/Projek DOS

Workshop

Building Data Pipeline using Airflow & Doris

Code. Schedule. Automate

Apache Airflow

DORIS

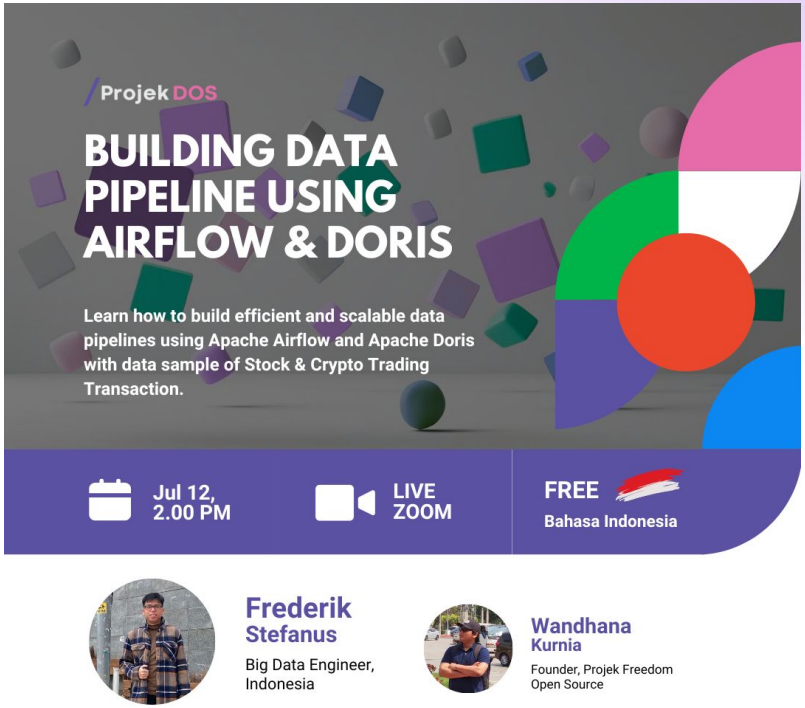


Table of contents

Projek DOS

Apache Airflow

DORIS

01
Projek Freedom Open Source

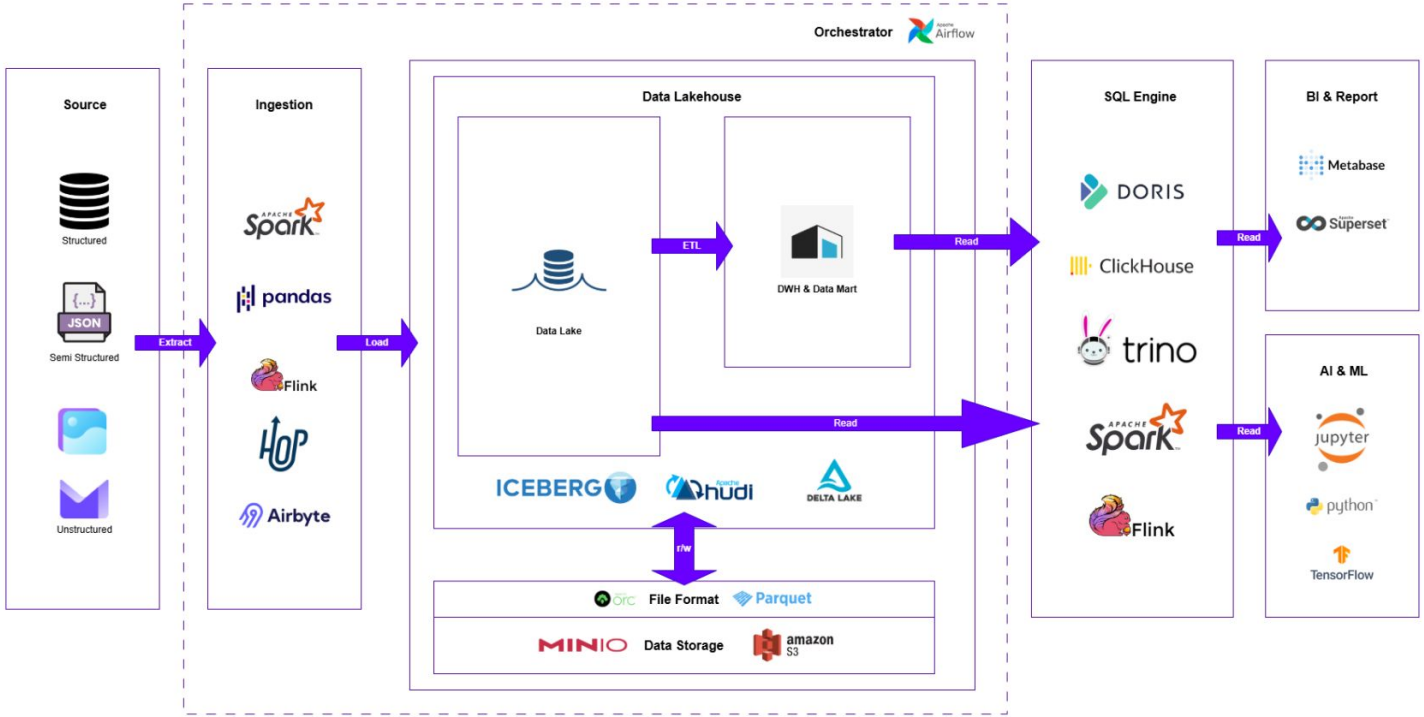
Intro /Projek DOS

Projek Freedom Open Source

Founded in 2025, as a **real world scenario** training environments initiative focused on open-source technology. The goal is to empower individuals with networking & knowledge about tools and technologies :

- Business Intelligence
- Data Warehouse
- Data Lake
- Big Data
- Machine Learning
- Artificial Intelligence
- More

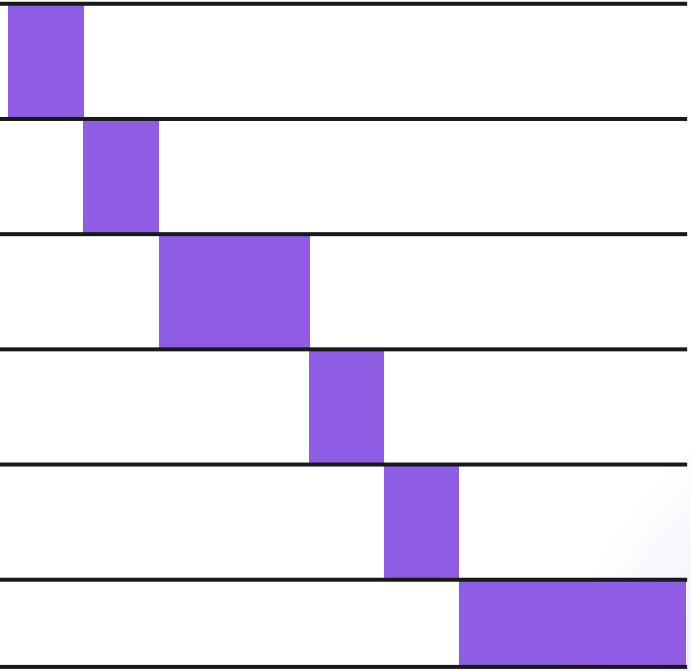
Real World Scenario Pipeline



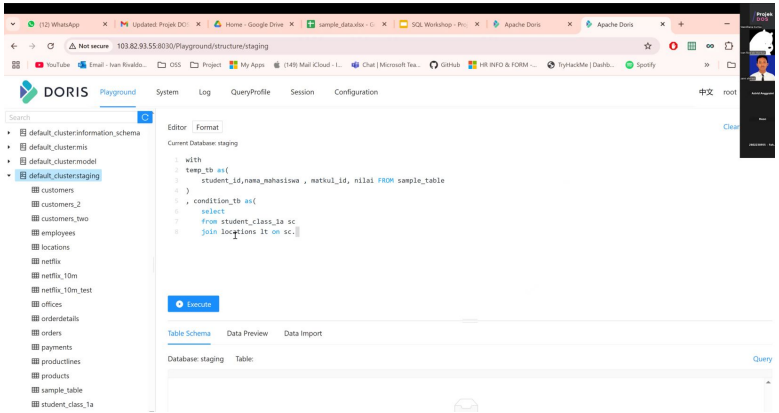
/ProjekDOS

Roadmap (2025)

Initiative	Objective	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
SQL	Mastering data querying, manipulation, and optimization for efficient database management									
ETL	Extracting, transforming, and loading data for seamless integration and processing									
Orchestrator	Managing and coordinating end-to-end data workflows automatically									
Business Intelligence	Leveraging analytics and visualization tools to drive data-driven decisions									
Machine Learning	Building predictive models and AI solutions for advanced data analytics.									
Other	Exploring more essential technologies									



