

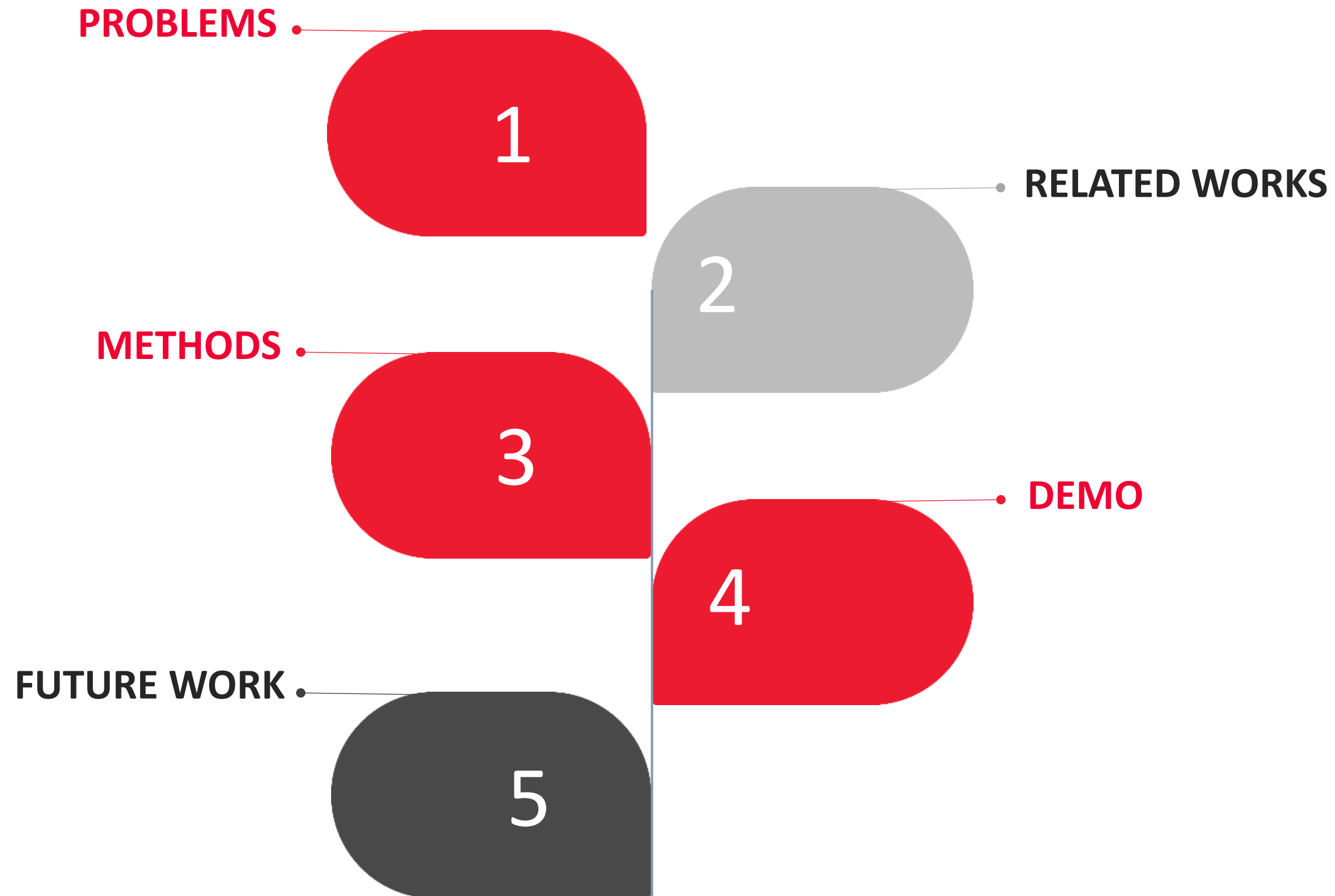
A woman with long dark hair, wearing a bright red raincoat, is smiling and looking to her right while holding a transparent umbrella. She is standing in the rain, with water droplets visible in the air. The background is a soft-focus view of trees and a building under a bright sky.

