/Projek DOS

Workshop

Building ETL using Apache Hop

Visual Data Engineering Made Simple

/Projek DOS

Table of contents

01
Introduction

Introduction

Projek Freedom Open Source

Founded in 2025, This Training Center is a **project-based, real world scenario** training environments initiative focused on education in open-source technology. The goal is to empower individuals, university, and corporate with knowledge about tools and technologies such as :

- Business Intelligence
- Data Warehouse
- Data Lake
- Big Data
- Blockchain
- Artificial Intelligence
- More

Open Partnership

University 

Integrating practical, hands-on training into workshops, and certification programs for students.

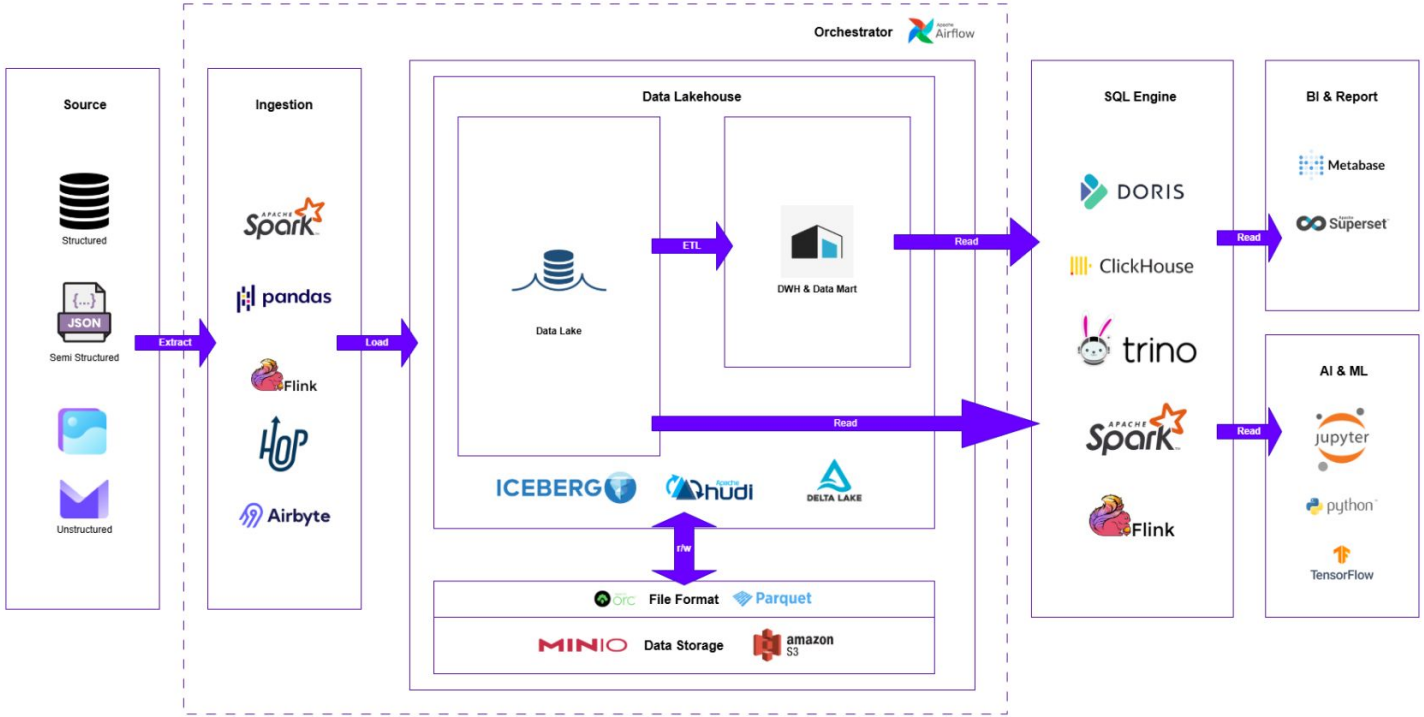
Corporate 

providing tailored training programs to upskill employees,and align talent with industry needs— talent acquisition partners

Government 

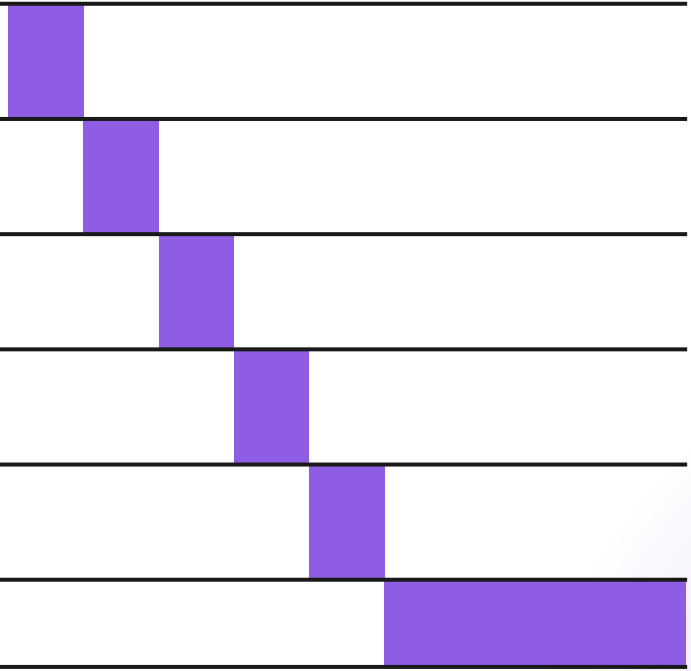
Supporting national and regional initiatives through specialized training programs, seminars, and collaborative education projects

End to End - Project Based Training

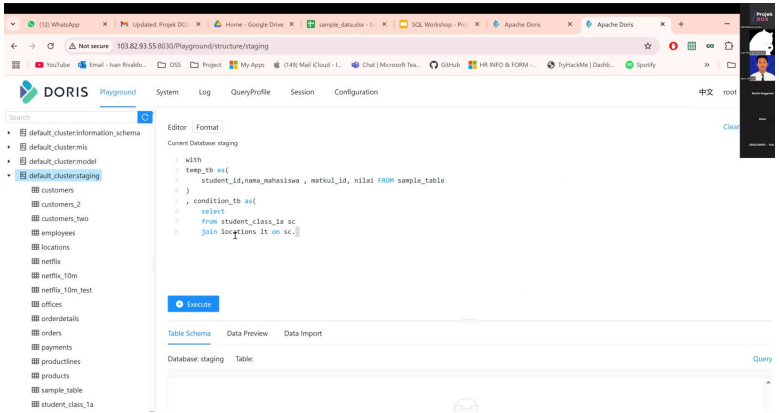


Roadmap (2025)

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 - Learn 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

02
ETL

About Me - Trainer

Firman Fakhri Mukti

ETL & BI Developer with 8+ years of experience in Banking and Insurance. Expert in building ETL pipelines, data warehouses, and BI dashboards to support data-driven decision-making

Firman Fakhri Mukti

/ProjekDOS

ETL Careers Path

- **ETL Developer** (Other Skills: SQL, Store Procedure)
- **Data Engineering** (Other Skills: Python, Orchestrator)
- **Data Architect** (Other Skills: Data Governance)