Previous Workshop - SQL (26, April 2025)

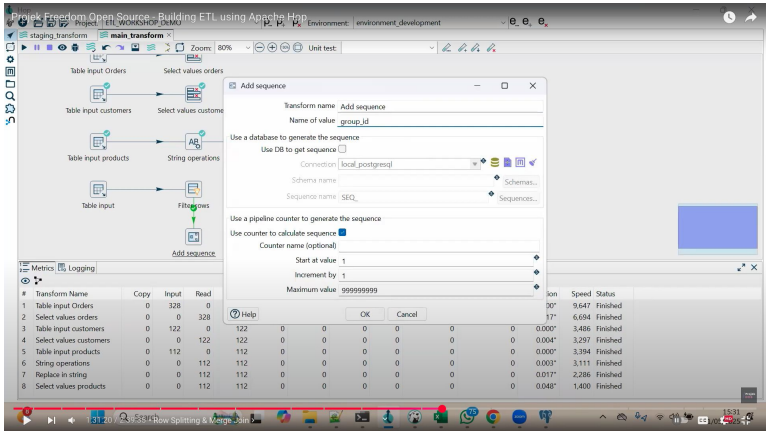


In this workshop, we explored key concepts of Query and hands-on SQL examples, including:

- String Manipulation
- Subquery
- CTE (Common Table Expression)
- Windowing Function

ProjekDOS

Previous Workshop - ETL (31, May 2025)



In this workshop, we explored key concepts of ETL and hands-on data transformation examples using Apache Hop, including:

- Extracting data from multiple sources
- Data Cleaning & Transformation
- Loading into target tables
- Run ETL workflows

Open Partnership & Collaborate

University

+ Hands-on training via workshops & labs.
+ Certification programs for students.
+ Support for building academic projects

Government

+ Training Program for digital and data literacy
+ Support for national/regional education initiatives.
+ Public sector upskilling & seminars.

Corporate

+ Custom training to upskill employees.
+ Practical sessions on trend technology stack
+ Role-based learning paths.

Community

+ Organizing meetups, hackathons, & open forums
+ Support for open-source contributions & learning
+ Knowledge sharing via local tech communities

Business Contact

info@projekdos.com

+6281385368844 (Whatsapp)

LinkedIn

/ProjekDOS

02
Introduction

About Me - Speaker

Frederik Stefanus

Big Data Engineer with 5+ years of experience in banking and telecom, focused on building end-to-end data pipelines. Skilled in leveraging tools like Apache Airflow, Spark, and SQL to deliver reliable, scalable data infrastructure.

Frederik Stefanus

/ProjekDOS

Fastest Growing Careers

In the next upcoming years, Job related to Data & Artificial Intelligence are the **Top Performer & Highest Demands** on Job Market.

(Source: World Economic Forum)

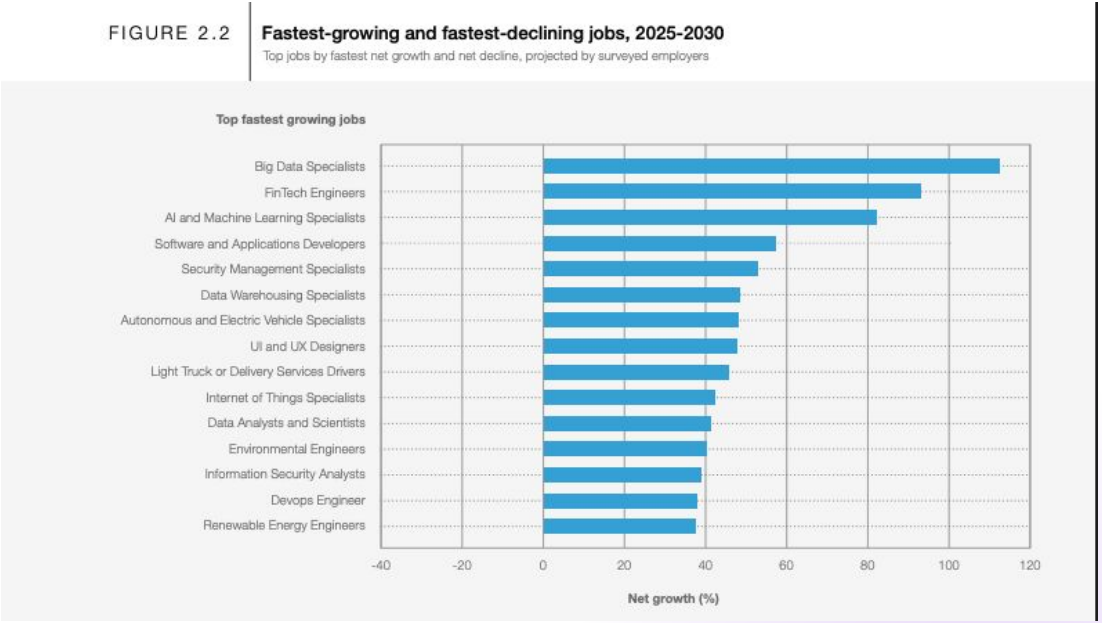


Image Credit: WEF

#1 Skills on Rise

Technological skills are projected to grow in importance more rapidly than any other type of skills. Among these, **AI and big data top the list as the fastest-growing skills**

(Source: World Economic Forum)

FIGURE 3.4 **Skills on the rise, 2025-2030**

Share of employers that consider skills to be increasing, decreasing, or remaining stable in importance. Skills are ranked based on net increase, which is the difference between the share of employers that consider a skill category to be increasing in use and those that consider it to be decreasing in use.

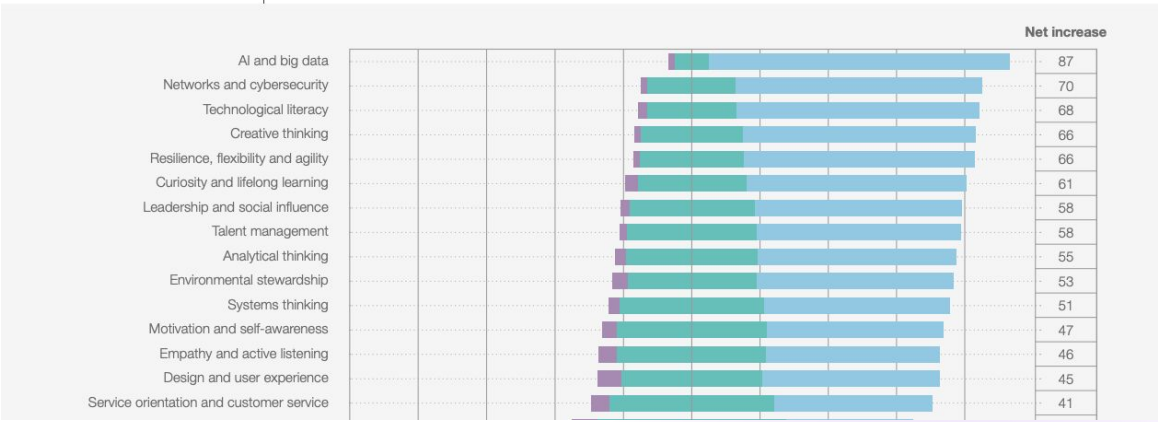


Image Credit: WEF

Salary Range 😀

With **5 years of hands-on experience** in the data field, IT Consultants in Indonesia are seeing competitive compensation. The average monthly salary typically falls between **IDR 17,000,000 to IDR 20,000,000 (even more)**, depending on the industry, experience, technical skill set, and company size.

(Source: persolkelly - Salary Guide 2024)

INFORMATION TECHNOLOGY				
Data Center Service Operation (DCSO) Specialist	D3	1	18,000,000	21,000,000
Data Center Service Provisioning (DCSP) Manager	S1	8	45,000,000	50,000,000
Data Center Techical Support (DCTS) Lead	D3/S1	5	35,000,000	40,000,000
Data Center Technical Support (DCTS) Manager	S1	8	45,000,000	50,000,000
Developer Community Manager	S1	3-5	33,000,000	36,000,000
Education Adoption Lead	S2	5-10	63,000,000	68,000,000
Enterprise Account Manager	S2	5-10	50,000,000	55,000,000
ESH ESG Senior Manager	S1	10	60,000,000	70,000,000
Field Data Collection Staff	High School	1-3	5,000,000	6,000,000
Finance Director	S1	15	85,000,000	95,000,000
Firmware Engineer	S1	1	10,000,000	15,000,000
Firmware Engineer Leader	S1	2	20,000,000	25,000,000
Fullstack Software Developer	S1	5	12.000.000	16.000.000
Education Adoption Specialist	S1	3-5	28,000,000	31,000,000
Head of End User Services	S1	>10	50,000,000	55,000,000
Head of HR	S1	10	30,000,000	35,000,000
HR Business Partner	S1	1-3	8,000,000	12,000,000
iOS Developer	S1	2-3	9,000,000	12,000,000
IT Consultant	S1	5	17,000,000	20,000,000
IT Developer Lead	S1	2-4	13,000,000	15,000,000
IT Manager	S1	10	45,000,000	50,000,000

Image Credit: persolkelly

03
Data Orchestration

Data Orchestration

Data orchestration refers to the process of scheduling & manage tasks.