VIETTEL DIGITAL TALENT PHASE 1 PROJECT

EXCEL CHATBOX

22 – Phan Trọng Đài  
Mentor: Phạm Nguyễn Phương  
VTIT

# Table of Contents

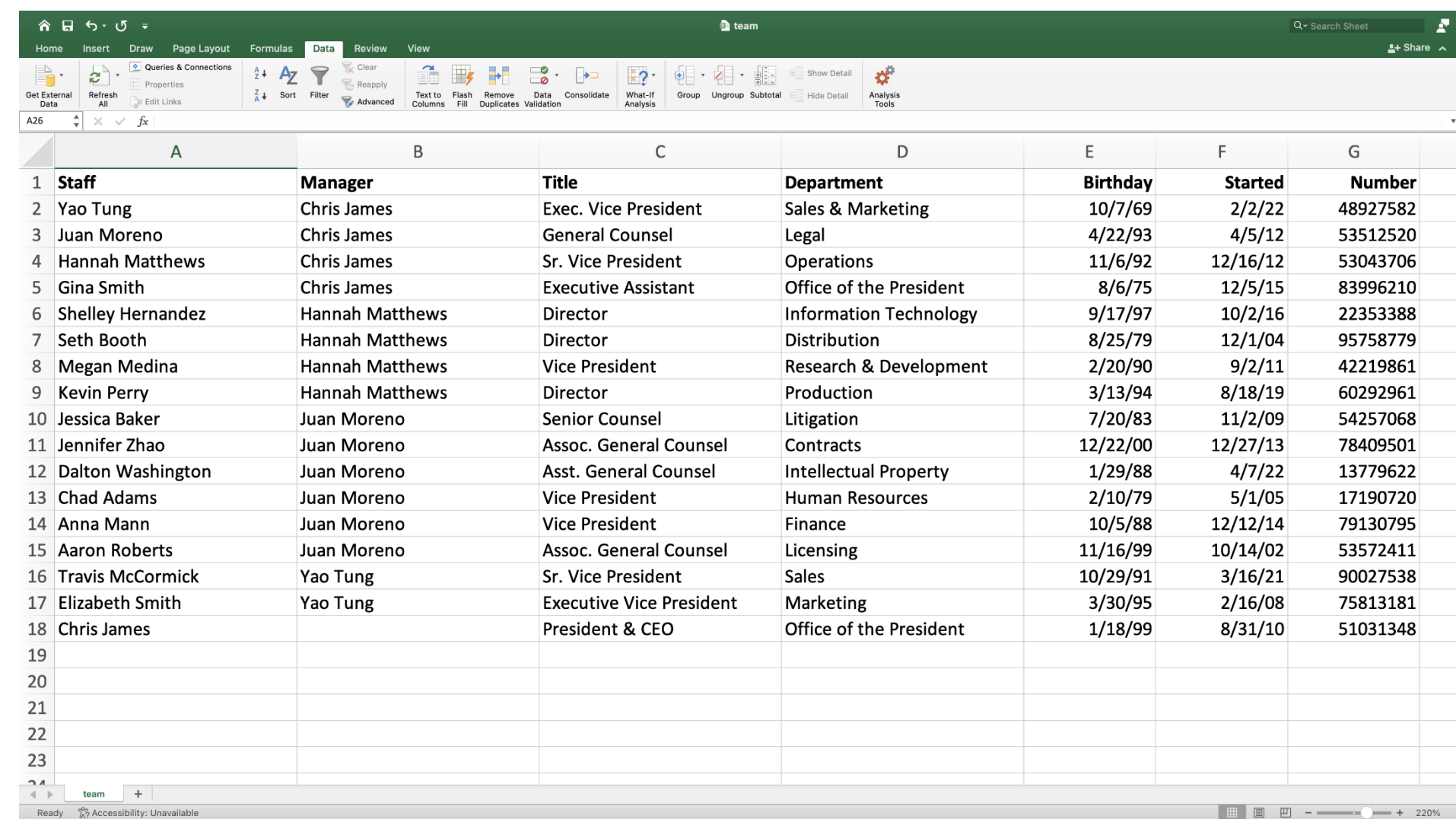






PROBLEMS

# Problem – Various structures



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
1	Staff	Manager	Title	Department	Birthday	Started	Number
2	Yao Tung	Chris James	Exec. Vice President	Sales & Marketing	10/7/69	2/2/22	48927582
3	Juan Moreno	Chris James	General Counsel	Legal	4/22/93	4/5/12	53512520
4	Hannah Matthews	Chris James	Sr. Vice President	Operations	11/6/92	12/16/12	53043706
5	Gina Smith	Chris James	Executive Assistant	Office of the President	8/6/75	12/5/15	83996210
6	Shelley Hernandez	Hannah Matthews	Director	Information Technology	9/17/97	10/2/16	22353388
7	Seth Booth	Hannah Matthews	Director	Distribution	8/25/79	12/1/04	95758779
8	Megan Medina	Hannah Matthews	Vice President	Research & Development	2/20/90	9/2/11	42219861
9	Kevin Perry	Hannah Matthews	Director	Production	3/13/94	8/18/19	60292961
10	Jessica Baker	Juan Moreno	Senior Counsel	Litigation	7/20/83	11/2/09	54257068
11	Jennifer Zhao	Juan Moreno	Assoc. General Counsel	Contracts	12/22/00	12/27/13	78409501
12	Dalton Washington	Juan Moreno	Asst. General Counsel	Intellectual Property	1/29/88	4/7/22	13779622
13	Chad Adams	Juan Moreno	Vice President	Human Resources	2/10/79	5/1/05	17190720
14	Anna Mann	Juan Moreno	Vice President	Finance	10/5/88	12/12/14	79130795
15	Aaron Roberts	Juan Moreno	Assoc. General Counsel	Licensing	11/16/99	10/14/02	53572411
16	Travis McCormick	Yao Tung	Sr. Vice President	Sales	10/29/91	3/16/21	90027538
17	Elizabeth Smith	Yao Tung	Executive Vice President	Marketing	3/30/95	2/16/08	75813181
18	Chris James		President & CEO	Office of the President	1/18/99	8/31/10	51031348
19							
20							
21							
22							
23							

- One value each cell
- Have simple header- the first row
- Can be treated like a CSV file