Demand IT Workers will continue to increase in the coming years. **(kemnaker)**

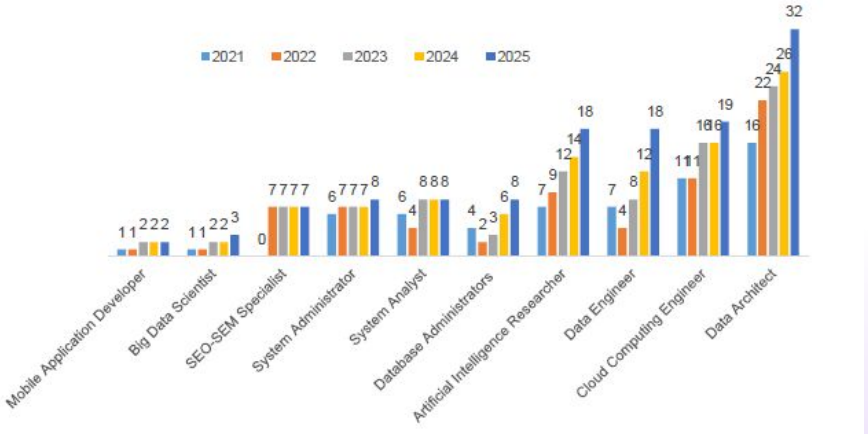


Image Credit: kemnaker

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**, depending on the industry, 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

Training Material

- What is ETL
- Most Popular ETL
- Why ETL is Important
- Who used ETL
- Pipeline Anatomy
- Multisource Integration
- Data Manipulation
- Data Load

What is ETL?

Definition:
ETL stands for **Extract, Transform, Load**. It is a data integration process that combines data from multiple sources into a central repository, such as a data warehouse. The process involves:

1. **Extracting** data from various sources like databases, APIs, or files.
2. **Transforming** the data by cleaning, formatting, and applying business rules to make it suitable for analysis.
3. **Loading** the transformed data into a target system for further use.

This process is fundamental in preparing data for analytics, reporting, and decision-making.

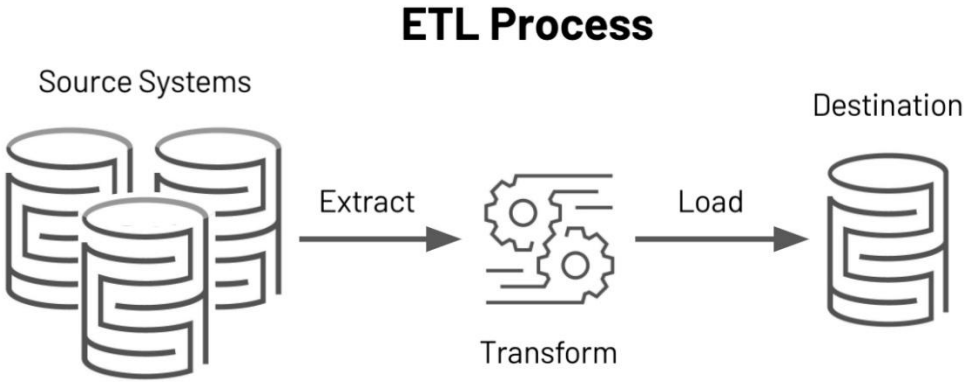


Image Credit: Extract Transform Load (ETL) – Databricks

Most Popular ETL

ETL tools are well-established in the data industry. Some of the most popular and mature solutions include:

- **SQL Server Integration Services (SSIS)** – A Microsoft tool for building enterprise-level data integration and workflow solutions.
- **IBM InfoSphere DataStage** – An enterprise-grade ETL platform from IBM, designed for large-scale and complex data environments.
- **Apache Hop** – An open-source ETL tool developed by the Apache Software Foundation.
- **Talend Data Integration** – A widely used ETL tool for enterprise versions, known for its flexibility and wide range of connectors.
- **Pentaho Data Integration (PDI)** – An Enterprise ETL solution with a graphical interface, also known as Kettle, used for batch and real-time data processing.

Why ETL is important

1. **Combines Data from Many Sources**

ETL helps you bring together data from different places — like Excel files, databases, or APIs — into one single storage. Think of it like collecting puzzle pieces from different boxes and putting them together into one clear picture.

2. **Cleans and Organizes Messy Data**

Raw data is often messy or incomplete. The “Transform” step in ETL fixes the data — making it clean, consistent, and ready for use.

3. **Makes Reports More Accurate**

Clean and complete data means better dashboards and reports — so teams can make smarter, data-based decisions.

4. **Saves Time and Effort**

ETL processes can run automatically on a schedule. That means no more manual data cleaning every day or week.

Who uses ETL?

/ProjekDOS

👨‍💻 **Data Engineers**
They design and build ETL pipelines to move and prepare data for use.

📊 **Data Analysts**
They use the cleaned and organized data from ETL to create reports, dashboards, and insights.

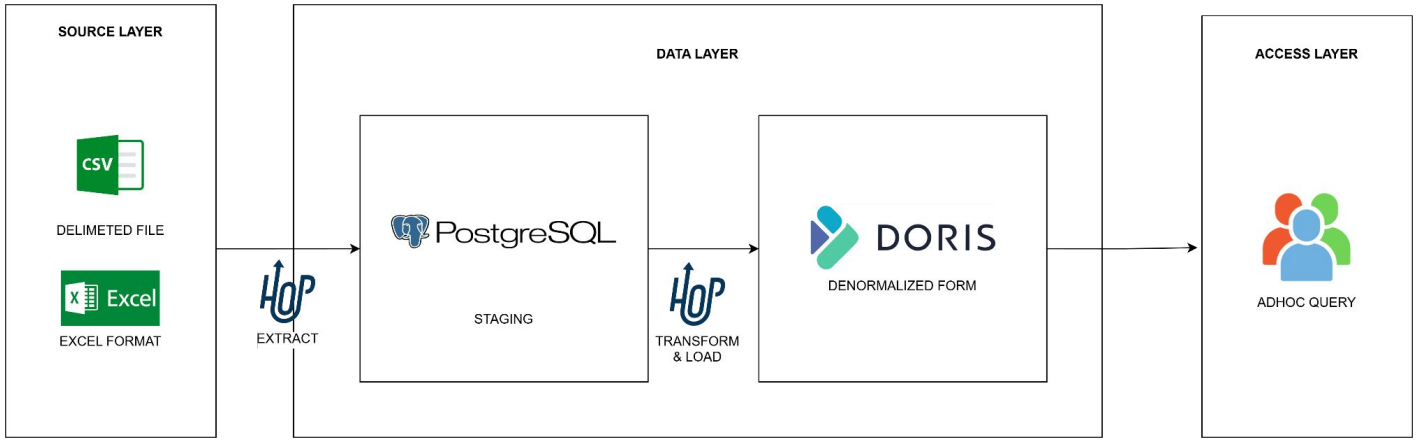
📈 **Business Intelligence (BI) Teams**
BI professionals rely on ETL to feed accurate data into analytics tools like Power BI or Tableau.

📂 **Data Scientists**
They need structured and clean data (often from ETL) to train machine learning models.

👔 **Business Users & Decision Makers**
They don't build ETL — but benefit from it! Thanks to ETL, they can view dashboards and reports to support decisions.

🏢 **Companies in All Industries**
Retail, banking, healthcare, startups, government — all use ETL to understand their data better.

Pipeline Anatomy



Tech Stack (ETL)

Apache Hop , short for **H**op **O**rchestration **P**latform, is a data orchestration and data engineering platform that aims to facilitate all aspects of data and metadata orchestration

Hop is an entirely new **open source** data integration platform that is easy to use, fast and flexible

Source :
https://hop.apache.org/manual/latest/getting-started/hop-what-is-hop.html

Tech Stack (Database)

PostgreSql , PostgreSQL is a powerful, open source object-relational database system that uses and extends the SQL language combined with many features that safely store and scale the most complicated data workloads.