A orchestration tool make business to schedule & focus on automation

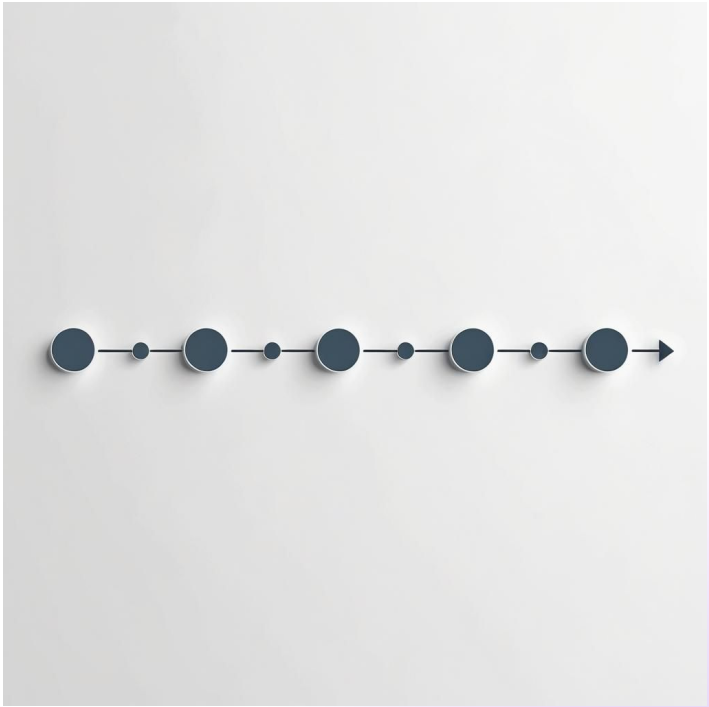


Image: Canva AI Prompt Generator

Popular Orchestrator Tools

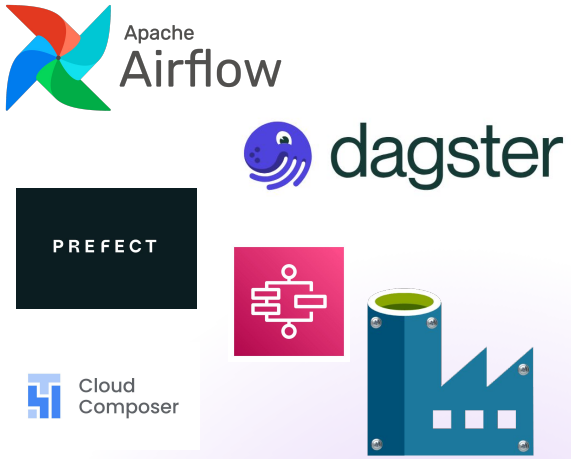
Most popular Open Source Orchestration tools in business :

1. **Apache Airflow -** https://github.com/apache/airflow
   - Most widely adopted
   - DAG-based workflows in Python.
   - Rich UI, extensible plugins

2. **Prefect (V1 OSS)** - https://github.com/PrefectHQ/prefect
3. **Dagster OSS** - https://github.com/dagster-io/dagster

Cloud Natives:

1. AWS Step Functions
2. Google Cloud Composer (based on Airflow)
3. Azure Data Factory



Business Relevance

✅ **Who Uses It**

Data Engineers, & Analysts who manage workflows with multiple steps or dependencies.

✅ **Why Use It**

To automate, schedule, and monitor complex data pipelines reliably. Business Efficiency

✅ **When to Use It**

When workflows need to run repeatedly, in order, and at scale

Image: Canva AI Prompt Generator

04
Conceptual Design

Medallion Architecture

“The medallion architecture describes a series of data layers that denote the quality of data stored in the lakehouse” - Microsoft

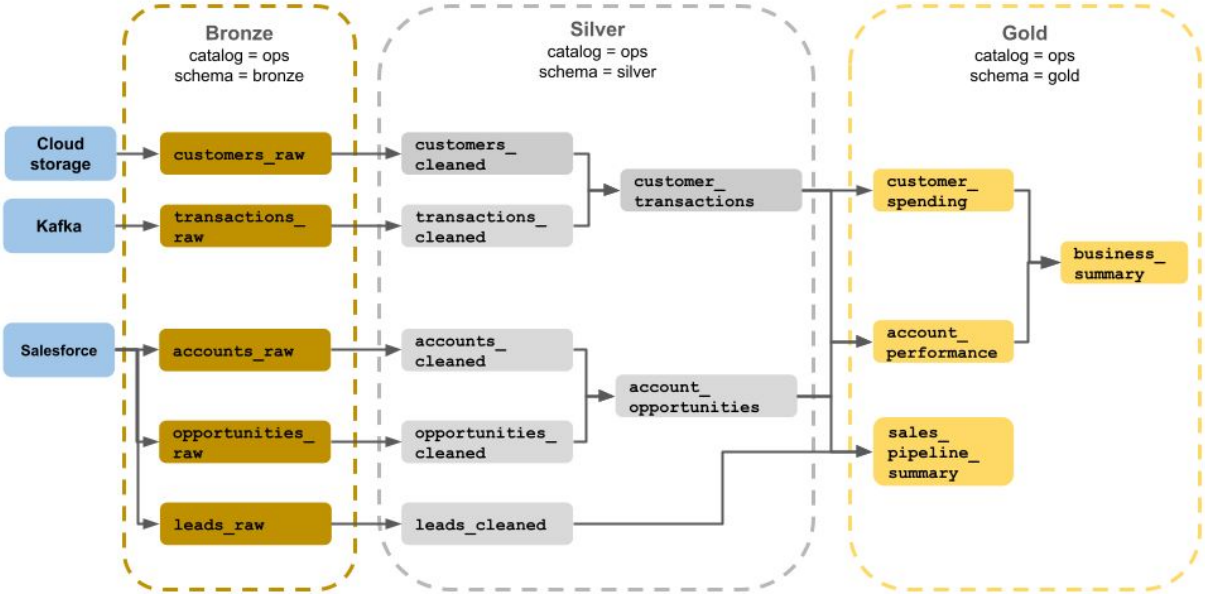


Image Credit: *Medallion Architecture*

Pipeline Anatomy

In this section we present **Medallion Architecture** using Trade Exchange Dummy Data end-to-end from source of structured files (csv), and semi-structured files (JSON) until Insight Layer dashboard

All Processing built in Apache Airflow that using Python Programming & SQL features.