Simple excel file

# Problem – Various structures

Purchase Orders							
OrderDate	Region	Name	Item	Units	UnitCost	Total	
1/6/2019	East	Jones Andrews	Pencil	95	1.99	189.05	
1/23/2019	Central	Kivell Jardine	Binder	50	19.99	999.5	
2/9/2019	Central	Jardine Jardine	Pencil	36	4.99	179.64	
2/26/2019		Gill Andrews	Pen	27	19.99	539.73	
4/1/2019	East	Jones Sorvino	Binder	60	4.99	299.4	
4/18/2019	Central	Andrews Gill	Pencil	75	1.99	149.25	
5/5/2019	Central	Jardine Sorvino	Pencil	90	4.99	449.1	
5/22/2019	West	Thompson Kivell	Pencil	32	1.99	63.68	
6/8/2019	East	Jones Morgan	Binder	60	8.99	539.4	
6/25/2019	Central	Morgan Jones	Pencil	90	4.99	449.1	
7/12/2019	East	Howard Kivell	Binder	29	1.99	57.71	
7/29/2019	East	Parent Gill	Binder	81	19.99	1,619.19	
8/15/2019	East	Jones Gill	Pencil	35	4.99	174.65	

- Complex header
- Need to deal with merge cells

*Merged header excel file*

# Problem – Various structures

	A	B	C
1	Họ và tên	Lớp	Trường
2	ABC	1	A
3	XYZ	2	B
4	DEF	3	A
5	IJK	4	A
6	MNP	5	B
7	UVX	6	A

*Flattened table*

	A	B	C	D	E	F	G	H
1	Cà phê	Có đá	Giá 2023			Giá 2024		
2	Cà phê thường	Không	100k	100k	100k	100k	100k	100k
3	Cà phê Đen	Không	200k	200k	200k	200k	200k	200k
4	Cà phê Đen	Có đá	300k	300k	300k	300k	300k	300k
5	Cà phê Sữa	Không	400k	400k	400k	400k	400k	400k
6	Cà phê Sữa	Có đá	500k	500k	500k	500k	500k	500k
7	Cà phê muối	Không	600k	600k	600k	600k	600k	600k

*Matrix table*

- Semantic meaning -> different query type
- Matrix Table much more complex and flexible
- Matrix Table used more in real life (in some fields)



# Problem – Various structures

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	T	Đối tác	Nguồn	Năm 2022												
2	T	Đối tác	Nguồn	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tổng
3	1	AL TEK														
4	1.1		Tự SX	100	150	200	250	300	350	400	450	500	550	600	650	700
5	1.2		OS	200	250	300	350	400	450	500	550	600	650	700	750	800
6	1.3		LK	300	350	400	450	500	550	600	650	700	750	800	850	900
7	1.4		Dịch vụ	400	450	500	550	600	650	700	750	800	850	900	950	1000
8	1.5		Phí QL	500	550	600	650	700	750	800	850	900	950	1000	1050	1100
9	1.6		Thương mại	600	650	700	750	800	850	900	950	1000	1050	1100	1150	1200
10	1.7		Khác	700	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300
11	2	ANT Group														
12	2.1		Tự SX	900	950	1000	1050	1100	1150	1200	1250	1300	1350	1400	1450	1500
13	2.2		OS	1000	1050	1100	1150	1200	1250	1300	1350	1400	1450	1500	1550	1600
14	2.3		LK	1100	1150	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700
15	2.4		Dịch vụ	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750	1800
16	2.5		Phí QL	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850	1900
17	2.6		Thương mại	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850	1900	1950	2000
18	2.7		Khác	1500	1550	1600	1650	1700	1750	1800	1850	1900	1950	2000	2050	2100
19	3	B&C														
20	3.1		Tự SX	1700	1750	1800	1850	1900	1950	2000	2050	2100	2150	2200	2250	2300
21	3.2		OS	1800	1850	1900	1950	2000	2050	2100	2150	2200	2250	2300	2350	2400
22	3.3		LK	1900	1950	2000	2050	2100	2150	2200	2250	2300	2350	2400	2450	2500
23	3.4		Dịch vụ	2000	2050	2100	2150	2200	2250	2300	2350	2400	2450	2500	2550	2600
24	3.5		Phí QL	2100	2150	2200	2250	2300	2350	2400	2450	2500	2550	2600	2650	2700
25	3.6		Thương mại	2200	2250	2300	2350	2400	2450	2500	2550	2600	2650	2700	2750	2800
26	3.7		Khác	2300	2350	2400	2450	2500	2550	2600	2650	2700	2750	2800	2850	2900

- Matrix table
- Merge cells

Our data

# Problem – Various structures

Chỉ tiêu	Số liệu	Phân loại	Loại hình
DOANH THU		Kế hoạch (Tập đoàn)	Sản xuất Dịch vụ Thương mại Khác
		Thực hiện	Sản xuất Dịch vụ Thương mại Khác
		% Hoàn thành	Sản xuất Dịch vụ Thương mại Khác
Lợi nhuận gộp	Tổng lợi nhuận gộp	Kế hoạch (Tập đoàn)	Sản xuất Dịch vụ Thương mại Khác
		Thực hiện	Sản xuất Dịch vụ Thương mại Khác
		% Hoàn thành	Sản xuất Dịch vụ Thương mại Khác