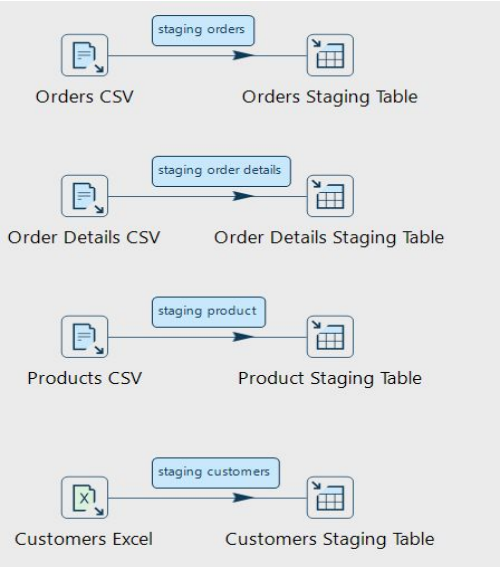
Tech Stack (OLAP)

Apache Doris , Apache Doris is an MPP-based data warehouse known for its high query speed. For queries on large datasets, it returns results in sub-seconds

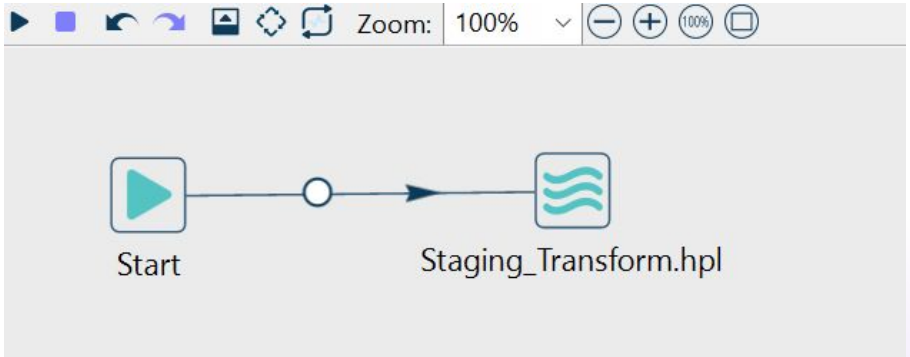
Source :
https://doris.apache.org/docs/gettingStarted/what-is-apache-doris

Pipelines Vs Workflows

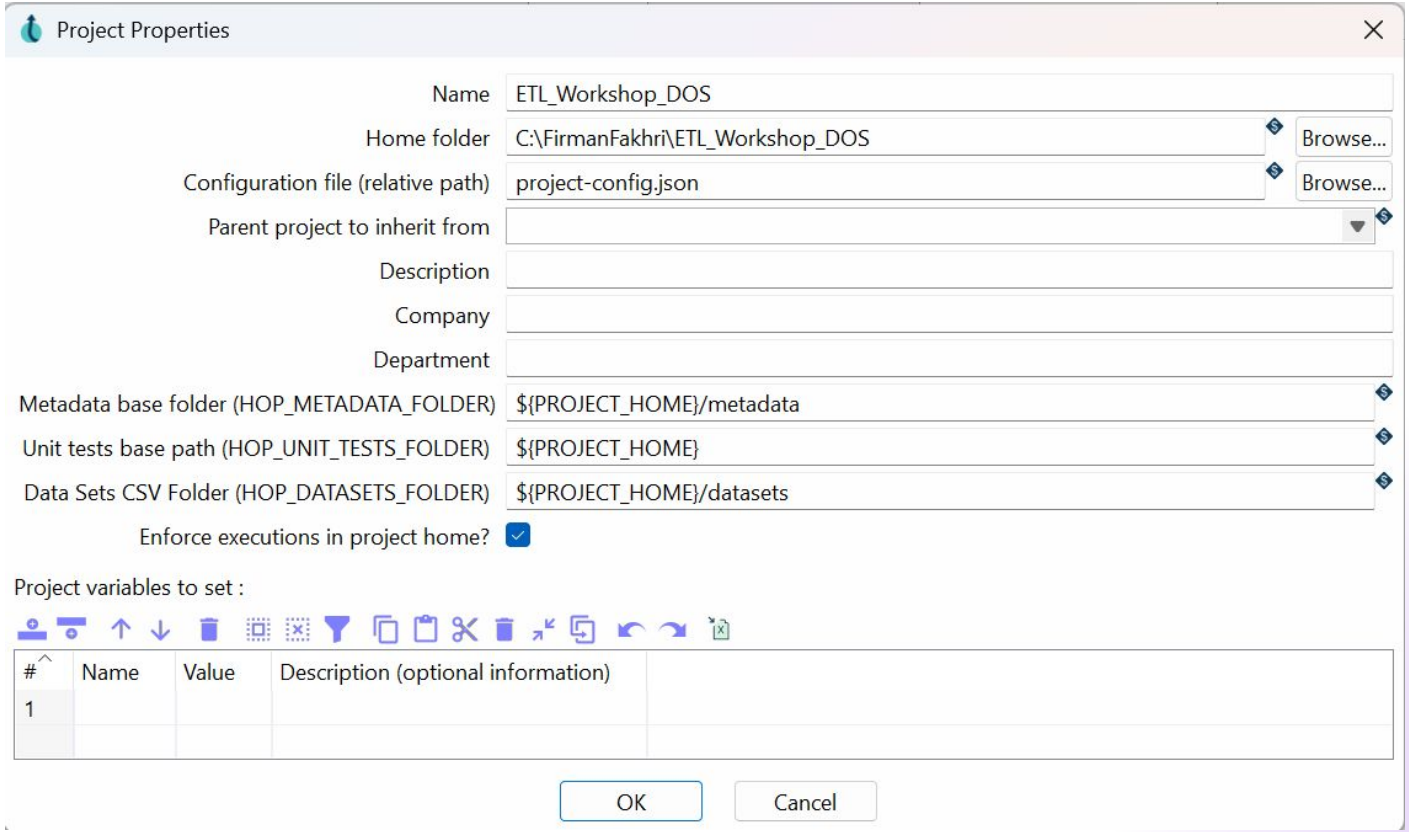
Pipelines are Hop's work horse: read from sources, write to targets and perform just about any manipulation on your data through hundreds of **transforms**.



Workflows are Hop's tool to orchestrate workflows and pipelines, perform environment validations, error handling and much more with the available **actions**