All Data layer (Raw, Refined, Business) use Apache Doris as MPP OLAP Database

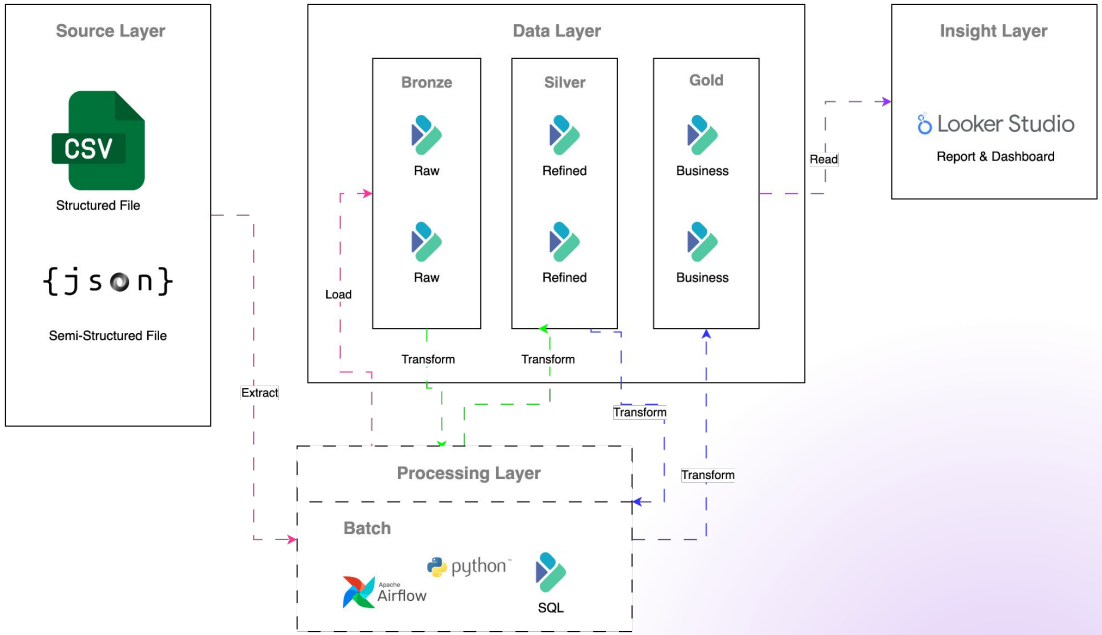
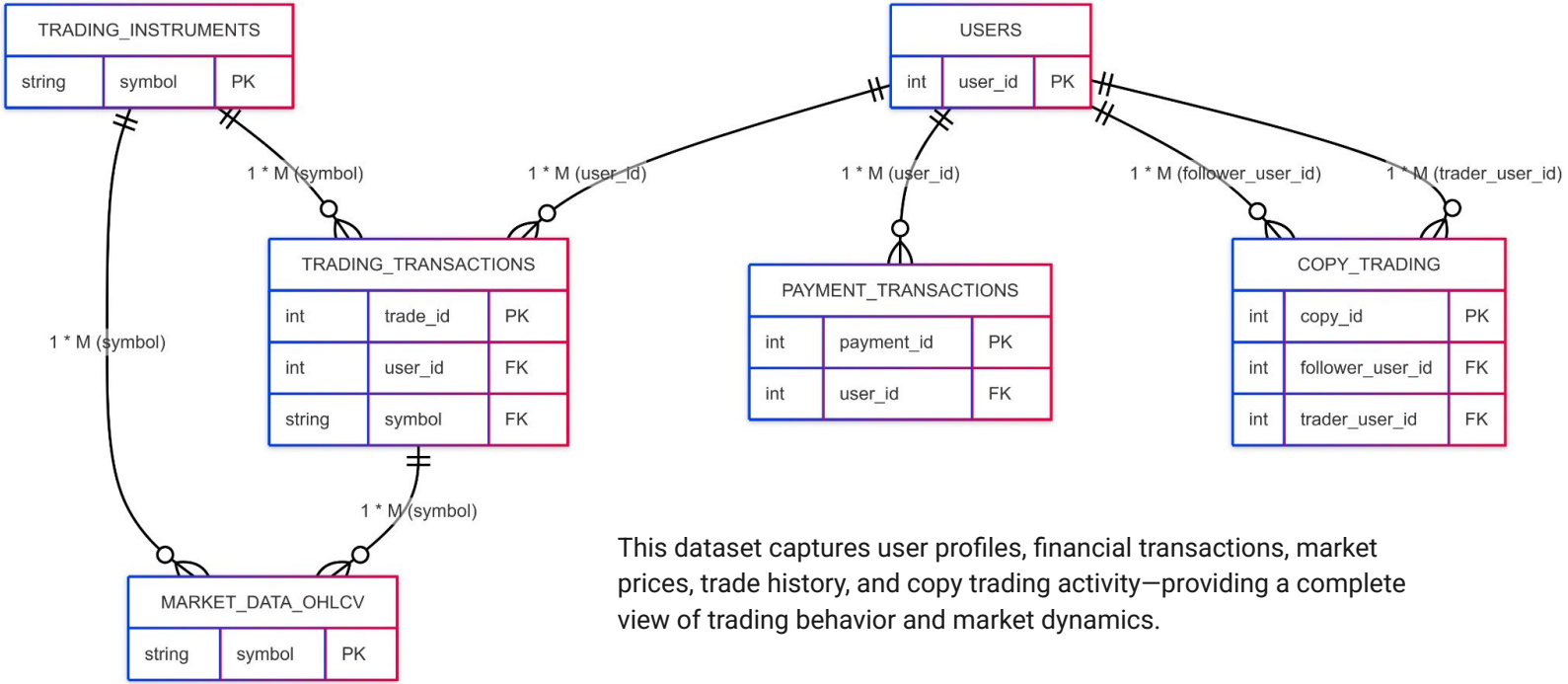


Image Creation: draw.io

ProjekDOS

ERD & Data Mapping



This dataset captures user profiles, financial transactions, market prices, trade history, and copy trading activity—providing a complete view of trading behavior and market dynamics.

Image Creation: mermaid.live

Business Mapping

No.	Schema	Table	Column	Position	Data Type	KEY	Sample Data	Source			
								path	file	type	Column
1	raw	users	user_id	1	text	KEY	USR_BE5FDEC2	~/source_data/	users.csv	csv	user_id
2	raw	users	email_hash	2	text		email_000001@domain.com	~/source_data/	users.csv	csv	email_hash
3	raw	users	registration_date	3	text		2023-08-02	~/source_data/	users.csv	csv	registration_date
4	raw	users	country_code	4	text		US	~/source_data/	users.csv	csv	country_code
5	raw	users	age_group	5	text		18-25	~/source_data/	users.csv	csv	age_group
6	raw	users	account_tier	6	text		Gold	~/source_data/	users.csv	csv	account_tier
7	raw	users	kyc_status	7	text		Verified	~/source_data/	users.csv	csv	kyc_status
8	raw	users	account_status	8	text		Inactive	~/source_data/	users.csv	csv	account_status
9	raw	users	last_login_date	9	date		2023-09-28	~/source_data/	users.csv	csv	last_login_date
10	raw	users	total_deposits	10	numeric(16,2)		7062.94	~/source_data/	users.csv	csv	total_deposits
11	raw	users	total_withdrawals	11	numeric(16,2)		3074.86	~/source_data/	users.csv	csv	total_withdrawals
12	raw	users	current_balance	12	numeric(16,2)		18516.69	~/source_data/	users.csv	csv	current_balance
13	raw	users	is_copy_trader	13	boolean		True	~/source_data/	users.csv	csv	is_copy_trader
14	raw	users	referral_code	14	text		REF_719176	~/source_data/	users.csv	csv	referral_code
15	raw	payment_transactions	payment_id	1	text	KEY	PAY_9963F903	~/source_data/	payment_transactions.csv	csv	payment_id
16	raw	payment_transactions	user_id	2	text		USR_6AA136F1	~/source_data/	payment_transactions.csv	csv	user_id
17	raw	payment_transactions	transaction_type	3	text		Deposit	~/source_data/	payment_transactions.csv	csv	transaction_type
18	raw	payment_transactions	amount	4	numeric(16,2)		3621.05	~/source_data/	payment_transactions.csv	csv	amount
19	raw	payment_transactions	currency	5	text		USD	~/source_data/	payment_transactions.csv	csv	currency
20	raw	payment_transactions	amount_usd	6	numeric(16,2)		6720.67	~/source_data/	payment_transactions.csv	csv	amount_usd
21	raw	payment_transactions	payment_method	7	text		PayPal	~/source_data/	payment_transactions.csv	csv	payment_method
22	raw	payment_transactions	status	8	text		Cancelled	~/source_data/	payment_transactions.csv	csv	status
23	raw	payment_transactions	transaction_date	9	datetime		2022-02-16 00:00:00	~/source_data/	payment_transactions.csv	csv	transaction_date
24	raw	payment_transactions	processing_time_hours	10	numeric(16,2)		51.9	~/source_data/	payment_transactions.csv	csv	processing_time_hours
25	raw	payment_transactions	fee_usd	11	numeric(16,2)		36.79	~/source_data/	payment_transactions.csv	csv	fee_usd
26	raw	payment_transactions	exchange_rate	12	numeric(7,5)		1.4524	~/source_data/	payment_transactions.csv	csv	exchange_rate
27	raw	payment_transactions	payment_provider	13	text		Stripe	~/source_data/	payment_transactions.csv	csv	payment_provider
28	raw	market_data_ohlcv	symbol	1	text	KEY	BTC	~/source_data/	market_data_ohlcv.json	JSON	symbol
29	raw	market_data_ohlcv	date	2	date	KEY	2024-01-01	~/source_data/	market_data_ohlcv.json	JSON	date

More on Spreadsheet: ProjekDOS-Demo-OrchestrationWorkshop.xlsx

ProjekDOS

Data Preview

```
select * from market_data_ohlcv
where date = '2024-01-01'
order by volume desc
limit 20
```

Execution Time: 34 ms

symbol	date	ohlc	volume	spread
USDJPY	2024-01-01	{"open":134.61681,"high":136.41505,"low":132.81857,"close":134.69323}	9983550	0.00524
TSLA	2024-01-01	{"open":214.52363,"high":223.03976,"low":206.0075,"close":209.75225}	8862388	0.00738
MSFT	2024-01-01	{"open":199.67385,"high":205.64536,"low":193.70233,"close":197.36434}	7567912	0.00173
GBPUSD	2024-01-01	{"open":1.42372,"high":1.44754,"low":1.39991,"close":1.41274}	7301816	0.00367
OIL	2024-01-01	{"open":133.7283,"high":136.40884,"low":131.04777,"close":131.91491}	6692675	0.00655
AMZN	2024-01-01	{"open":286.30174,"high":291.67539,"low":280.9281,"close":287.7531}	6406386	0.00345
GOOGL	2024-01-01	{"open":239.64747,"high":242.18667,"low":237.10827,"close":240.15712}	5942586	0.00137
SPY	2024-01-01	{"open":368.52729,"high":377.07246,"low":359.98212,"close":364.96927}	4789368	0.00169
GLD	2024-01-01	{"open":431.5084,"high":439.84675,"low":423.17004,"close":430.8612}	4697109	0.00368
EURUSD	2024-01-01	{"open":1.32245,"high":1.34483,"low":1.30007,"close":1.31542}	4462396	0.0071
ETH	2024-01-01	{"open":3066.29169,"high":3139.97798,"low":2992.6054,"close":3098.78072}	2864149	0.00854
BTC	2024-01-01	{"open":44548.3515,"high":46034.12872,"low":43062.57427,"close":45462.98397}	2154633	0.00198
AAPL	2024-01-01	{"open":121.45156,"high":124.11933,"low":118.78379,"close":122.79023}	1343831	0.00933

05
Technology Stack

Apache Airflow

“Apache Airflow is an open-source platform for developing, scheduling, and monitoring batch-oriented workflows.