F	G	H	I	J	K	L	M	N	O
Năm 2024									
6 Tháng đầu năm									
Quý I				Quý II				Cộng	
Tháng 01	Tháng 02	Tháng 03	Cộng	Tháng 04	Tháng 05	Tháng 06	Cộng	Cộng	Tháng

TT	Chỉ tiêu	Số liệu	Phân loại	Loại hình
11.3			% Hoàn thành	
12	Nợ phải trả/ Vốn chủ sở hữu		Kế hoạch (Tập đoàn)	
12.1			Thực hiện	
12.2			% Hoàn thành	
12.3				
13	Lô lũy kế			
14	Nộp ngân sách Nhà nước			
14.1		Tổng số phải nộp	Kế hoạch (Tập đoàn)	
14.1.1			Thực hiện	
14.1.2			% Hoàn thành	
14.1.3				
14.2		Số nộp tại Thanh Hóa		
15	Tổng tài sản			
16	Vốn chủ sở hữu			
17	Vốn vay			

- Undefined field
- Complex hierarchical header
- Flexible field



# Problem – Security

A	B	C	D	E	F	G	H	I	J	
TT	Đối tác	Nguồn	Năm 2022							
			Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	Tháng	
1	ALTEK									
1.1		Tư SX	100	150	200	250	300	350	400	4
1.2		OS	200	250	300	350	400	450	500	5
1.3		LK	300	350	400	450	500	550	600	6
1.4		Dịch vụ	400	450	500	550	600	650	700	7
1.5		Phí QL	500	550	600	650	700	750	800	8
1.6		Thương mại	600	650	700	750	800	850	900	9
1.7		Khác	700	750	800	850	900	950	1000	1

- Sensitive internal information!!!

# Problem – User query

	A	B	C	D	E	F	G	H
1	Cà phê	Có đá	Giá 2023			Giá 2024		
2	Cà phê thường	Không	100k	100k	100k	100k	100k	100k
3	Cà phê Đen	Không	200k	200k	200k	200k	200k	200k
4	Cà phê Đen	Có đá	300k	300k	300k	300k	300k	300k
5	Cà phê Sữa	Không	400k	400k	400k	400k	400k	400k
6	Cà phê Sữa	Có đá	500k	500k	500k	500k	500k	500k
7	Cà phê muối	Không	600k	600k	600k	600k	600k	600k

- Năm 2024 giá của cà phê đen là bao nhiêu?

Cà phê	Có đá	Giá 2023			Giá 2024		
Cà phê thường	Không	100k	100k	100k	100k	100k	100k
Cà phê Đen	Không	200k	200k	200k	200k	200k	200k
Cà phê Đen	Có đá	300k	300k	300k	300k	300k	300k
Cà phê Sữa	Không	400k	400k	400k	400k	400k	400k
Cà phê Sữa	Có đá	500k	500k	500k	500k	500k	500k
Cà phê muối	Không	600k	600k	600k	600k	600k	600k

- Loại cà phê mặn mặn hơn bình thường 1 tí giá bao nhiêu vào 2023

	A	B	C	D	E	F	G	H
1	Cà phê	Có đá	Giá 2023			Giá 2024		
2	Cà phê thường	Không	100k	100k	100k	100k	100k	100k
3	Cà phê Đen	Không	200k	200k	200k	200k	200k	200k
4	Cà phê Đen	Có đá	300k	300k	300k	300k	300k	300k
5	Cà phê Sữa	Không	400k	400k	400k	400k	400k	400k
6	Cà phê Sữa	Có đá	500k	500k	500k	500k	500k	500k
7	Cà phê muối	Không	600k	600k	600k	600k	600k	600k

- Cà phê có đá giá bao nhiêu

Cà phê	Có đá	Giá 2023			Giá 2024		
Cà phê thường	Không	100k	100k	100k	100k	100k	100k
Cà phê Đen	Không	200k	200k	200k	200k	200k	200k
Cà phê Đen	Có đá	300k	300k	300k	300k	300k	300k
Cà phê Sữa	Không	400k	400k	400k	400k	400k	400k
Cà phê Sữa	Có đá	500k	500k	500k	500k	500k	500k
Cà phê muối	Không	600k	600k	600k	600k	600k	600k

- Cfm có giá bn vào 2023

# Problem – Multi files

		Cà phê	Có đá	Giá 2023			
	A	B	C	D	E		00k
1	Cà phê	Có đá	Giá 2024				00k
	A	B	C	D	E	100k	00k
1	Cà phê	Có đá	Giá 2025			200k	00k
2	Cà phê thường	Không	100k	100k	100k	300k	00k
3	Cà phê Đen	Không	200k	200k	200k	400k	00k
4	Cà phê Đen	Có đá	300k	300k	300k	500k	00k
5	Cà phê Sữa	Không	400k	400k	400k	600k	
6	Cà phê Sữa	Có đá	500k	500k	500k		
7	Cà phê muối	Không	600k	600k	600k		

- Cà phê đen đá năm 2025 có giá bao nhiêu?
- Cà phê đen đá năm 2025 và 2024 có giá như thế nào?
- Cho tôi giá cà phê đen 2024 và cà phê sữa 2023.