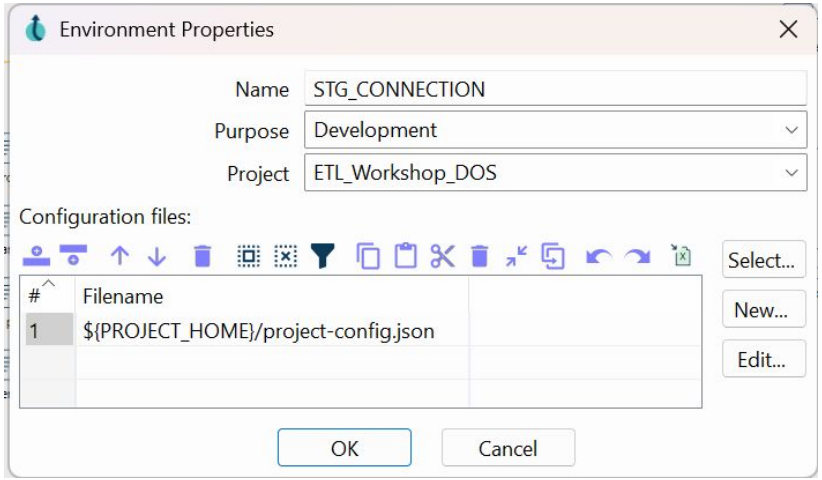


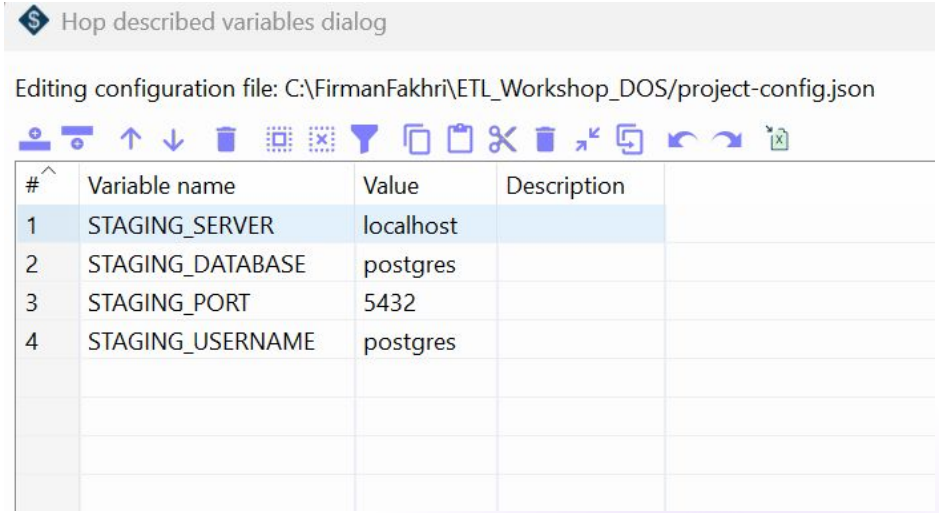
Project Setup



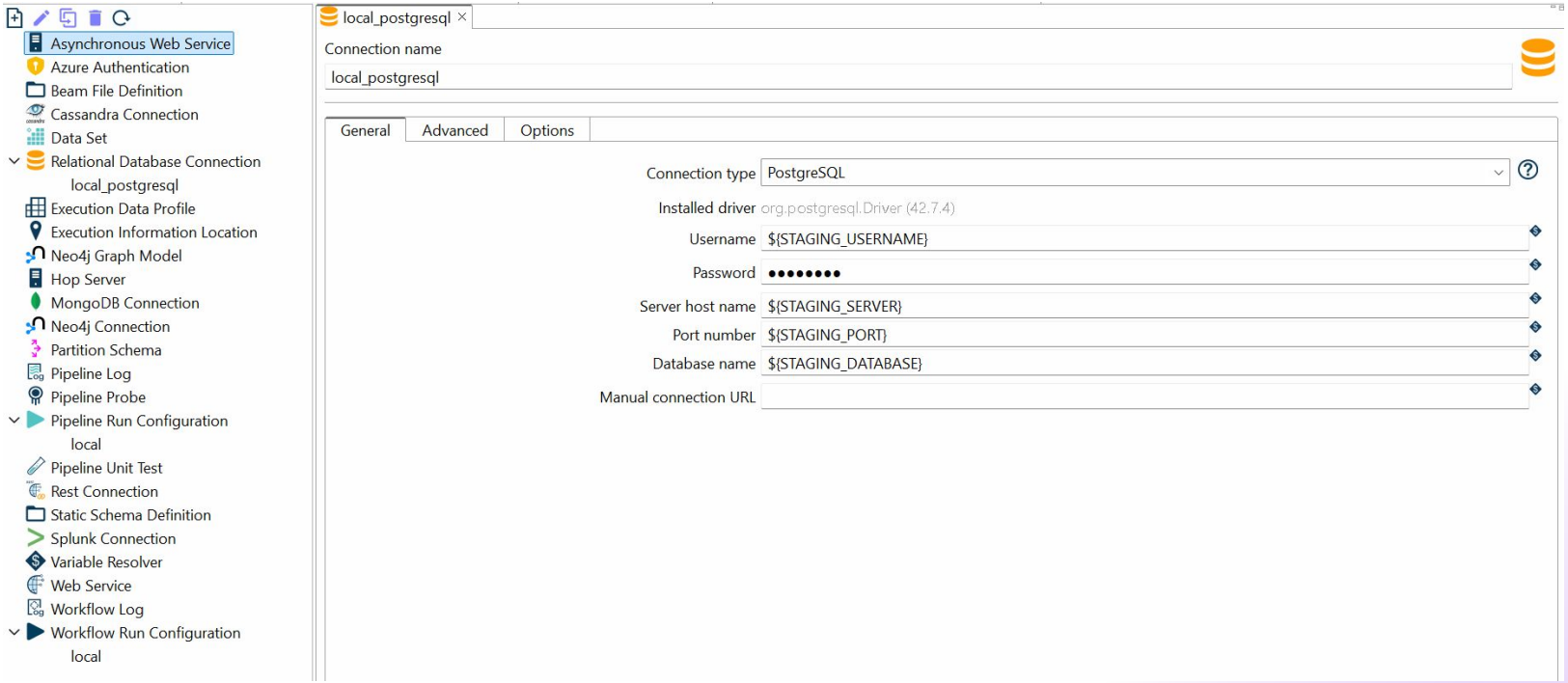
/ProjekDOS

Project Environment Variable

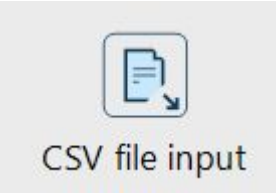




Metadata Connection



Transform (CSV File Input)



CSV file input

Transform name: Products CSV

Filename: ${PROJECT_HOME}/files/20051231.products.csv [Browse...]

Delimiter: | [Insert TAB]

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☑

Header row present? ☑

Add filename to result ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

File encoding:

Schema definition:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	productCode	String		168		$	.	,	none
2	productName	String		43		$	.	,	none
3	productLine	String		16		$	.	,	none
4	productScale	String		6		$	.	,	none
5	productVendor	String		25		$	.	,	none
6	productDescription	String		391		$	.	,	none
7	quantityInStock	Integer	#	15	0	$	.	,	none
8	buyPrice	Number	#.#	15	1	$	.	,	none
9	MSRP	Number	#.#	15	1	$	.	,	none

Transform (Table Output)

Table output

Transform name: Product Staging Table
Connection: local_postgresql
Target schema: staging (Browse...)
Target table: products (Browse...)
Commit size: 1000
Truncate table: ☑
Truncate on first row: ☐
Ignore insert errors: ☐
Specify database fields: ☑

Main options | Database fields

Fields to insert:

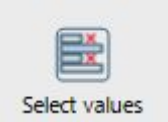
Get fields
Enter field mapping

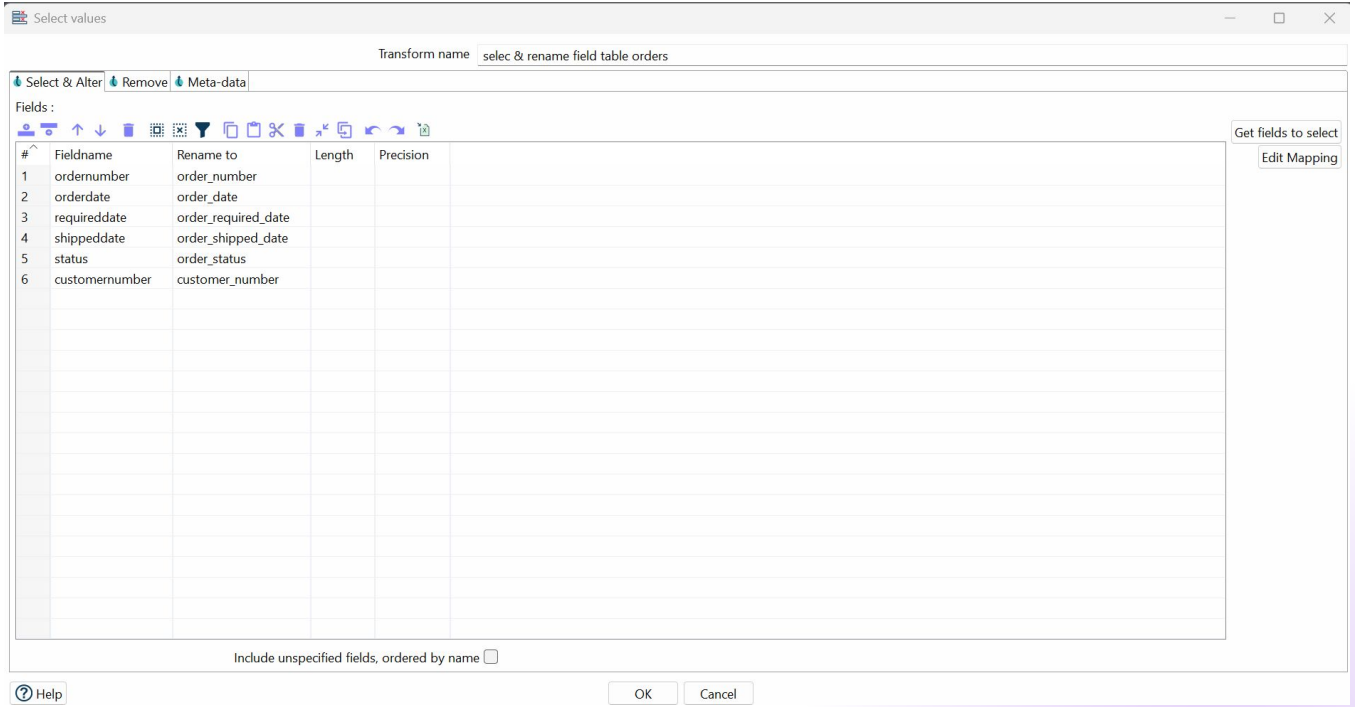
#	Table field	Stream field
1	productCode	productCode
2	productName	productName
3	productLine	productLine
4	productScale	productScale
5	productVendor	productVendor
6	productDescription	productDescription
7	quantityInStock	quantityInStock
8	buyPrice	buyPrice
9	MSRP	MSRP

Mapping Table

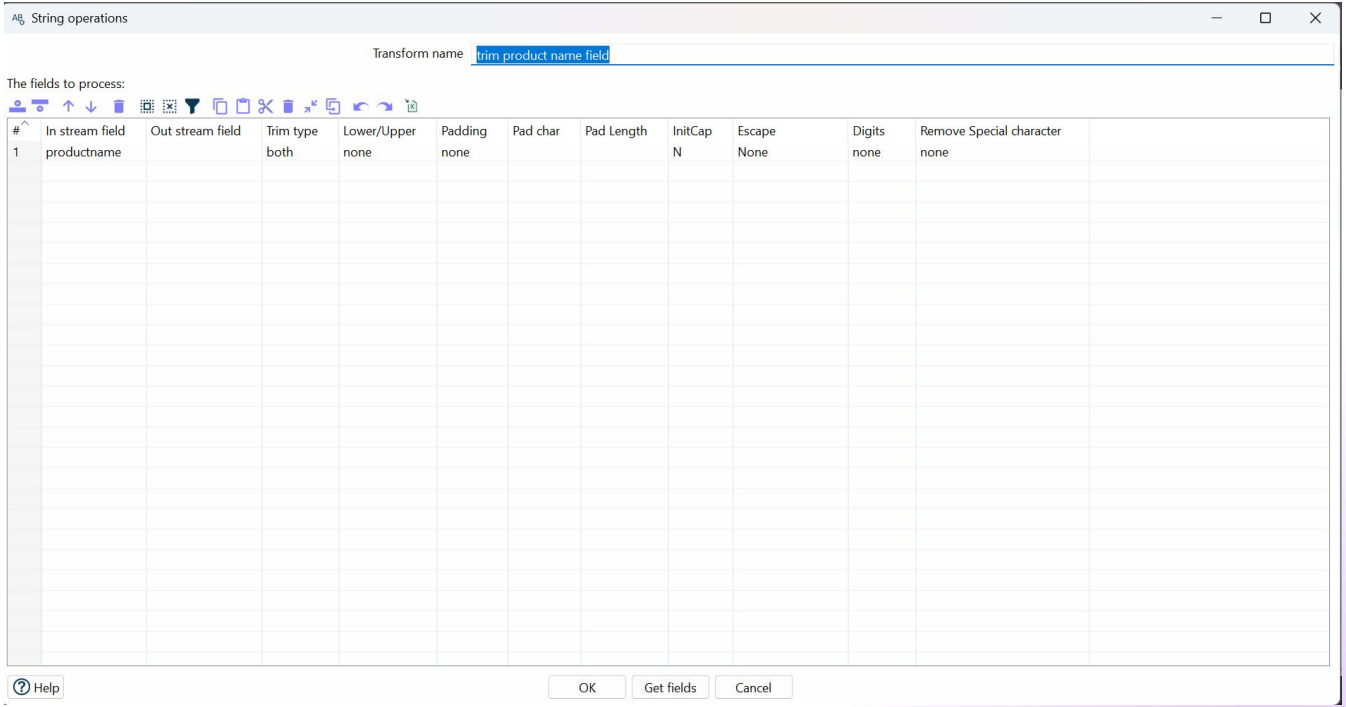
No.	Column	Source		
		Table	Column	Business Rule
1	order_date	orders	orderDate	
2	order_required_date	orders	requiredDate	
3	order_shipped_date	orders	shippedDate	
4	order_number	orders	orderNumber	
5	order_status	orders	status	
6	order_quantity	orderDetails	quantityOrdered	join orders to order detail on ordernumber = ordernumber
7	order_price_each	orderDetails	priceEach	case when currency USD then priceEach * 1 else when currency IDR then priceEach / 16500 end
8	order_amount	orderDetails	order_amount	quantityOrdered * priceEach
9	order_line_number	orderDetails	order_line_number	
10	product_code	products	productCode	join orders to order detaill on ordernumber = ordernumber --> join order detail to product on productcode = productcode
11	product_name	products	productName	cleansing special character (space,#,$)
12	product_scale	products	productScale	
13	product_description	products	productDescription	
14	product_category	products	productLines	
15	customer_number	customers	customerNumber	join orders to customers on customernumber = customernumber
16	customer_name	customers	customerName	
17	customer_city	customers	city	
18	customer_state	customers	state	
19	customer_country	customers	country	

Transform (Select Values)

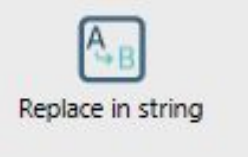




Transform (String Operations)



Transform (Replace in String)