A web-based UI helps you visualize, manage, and debug your workflows.” - Airflow

https://airflow.apache.org/docs/apache-airflow/stable/index.html

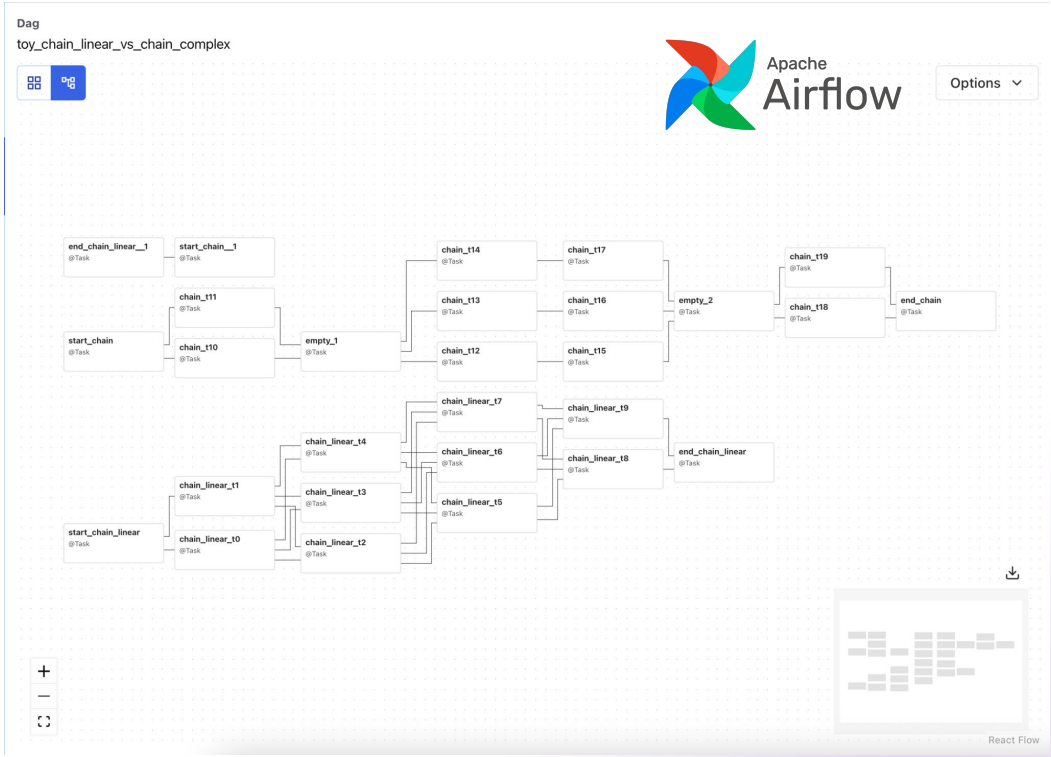


Image Credit: Apache Airflow

ProjekDOS

Webserver

Apache Airflow

Airflow webserver is the component that provides the **web-based user interface (UI)** for Apache Airflow.

- View, trigger, pause/resume DAGs
- Monitor DAG runs and tasks
- View logs
- Manage connections, variables

Hosts the Airflow UI you access in your browser (usually at h**ttp://localhost:8080** or a server IP)

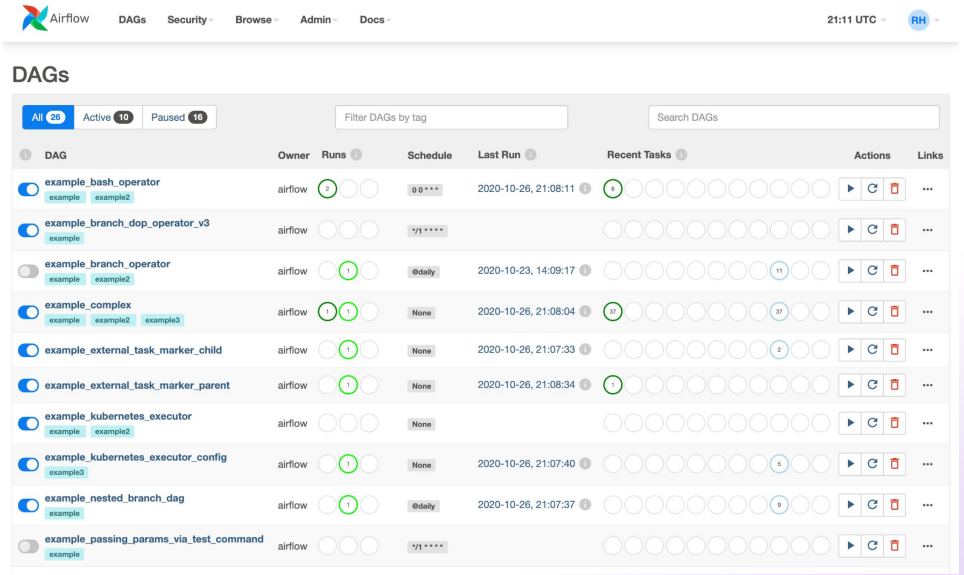


Image Credit: Apache Airflow

ProjekDOS

DAG

Apache Airflow

"A DAG is a model that encapsulates everything needed to execute a workflow

A DAG is defined in a Python script, which represents the DAGs structure (tasks and their dependencies) as code.

The DAG itself doesn't care about *what* is happening inside the tasks; it is merely concerned with *how* to execute them - the order to run them in, how many times to retry them, if they have timeouts, and so on." - Airflow

Image Credit: Airflow

Image Credit: Apache Airflow

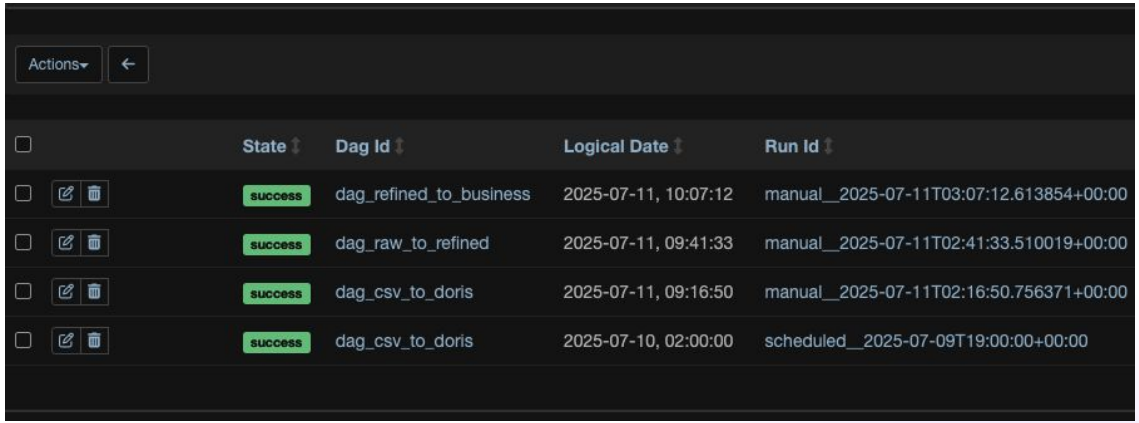
DAG Runs

Apache Airflow

“A DAG Run is an object representing an instantiation of the DAG in time.

Any time the DAG is executed, a DAG Run is created and all tasks inside it are executed. The status of the DAG Run depends on the tasks states.

