automatic schema and data contracts
Testing Framework	Custom	Built-in	Built-in (including unit tests)
Scalability	Depends on implementation	Can be costly for large datasets	Efficient for large datasets
Cost	No licensing costs	Open-source (paid options available)	Open-source (paid options available)
Community & Ecosystem	N/A	Large, active community	Smaller, growing community

Table 1: Side-by-side comparison of data transformation tools

Figure 1: Overview of the OMOP CDM Conversion Pipeline at Siriraj Hospital

Figure 2: Visualization of data lineage automatically generated by SQLMesh

Figure 3: Illustration of SQLMesh Model Configuration

SQLMesh: Streamlining OMOP CDM Conversion

SQLMesh has optimized data transformation at Siriraj Hospital, boosting efficiency and reliability. This transition standardizes data for research while offering developers enhanced tools.

References

https://www.ohdsi.org/wp-content/uploads/2025/01/20_Nongnaphat-Wongpiyachai_From-dbt-to-SQLMesh.pdf

Data Modeling Tools

dbt & SQLMesh ทำด้านปั้นข้อมูลโดยเฉพาะ

Mahidol University
Faculty of Medicine
Siriraj Hospital

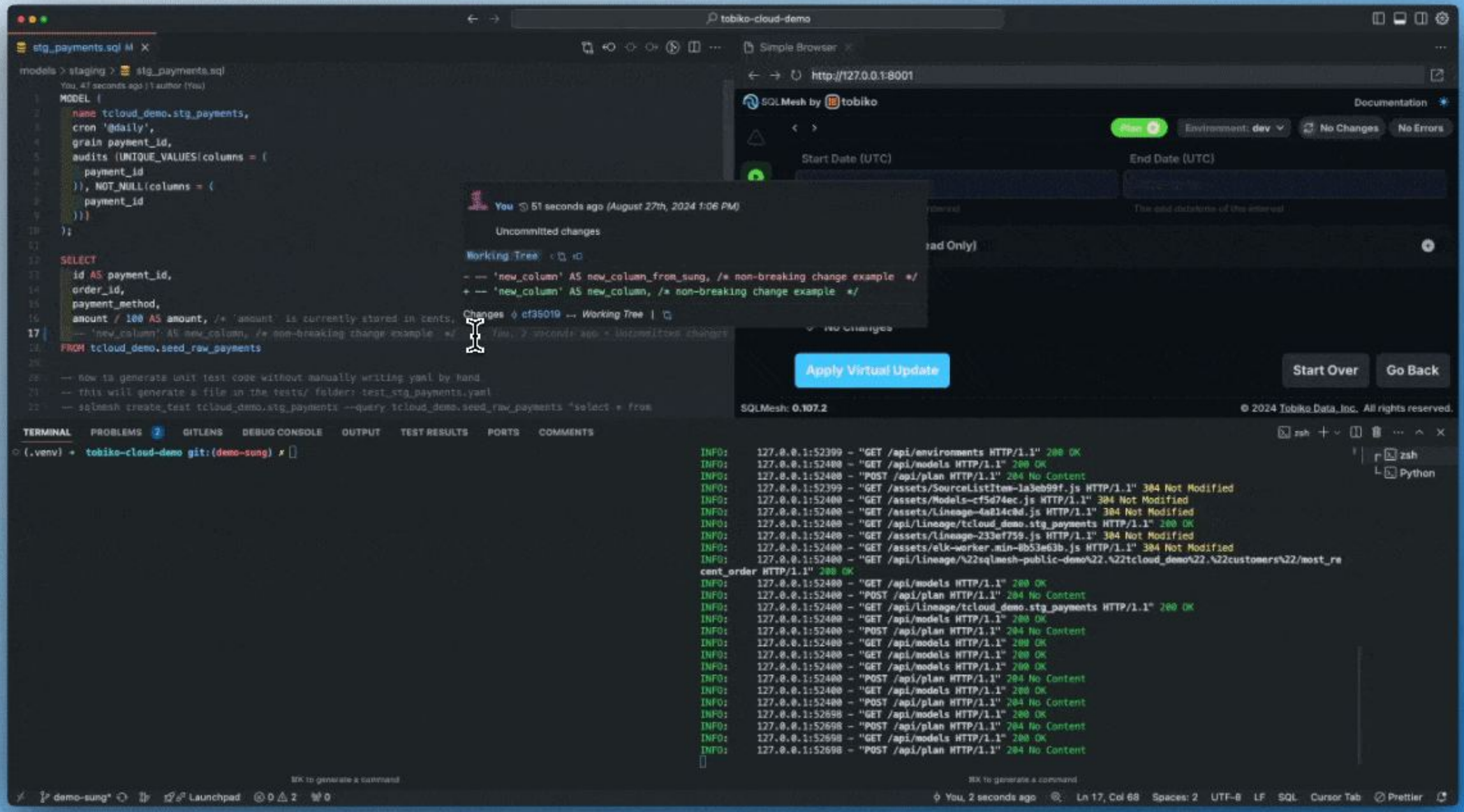
Tools	ความเหมาะสม	ข้อจำกัด
1. Custom scripts, e.g., SQL, Python, R, Java	ต้องการอะไรก็เขียนเอาเองได้เลย flexible ที่สุด	เหนื่อยทั้งเขียน ดูแล ซ่อม technical debt บาน
2. Commercial ETL tools, e.g., MS SSIS, Talend, Pentaho	ถ้ามีคนใช้เครื่องมือเหล่านี้เป็นอยู่แล้ว ในองค์กรก็ง่าย	ถ้าไม่มี ก็เสียตังค์ เสียเวลา ส่วนใหญ่เป็น GUI เยอะ maintain ยาก
3. Cloud ETL tools, e.g., Azure Data Factory, AWS Glue, GCP Cloud Dataflow, Dataform	ถ้าใช้ Cloud เจ้านั้น ๆ อยู่แล้ว ก็ integrate ได้ดี	ไม่ได้เกิดมาเพื่อปั้นโมเดลข้อมูลโดยเฉพาะ ขาด features ช่วยเหลือ
4. dbt	เป็น free open source library ตัวแรก ๆ ที่มี features data modeling ครบครัน	ไม่เข้าใจ SQL โดยตรง เป็นแค่ Jinja templating engine หลาย ๆ features ต้องเสียตังค์เพิ่ม
5. SQLMesh	พัฒนาแก้ไขข้อจำกัดของ dbt เข้าใจ SQL semantic และเปิดฟรีเกือบทุก features (slides 17)	ยังไม่ stable version (v0.196) มี learning curve (แต่ไม่ยากกว่า dbt)