Replace in string

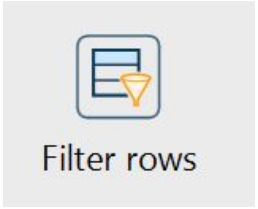
Transform name: Replace Character

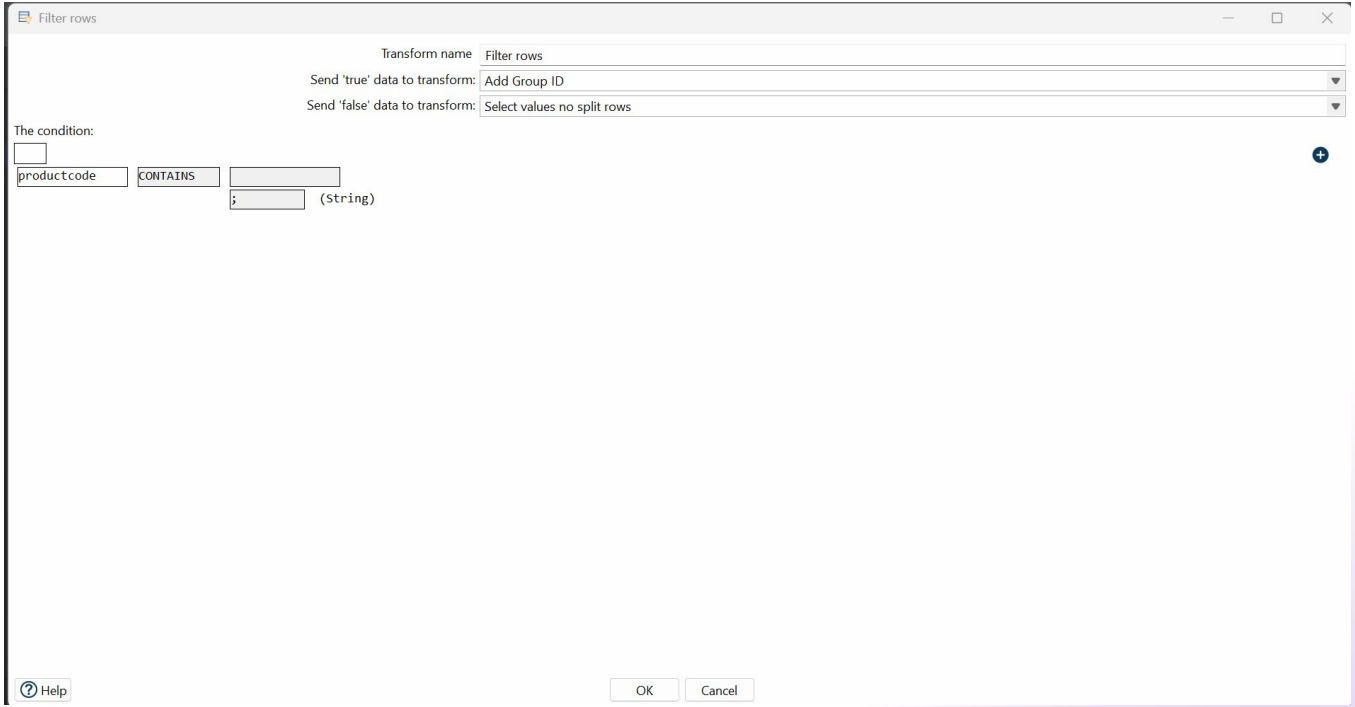
Fields string

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive	Is Unicode
1	productname		N	#		N		N	N	N
2	productname		N	$		N		N	N	N

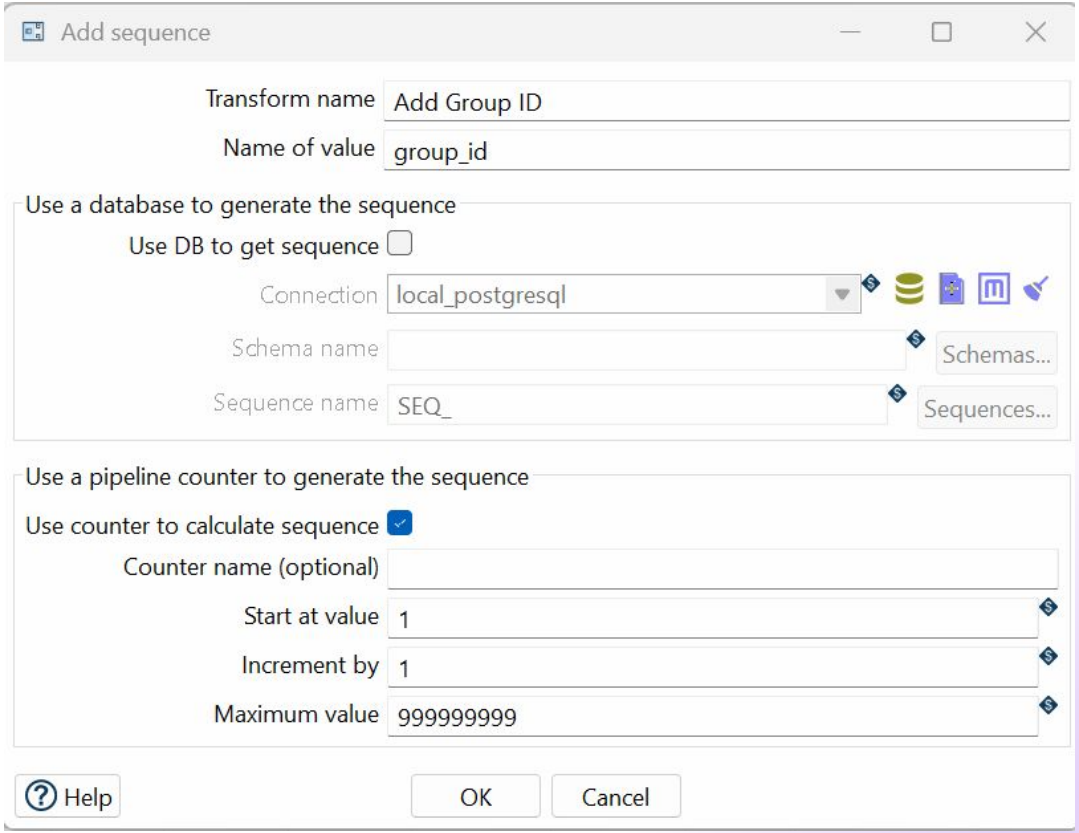
Help | OK | Get fields | Cancel

Transform (Filter rows)