Each DAG Run is run separately from one another, meaning that you can have many runs of a DAG at the same time.” - Airflow



Cron Expression

Preset	Description	Cron Expression
None	No automatic schedule. Only triggered manually or externally.	*(blank)*
@once	Run one time only.	*(blank)*
@hourly	Run every hour, at the start of the hour.	0 * * * *
@daily	Run Daily at midnight.	0 0 * * *
@weekly	Run Weekly at midnight on Sunday.	0 0 * * 0
@monthly	Run Monthly at midnight on the 1st day.	0 0 1 * *
@yearly	Run Yearly at midnight on January 1.	0 0 1 1 *

Cron Generator: https://crontab.guru/

Apache Doris

“Apache Doris , Apache Doris is an MPP-based data warehouse known for its high query speed. It can be used for report analysis, ad-hoc queries, unified data warehouse, and data lake query acceleration” - Doris

https://doris.apache.org/docs/gettingStarted/what-is-apache-doris

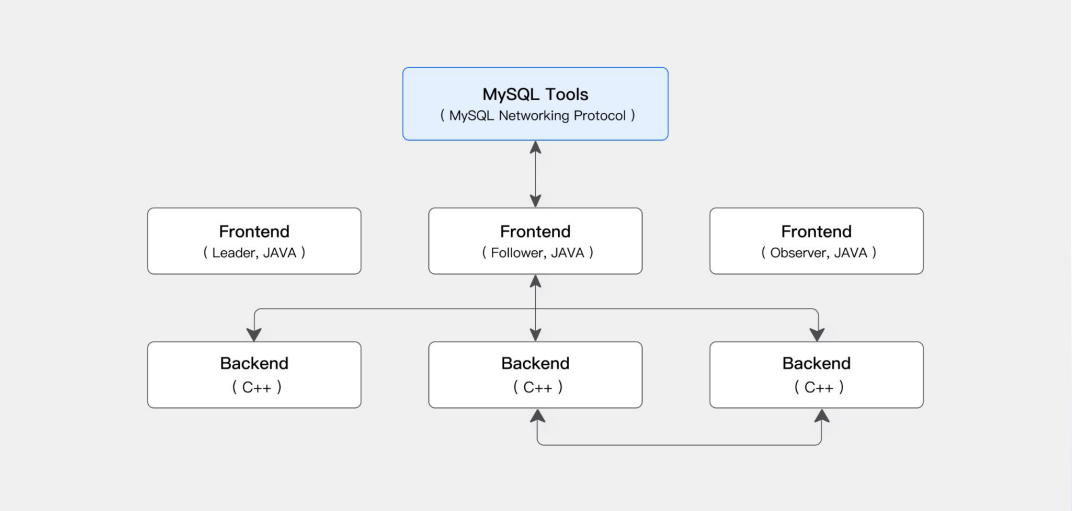


Image Credit: Apache Doris

Doris Architecture

“Architecture Apache Doris is simple and neat with only two types of processes:

- Frontend (FE): It is responsible for user request processing, query parsing and planning, metadata management, and node management.
- Backend (BE): It is used to store data and execute query plans “

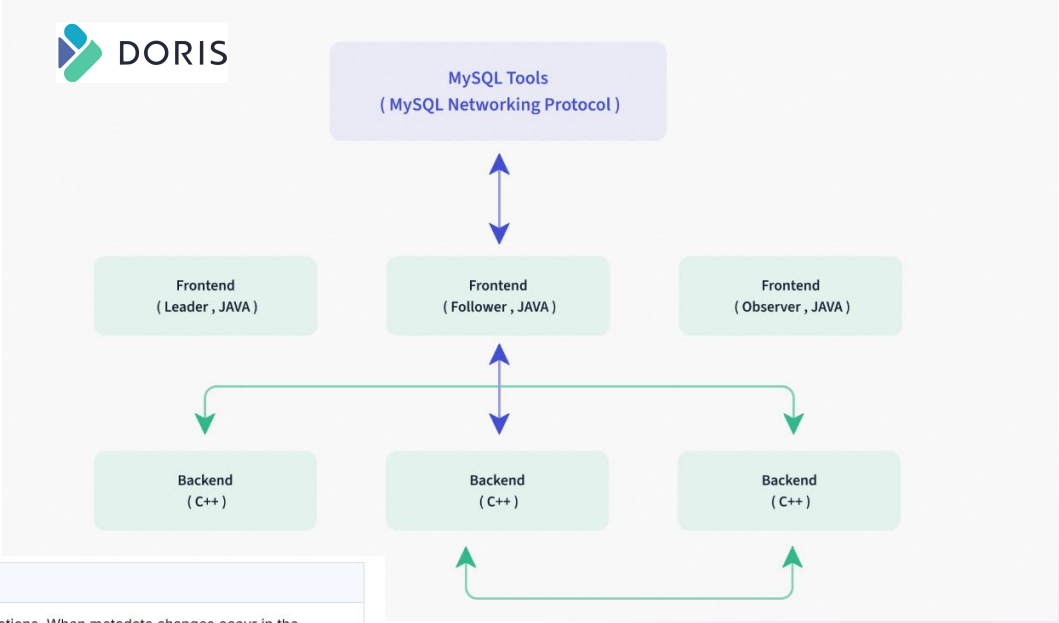


Image Credit: Alibaba

Role	Function
Master	The FE Master node is responsible for metadata read and write operations. When metadata changes occur in the Master, they are synchronized to Follower or Observer nodes via the BDB JE protocol.
Follower	The Follower node is responsible for reading metadata. If the Master node fails, a Follower node can be selected as the new Master.
Observer	The Observer node is responsible for reading metadata and is mainly used to increase query concurrency. It does not participate in cluster leadership elections.

DWH Benchmark

DORIS

"Benchmark Performance of Doris on the Star Schema Benchmark (SSB) 1000GB test set. (5,999,989,709 rows)
https://www.cs.umb.edu/~poneil/StarSchemaB.PDF

Doris tested 13 queries on the Star Schema Benchmark standard test dataset and result"

Hardware	Configuration Instructions
Number of Machines	4 Aliyun Virtual Machine (1FE, 3BEs)
CPU	Intel Xeon (Ice Lake) Platinum 8369B 32C
Memory	128G
Disk	Enterprise SSD (PL0)

Q3. In our third query flight, we want to place restrictions on three dimensions, including the remaining dimension, customer. We base our query on TPCQ5. The query is intended to provide revenue volume for lineorder transactions by customer nation and supplier nation and year within a given region, in a certain time period.

Q3 select c_nation, s_nation, d_year, sum(lo_revenue) as revenue from customer, lineorder, supplier, date
where lo_custkey = c_custkey
and lo_suppkey = s_suppkey
and lo_orderdate = d_datekey
and c_region = 'ASIA' and s_region = 'ASIA'
and d_year >= 1992 and d_year <= 1997
group by c_nation, s_nation, d_year
order by d_year asc, revenue desc;

Q3.1 Q3 as written: c_region = 'ASIA' so FF = 1/5 for customer, FF = 1/5 for supplier, and 6-year period FF = 6/7 for d_year; Thus LINEORDER FF = (1/5)*(1/5)*(6/7) = 6/175 and the number of lineorder rows selected, for SF = 1, is (6/175)*6,000,000 ≈ 205,714.

Sample Q3.1

Total: 19390 ms means all queries combined took just under **20 seconds** to execute.

Query	Doris 2.0.15.1 (ms)	
q1.1	330	Simple Queries
q1.2	80	
q1.3	80	
q2.1	1780	Medium Complex Queries
q2.2	1970	
q2.3	1510	
q3.1	4000	More Complex Queries
q3.2	1720	
q3.3	1510	
q3.4	160	
q4.1	4010	Large Scan & Heavy Queries
q4.2	840	
q4.3	400	
Total	19390	