Related work

# Related work

## Rule based

- Pandas + OpenPyxl
- Problems with matrix table and merge cells
- Need to define structure first
- Problems with natural language query

## DL Approach

- Mainly focus simple excel files
- Do not focus on handling query

# Related work

## NL2SQL

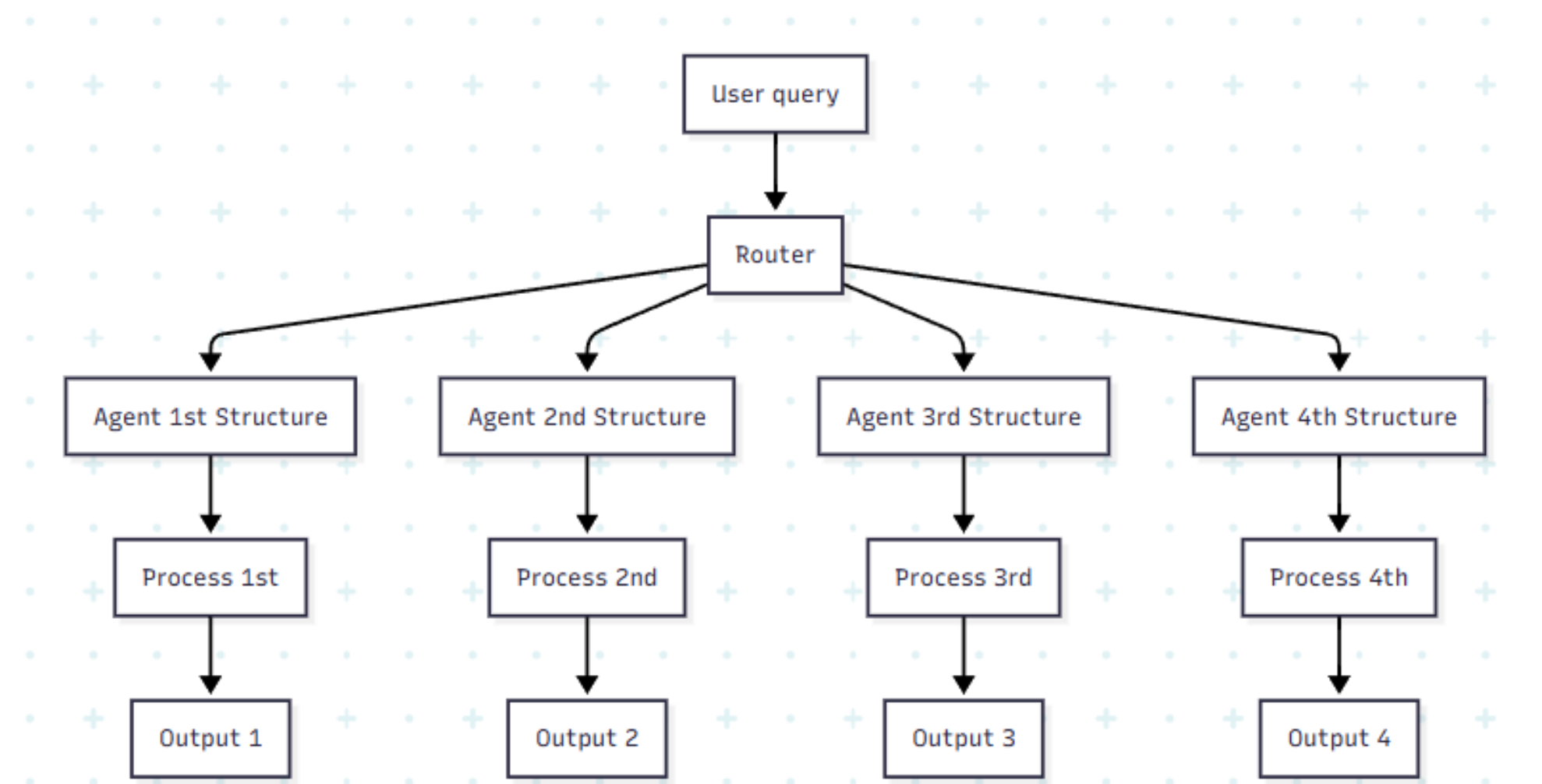
- Strong in handle query
- Problem with converting excel (matrix table) into SQL-friendly data