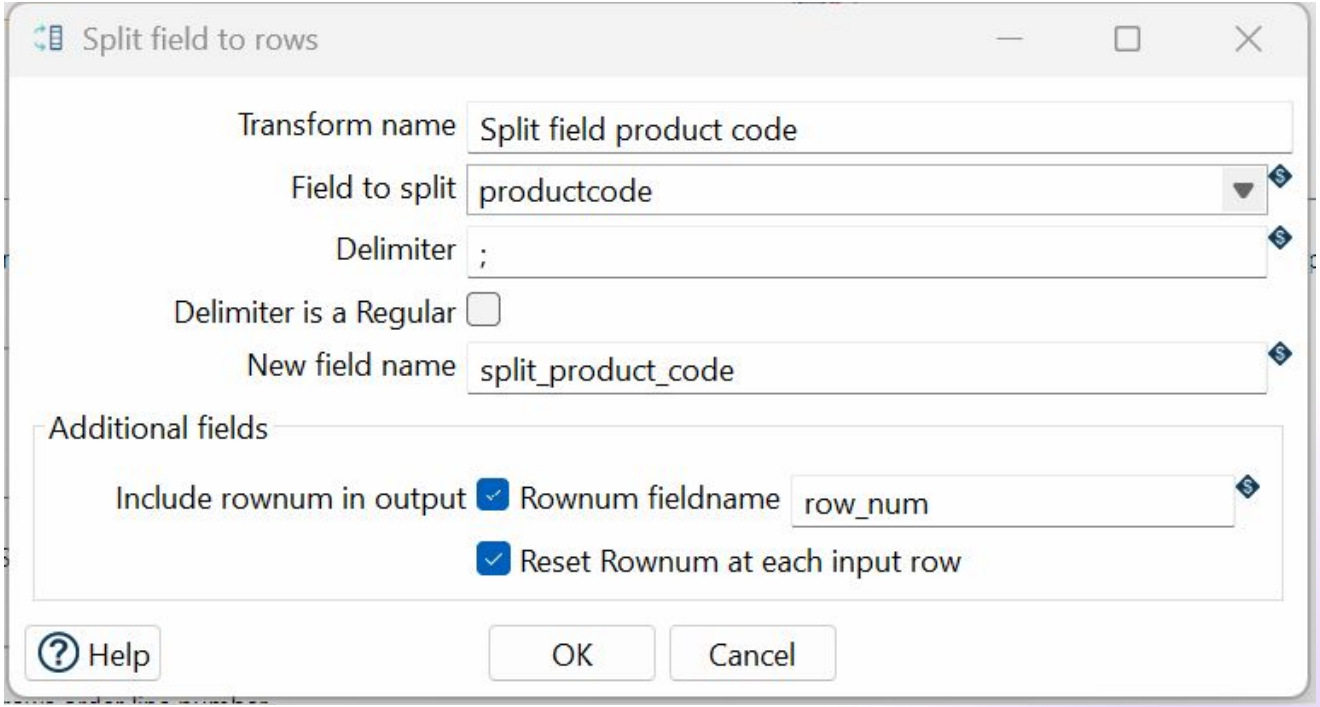




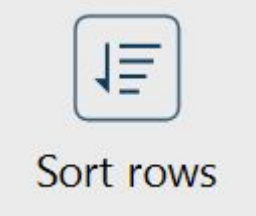
Transform (Add sequence)

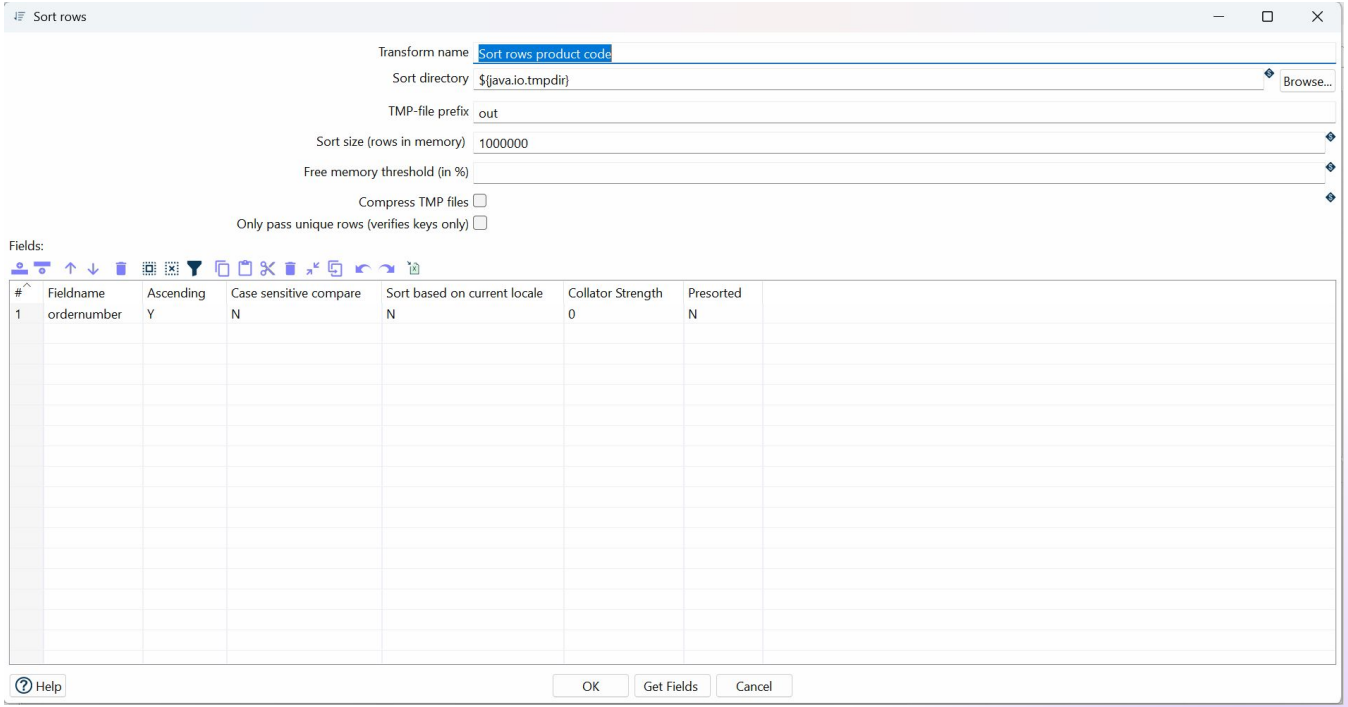


Transform (Split Field to Rows)

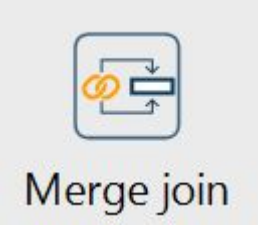


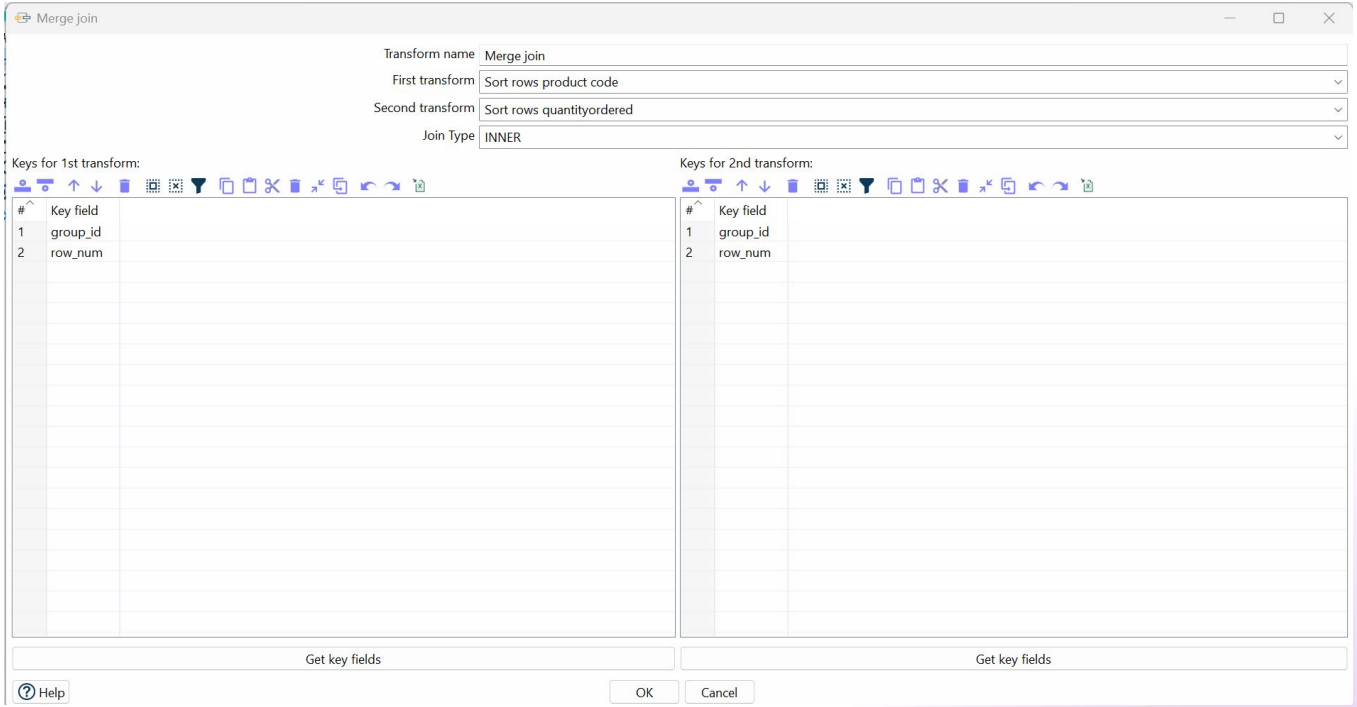
Transform (Sort Rows)