Image Credit: Doris Benchmark

VeloDB Studio

“VeloDB is a modern real-time data warehouse for blazing-fast analytics on large-scale data, & **VeloDB Studio** is a GUI tool designed for Apache Doris and its compatible databases, simplifying data development and management” - VeloDB

https://www.velodb.io

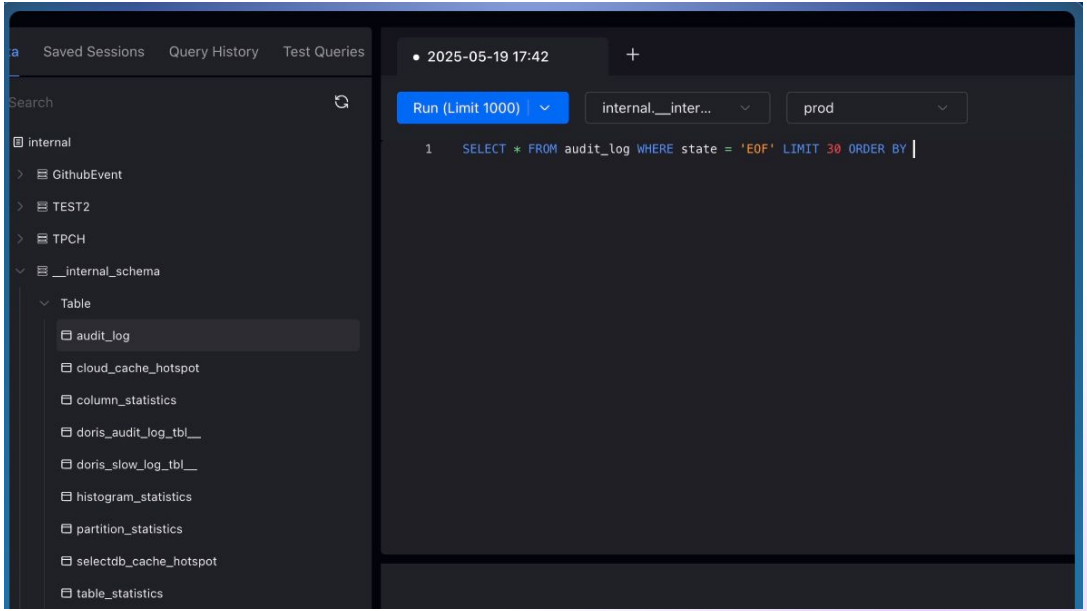
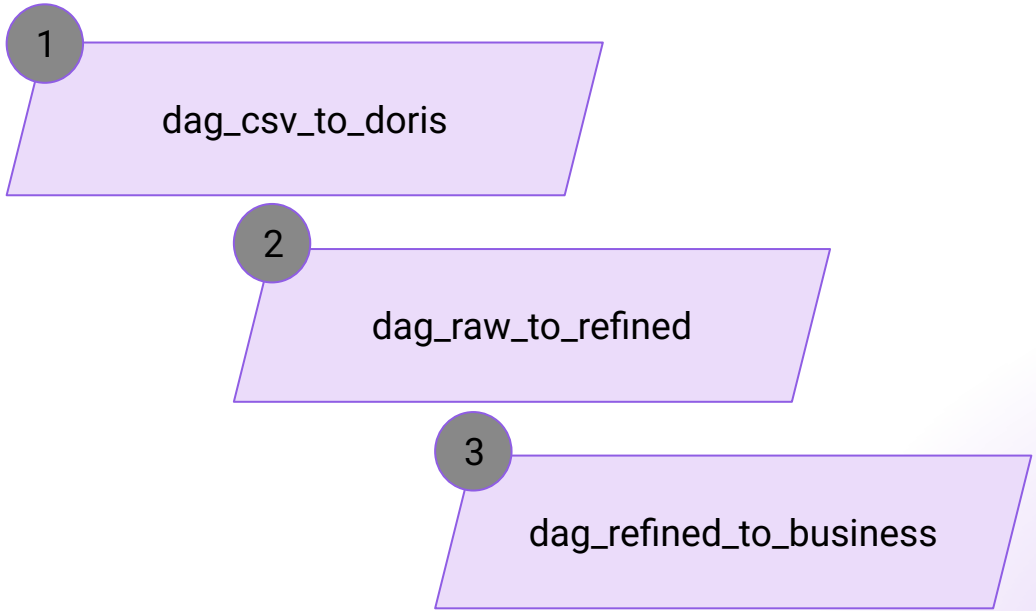


Image Credit: VeloDB

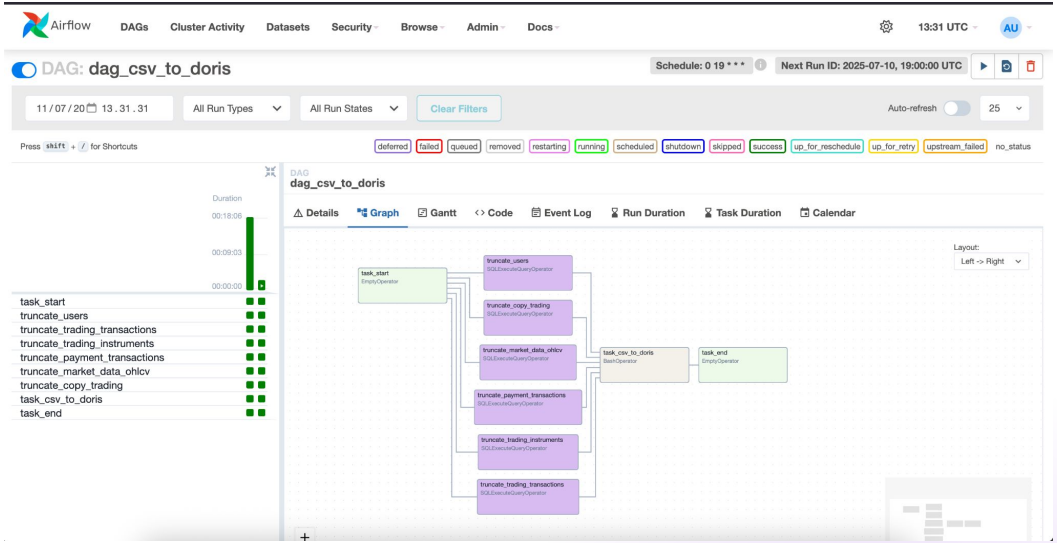
06
Airflow & Doris Hands On

Dags run sequence (Overview)



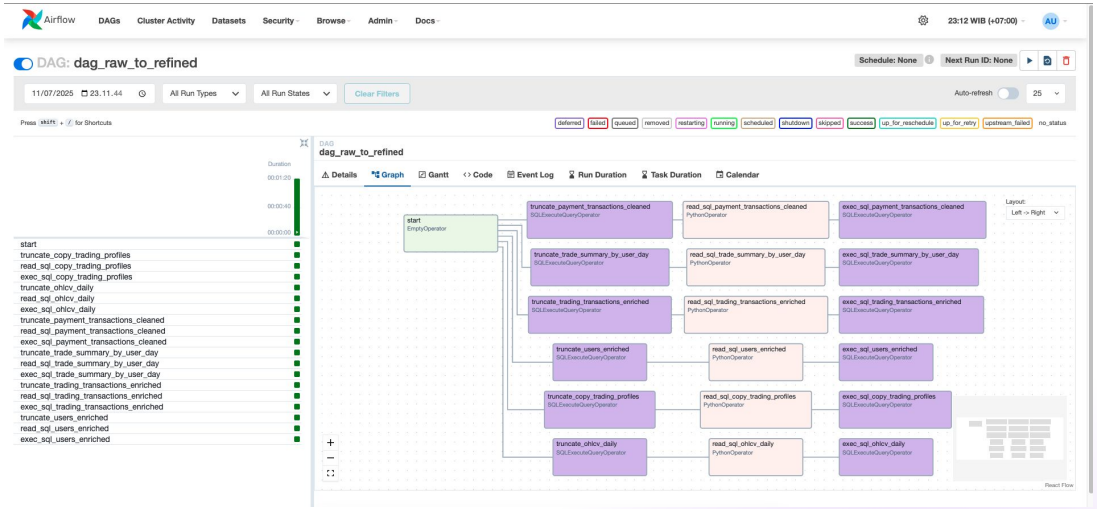
1. DAG Stream Load to Doris

This DAG will write all the tables from CSV, & JSON source files to Apache Doris using the built-in API provided by Apache Doris. All data written to this table is as-is, meaning there is no transformation. This represents the first layer of the medallion architecture.



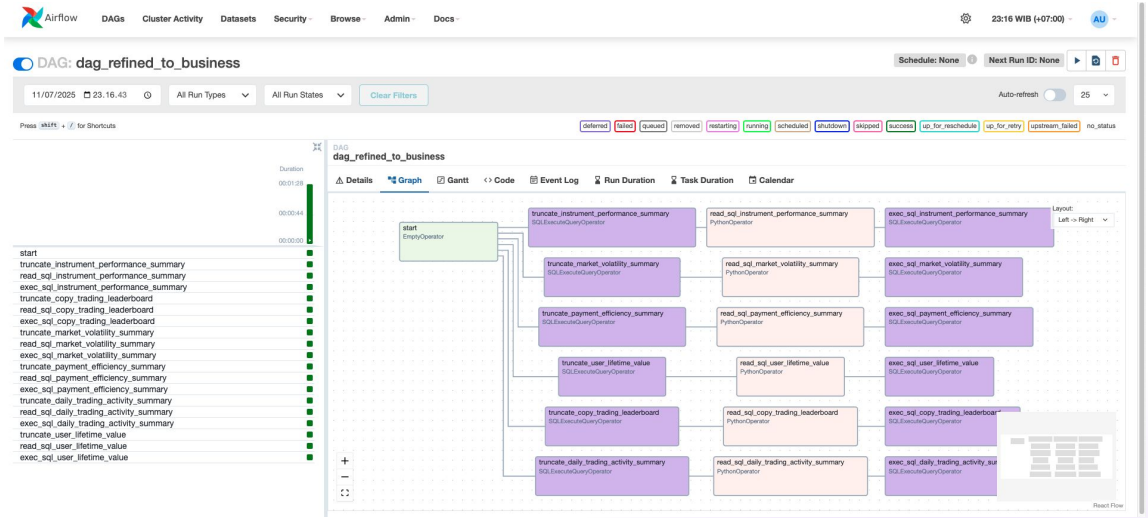
2. DAG Raw to Refined

This is the second layer of the Medallion architecture (Silver Layer). For example, the table copy_trading_profiles summarizes traders' profiles, allowing us to determine how many followers a trader has and how much profit they have earned.