แล้ว SQLMesh ใช้งานยังไง (แบบย่อ ๆ)

แบบเต็ม ๆ (1 ชั่วโมง): https://www.youtube.com/watch?v=J0_2hkXz-HY

https://github.com/sidataplus/demo-etl-sqlmesh-omop-synthea/

1. สร้าง Python environment แล้ว pip install sqlmesh (แต่ตอนแปลงข้อมูลใช้แค่ SQL ก็ได้)
2. ตั้งค่าการเชื่อมต่อกับ database ใน config.yaml
3. เขียน SQL ในการแปลงข้อมูลตามที่ต้องการแต่ละ table
4. เขียน Model Definition (metadata) ในไฟล์ .sql เดียวกัน บอกให้ SQLMesh จัดการข้อมูลอย่างไร เช่น incremental load, data type, audit
5. เขียน SQL table อื่นต่อไปเรื่อย ๆ แล้ว SQLMesh จะนำไปต่อกันให้เป็น DAG Lineage
6. สั่ง `sqlmesh plan dev` เพื่อให้ SQLMesh render SQL scripts แล้วรันบน DB ตาม DAG (ไม่มีข้อมูลเข้าเครื่อง dev) บน dev environment (schema omop__dev.[table_name]) ระหว่างที่รันสามารถให้เช็ค audit ตรวจสอบคุณภาพข้อมูลไปพร้อมกัน
7. เช็คข้อมูลใน dev env จนพอใจ แล้วรัน `sqlmesh plan` เพื่อให้ promote เข้า prod env (schema omop.[table_name])
8. ตั้ง schedule อัพเดทข้อมูลด้วย `sqlmesh run`



https://sqlmesh.readthedocs.io/

Features ช่วยการแปลงข้อมูล

ที่ tools อื่นไม่มี

1. **Stateful Virtual Data Environment** ช่วยแยก dev, prod ให้ โดยยังอยู่ใน DB เดียวกัน หรือจะแยก DB (multiple gateway) ก็ทำได้ เก็บ state history ทำให้ dev หลายคนทำงานร่วมกันได้อย่างปลอดภัย
2. **SQL Semantic Understanding** เข้าใจ SQL ช่วยให้เจอ bug ก่อนไปรันบน DB ช่วยให้ track change by column สร้าง column lineage ได้
3. **SQL Transpilation** ใช้ SQLGlot แปล syntax ข้าม DBMS ได้
4. **Data contracts** เช็คว่าการแปลงข้อมูลสมบูรณ์ทั้ง pipeline ผ่าน audits ถึงจะเปิดให้ access ได้ (dbt แปลงได้แค่ไหนก็ปล่อยเลย)

	dbt Core	SQLMesh
SQL Models	✓	✓
Python Models	✓	✓
Jinja support	✓	✓
Jinja macros	✓	✓
Python macros	✗	✓
Virtual Data Environments	✗	✓
Open-source CI/CD bot	✗	✓
Blue/Green Deployments	✗	✓
Multi-repository support	✗	✓
SQL transpilation	✗	✓
CLI	✓	✓
SQL semantic validation	✗	✓
Unit tests	✓	✓
Table diff	✗	✓
Data audits	✓	✓
Schema contracts	✓	✓
Data contracts	✗	✓
VS Code extension	✗	✓

https://www.tobikodata.com/sqlmesh

Demo Siriraj

อยากลองใช้ SQLMesh บ้าง ทำอย่างไร?

1. **SQLMesh Docs Get Started:** https://sqlmesh.readthedocs.io/en/stable/quick_start/
2. **Demo Synthea -> OMOP:** https://github.com/sidataplus/demo-etl-sqlmesh-omop-synthea/ ใช้ synthetic data บน DuckDB ทำตาม instructions ใน README ได้
3. เริ่มแปลงข้อมูลจาก table ง่าย ๆ / หลัก ๆ ก่อน

OMOP Conversion order (suggested)

1. local concept master
2. person
3. location
4. care_site
5. provider
6. visit_occurrence
7. observation_period
8. death
9. condition_occurrence
10. observation
11. procedure_occurrence
12. drug_exposure
13. condition_era
14. drug_era
15. measurement
16. cost
17. payer_plan_period