## LLM Approach

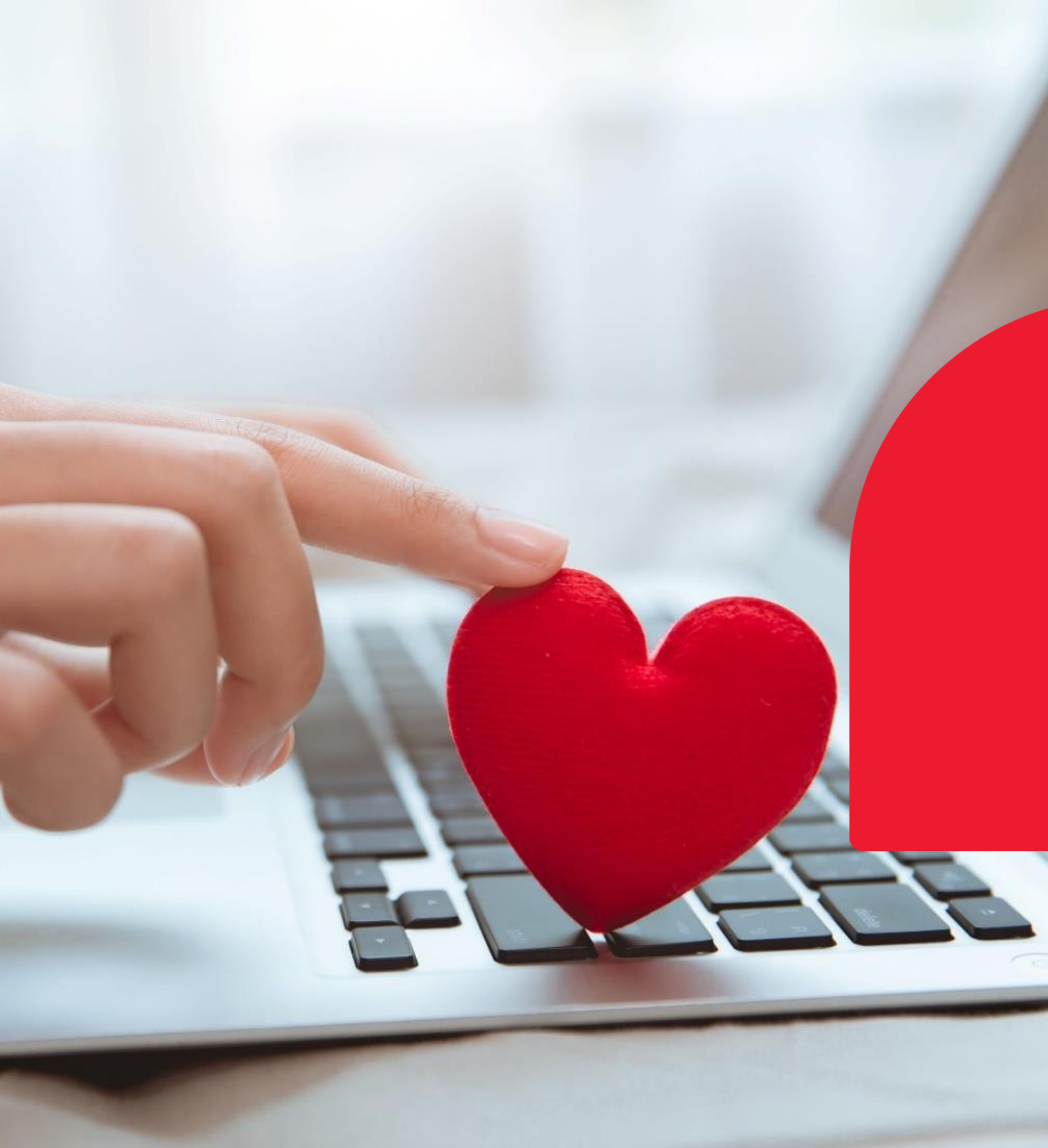
- Sheetcopilot, Spreadsheetcoder, InstructExcel
  - Do not handle matrix table
  - Security concern
  - Reliability concern
- Llama-index, Llama-parse (AI api)
  - Security concern
  - Reliability concern (Have a hard time for AI to interpret complex structure)