3. DAG Refined to Business

DAG Process from Silver to Third Layer: **Gold Layer** in the Medallion architecture represents curated, **business-ready data** that directly supports analytics, reporting, and decision-making.



DDL & ELT Script

1. Prepare Doris DDL Script

Doris DDL scripts must be prepared according to the data mapping as a foundation before implementing the ETL process.

2. Prepare Doris Load Script

Doris stream load serves as the primary mechanism for ingesting raw data into the platform, forming the foundation of the ETL pipeline.

```
## raw.payment_transactions (CSV)
curl --location-trusted \
  -u engineer:"engineer" \
  -H "Expect:100-continue" \
  -H "column_separator:," \
  -H "columns:payment_id,user_id,transaction_type,amount
  processing_time_hours,fee_usd,exchange_rate,payment_pr
  -T /home/administrator/data/files/payment_transactions
  -H "max_filter_ratio:0.1" \
  -XPUT http://localhost:8030/api/raw/payment_transactio

## raw.market_data_ohlcv (JSON)
curl --location-trusted \
  -u root:"" \
  -H "Expect:100-continue" \
  -H "format: json" \
  -H "strip_outer_array: true" \
  -H "read_json_by_line: false" \
  -H "columns: symbol, date, ohlc, volume, spread" \
  -H "max_filter_ratio: 0.1" \
  -T /home/administrator/data/files/market_data_ohlcv.js
  -XPUT http://localhost:8030/api/raw/market_data_ohlcv/
```

3. Prepare Doris SQL Script

Preparing the Doris SQL scripts according to the data mapping is essential before proceeding with the orchestration setup in Airflow.

insert_doris_medallion_lakehouse.sql

Orchestrator > git > script > insert_doris_medallion_lakehouse.sql

```sql
-- User lifetime value
INSERT INTO business.user_lifetime_value
SELECT
    u.user_id,
    u.account_tier,
    u.country_code,
    COUNT(t.trade_id) AS total_trades,
    COALESCE(SUM(t.net_profit), 0) AS total_net_profit,
    COALESCE(SUM(t.pnl_usd), 0) AS gross_pnl,
    COALESCE(SUM(t.commission_usd + t.swap_usd), 0) AS total_fees,
    COALESCE(p.total_deposits, 0) AS total_deposits,
    COALESCE(p.total_withdrawals, 0) AS total_withdrawals,
    u.current_balance,
    CASE
        WHEN COALESCE(p.total_deposits, 0) > 0
        THEN COALESCE(SUM(t.net_profit), 0) / p.total_deposits
        ELSE NULL
    END AS roi_ratio,
    CASE
        WHEN COALESCE(SUM(t.net_profit), 0) > 0 THEN 'profitable'
        WHEN COALESCE(SUM(t.net_profit), 0) < 0 THEN 'losing'
        ELSE 'neutral'
    END AS trader_type
FROM refined.users_enriched u
LEFT JOIN refined.trading_transactions_enriched t ON u.user_id = t.user_id
LEFT JOIN (
    SELECT
        user_id,
        SUM(CASE WHEN transaction_type = 'Deposit' THEN amount_usd ELSE 0 END) AS total_deposits,
        SUM(CASE WHEN transaction_type = 'Withdrawal' THEN amount_usd ELSE 0 END) AS total_withdrawals
    FROM refined.payment_transactions_cleaned
    GROUP BY user_id
) p ON u.user_id = p.user_id
GROUP BY u.user_id, u.account_tier, u.country_code, p.total_deposits, p.total_withdrawals, u.current_balance;
```

4. Prepare Airflow DAG Script

Prepare the Airflow DAG script to automate and orchestrate the data pipeline, ensuring each component runs in sequence based on the defined workflow.

Parsed at: 2025-07-11, 15:11:58 WIB

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.models import Variable
from airflow.operators.empty import EmptyOperator
from airflow.providers.common.sql.operators.sql import SQLExecuteQueryOperator
from datetime import datetime, timedelta
import pendulum


param = Variable.get('dag_csv_to_doris', deserialize_json=True)
shell_path=param['shell_path']
items = [
        "users",
        "trading_transactions",
        "trading_instruments",
        "payment_transactions",
        "market_data_ohlcv",
        "copy_trading"
    ]

args = {
    'owner': 'frederik',
    'depends_on_past': False,
    'email_on_failure': False,
    'email_on_retry': False,
```

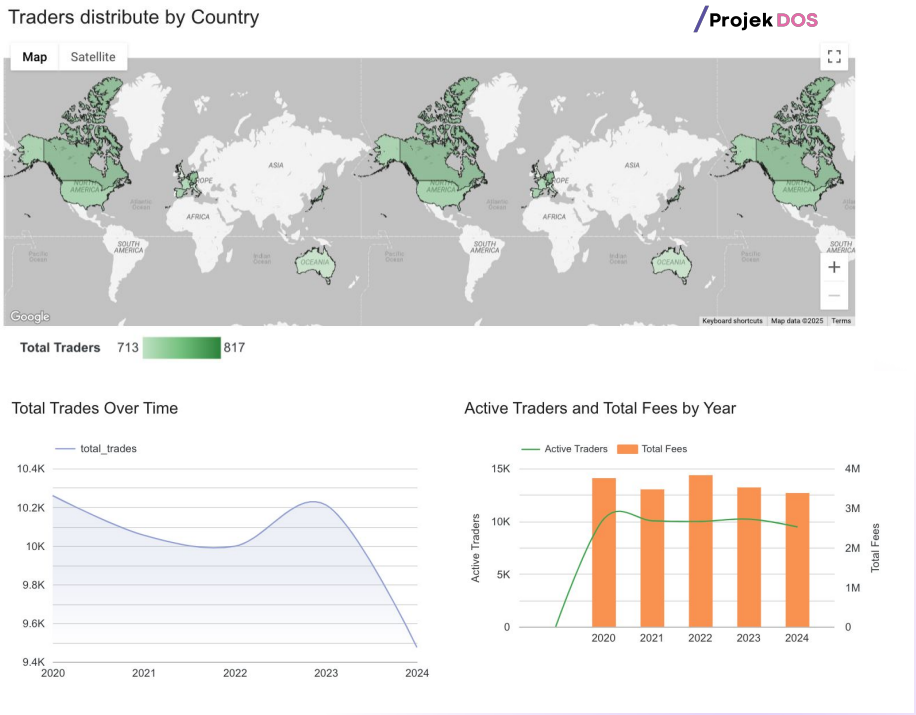
5. Demo

Insight Layer (Bonus)

After all populated data ready on Doris Business Layer. We can prepare & build Dashboard for Decision Making Purposes.

Trade Exchange Dashboard (Looker) -

https://lookerstudio.google.com/reporting/e76b66c4-773c-4c0c-a5d0-195e93b5af29



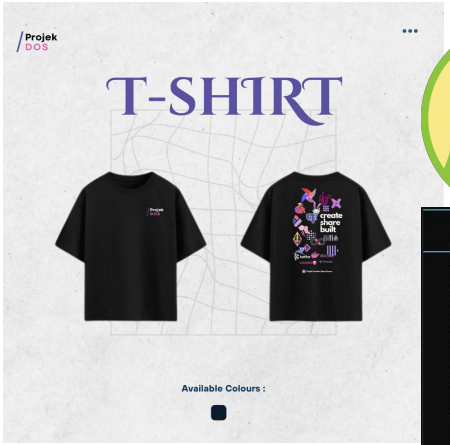
06
Summary & Quiz

Summary

Participants revisited key concepts of orchestration and data warehousing within the medallion architecture framework and learned to apply these principles using open-source tools such as Apache Airflow and Doris.

The session also provided valuable insights into career opportunities in IT, particularly in data-focused roles.

Quiz



DORIS

Apache Airflow

- URL: https://github.com/projekdos
- Repository: orchestrator_workshop_airflow_doris_batch1
- Read the Instruction (Mapping file)
- Solve & Build

Please submit the result to info@projekdos.com / projek.freedomopensource@gmail.com . The winner will get Free Merchandise from us!

Thanks!

Business Contact:

info@projekdos.com

+6281385368844 (Whatsapp)

Whatsapp Community

LinkedIn

Youtube Channel

/ProjekDOS