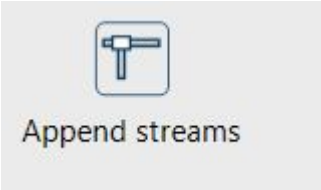


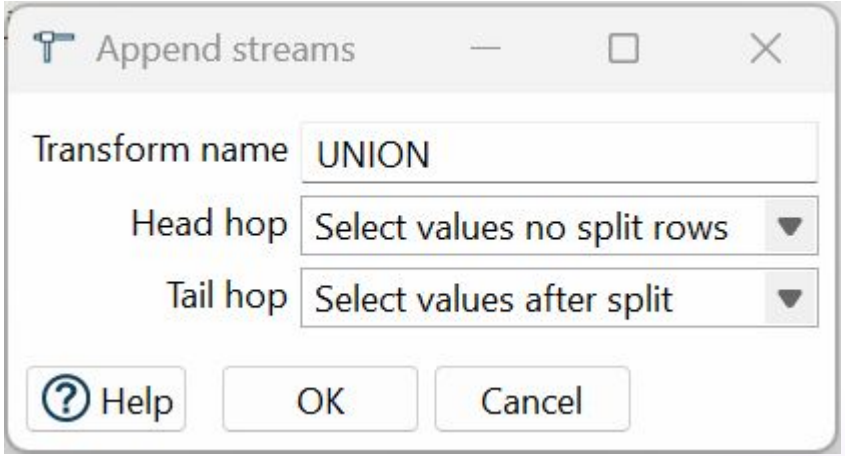
Transform (Merge join)





Transform (Append Stream)





Transform (Formula)

Formula

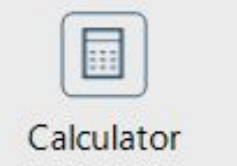
Formula name: Currency Conditional

Fields:

#	New field	Formula	Value type	Length	Precision	Replace value	Set Null to #N/A
1	order_price_each_update	IF([currency]="USD", [order_price_each]*1, IF([currency]="IDR", [order_price_each]/16500, NA()))	Number				N

Help OK Cancel

Transform (Calculator)



Calculator

Transform name: calculate order amount

Throw an error on non existing files

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	Conversion mask	Decimal symbol	Grouping symbol	Currency syr
1	order_amount	A * B	order_price_each_update	order_quantity		Number			N				

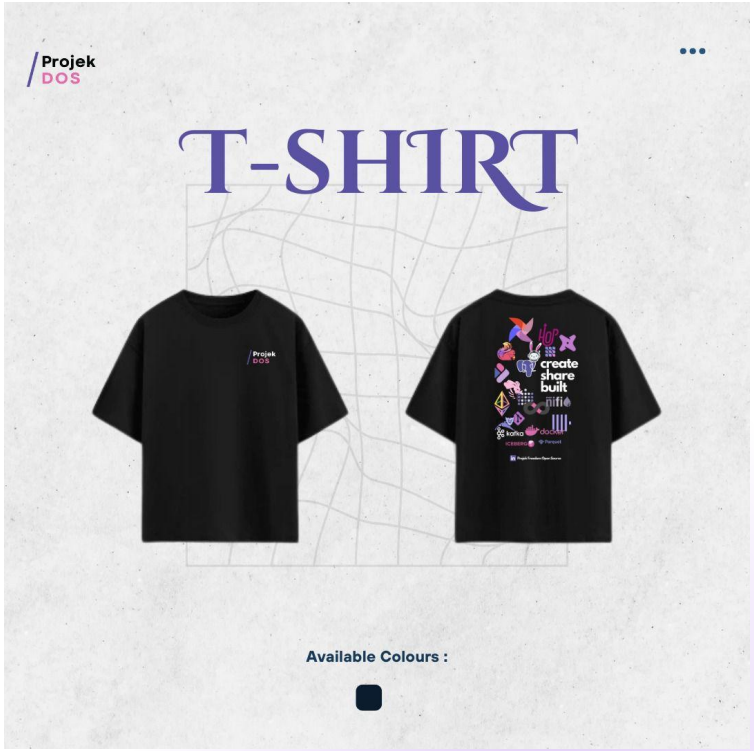
Help | OK | Cancel

Demo

Quiz

- URL: https://github.com/projekdos
- Repository: etl_workshop_apache_hop_batch1
- Read the Instruction
- Solve the Problem

Please submit the result to **info@projekdos.com.** The winner will get Free Merchandise from us!



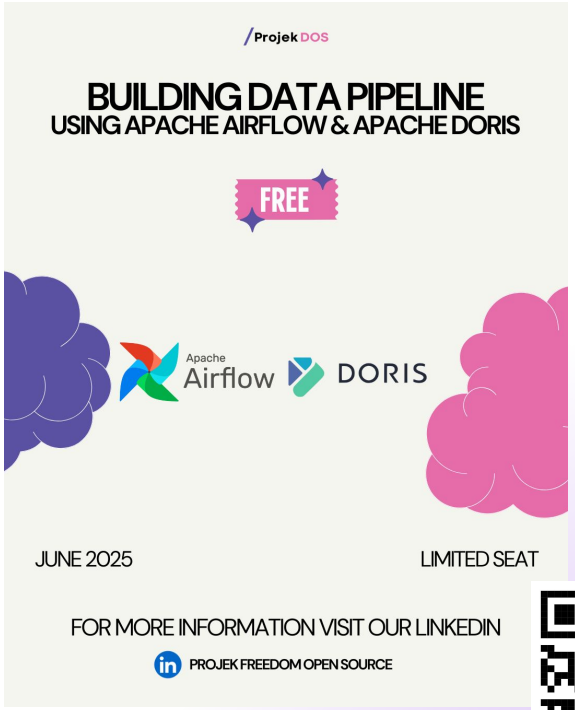
What’s next

Building Data Pipeline using Apache Airflow & Apache Doris – FREE (June 2025)

Understand how data orchestrators like Apache Airflow manage workflows from data sources to your target systems. Perfect for beginners and aspiring data engineers.

What You’ll Learn:

- What is data orchestration and why it matters
- Introduction to Apache Airflow & DAG concepts
- Hands-on: build and schedule your first data pipeline
- Simple integration using Python, SQL & Apache Doris



Thanks!

Business Contact:

info@projekdos.com

+6281385368844 (Whatsapp)

Whatsapp Community

LinkedIn

Youtube Channel

ProjekDOS