# Related work – Agents for specific structure

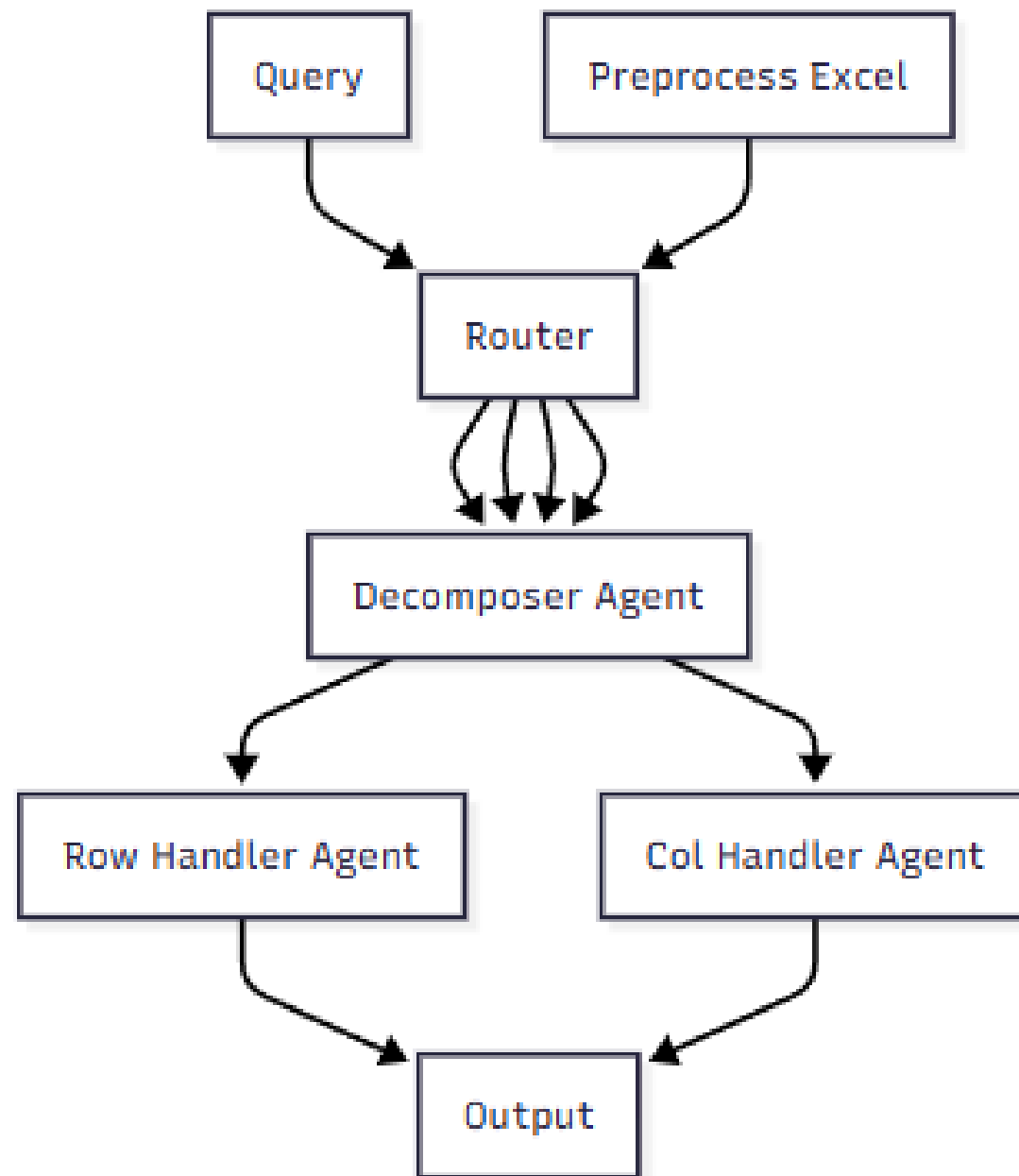


- Flexible
- Scalability ~ New template agents for new file structures



# Method

# Method – Main Architecture



- One single multi-agent flow for different structure
- Based on LLM and inductive bias of Matrix table (rule based) to unify multi separated agents into one!
- Much more flexible [No need to change code – code independency]



# Method – Preprocess

FEATURE ROW		FEATURE COL	
FEATURE ROW VALUE	Thành Phố	Phân loại	Giới Tính
			Nam Nữ
	TPHCM		
		Cao	
		TB	
	Hà nội		
		Cao	
	Cao bằng		
		TB	
	Hà Giang		
		TB	
		Cao	
	Vũng Tàu		
		TB	

- Identify header based on rule based
- Classify feature row and feature col based on LLM (need semantic and structure understanding!)

```
### Content
Thành Phố,          Phân loại,          Giới tính,Giới tính
Unnamed: 0_level_1,Unnamed: 1_level_1,Nam      ,Nữ

### Name of Feature Rows
Thành Phố,Phân loại,Quận,Phường

### Name of Feature Cols
Giới tính
```

# Method – Preprocess

```
### Feature Rows
['Cà phê', 'Loại', 'Nhập Khẩu ']  
  
### Row Hierarchy  
cà phê: cà phê thường  
    loại: loại 1  
        nhập khẩu : việt nam, brazil, mỹ  
    loại: loại 2  
        nhập khẩu : việt nam, mỹ  
cà phê: cà phê đen  
    loại: loại 2  
        nhập khẩu : việt nam
```

```
### Feature Column  
['thời gian', 'thu nhập']  
  
### Column Hierarchy  
level_1: thời gian  
    level_2: hè, đông  
level_1: thu nhập  
    level_2: thấp, trung bình, cao
```

- Use DFS to continue to extract other schema information
- Note: Try many ways to represent – above is my best for LLM understanding

# Method – Decomposer

```
### Query
cho tôi biết cà phê đen tại việt nam có giá như thế nào vào những tháng hè

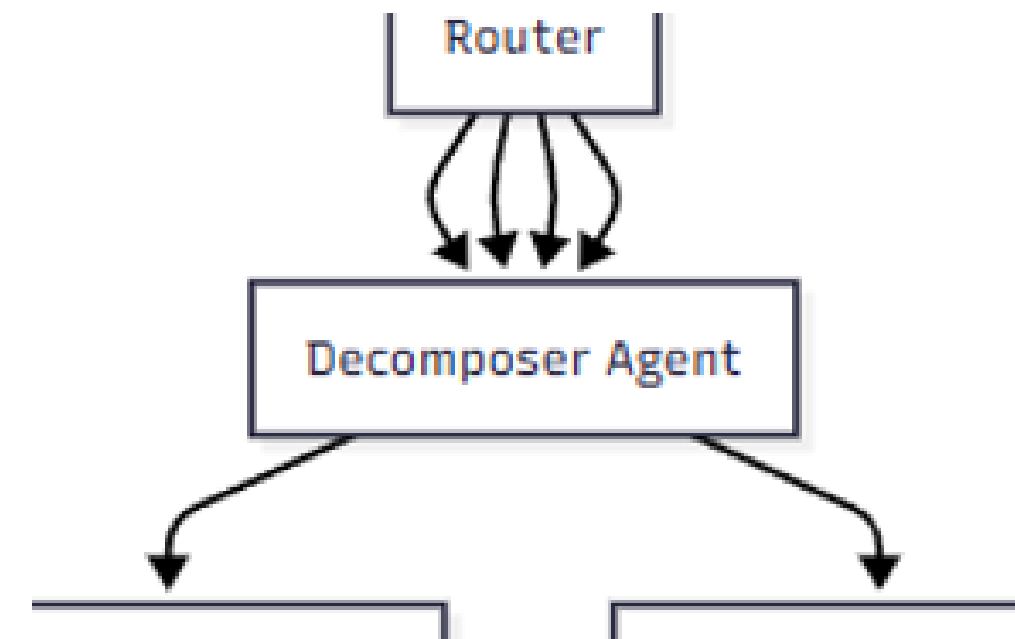
### Row Hierarchy
cà phê: cà phê thường
    loại: loại 1
        nhập khẩu : việt nam, brazil, mỹ
    loại: loại 2
        nhập khẩu : việt nam, mỹ
cà phê: cà phê đen
    loại: loại 2
        nhập khẩu : việt nam

### Column Hierarchy
level_1: thời gian
    level_2: hè, đông
level_1: thu nhập
    level_2: thấp, trung bình, cao

### Row Keywords
- cà phê đen
- việt nam

### Col Keywords
- tháng hè
```

- LLM based
- Input:
  - User query
  - Row Hierarchy
  - Column Hierarch
- Output: Row and Col keywords





# Method – Row Handler

```
### Query
cho tôi biết cà phê đen tại việt nam có giá như thế nào vào những tháng hè

### Row Hierarchy
cà phê: cà phê thường
    loại: loại 1
        nhập khẩu : việt nam, brazil, mỹ
    loại: loại 2
        nhập khẩu : việt nam, mỹ
cà phê: cà phê đen
    loại: loại 2
        nhập khẩu : việt nam

### Row Keywords
['cà phê đen', 'việt nam']

### Row Identifier
cà phê: cà phê đen
    loại: Undefined
        nhập khẩu : việt nam
```

- LLM based
- Input:
  - Query
  - Row hierarchy
  - Row Keywords
- Output: Row Identifier



# Method – Col Handler

```
### Query
cho tôi biết cà phê đen tại việt nam có giá như thế nào vào những tháng hè

### Column Hierarchy
level_1: thời gian
    level_2: hè, đông
level_1: thu nhập
    level_2: thấp, trung bình, cao

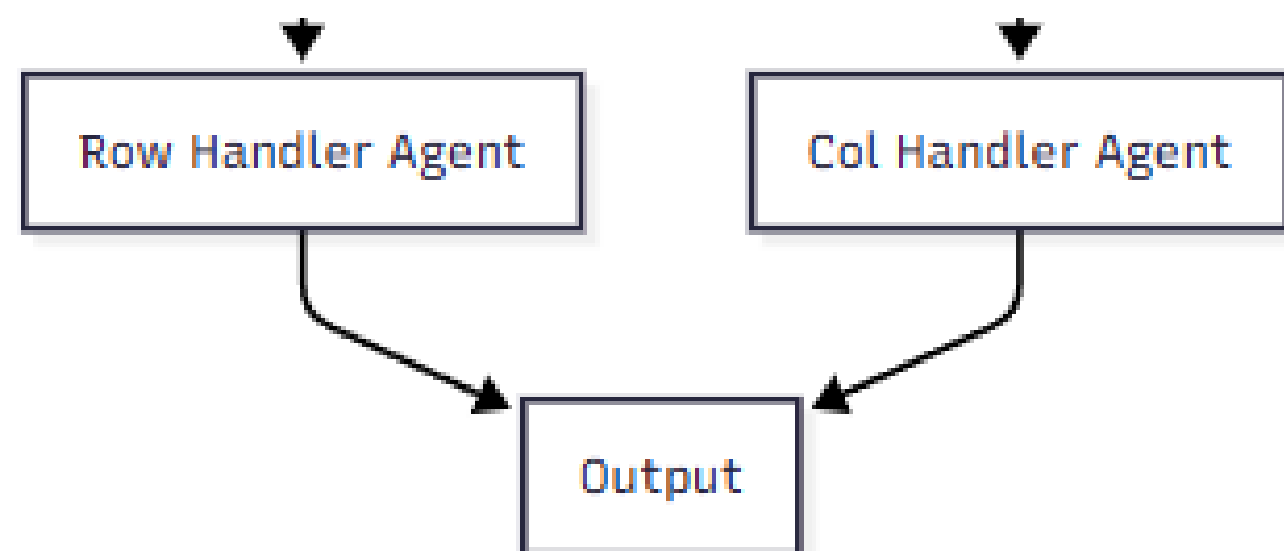
### Col Keywords
['tháng hè']

### Col Identifier
level_1: thời gian
    level_2: hè
```

- LLM based
- Input:
  - Query
  - Col hierarchy
  - Col Keywords
- Output: Col Identifier



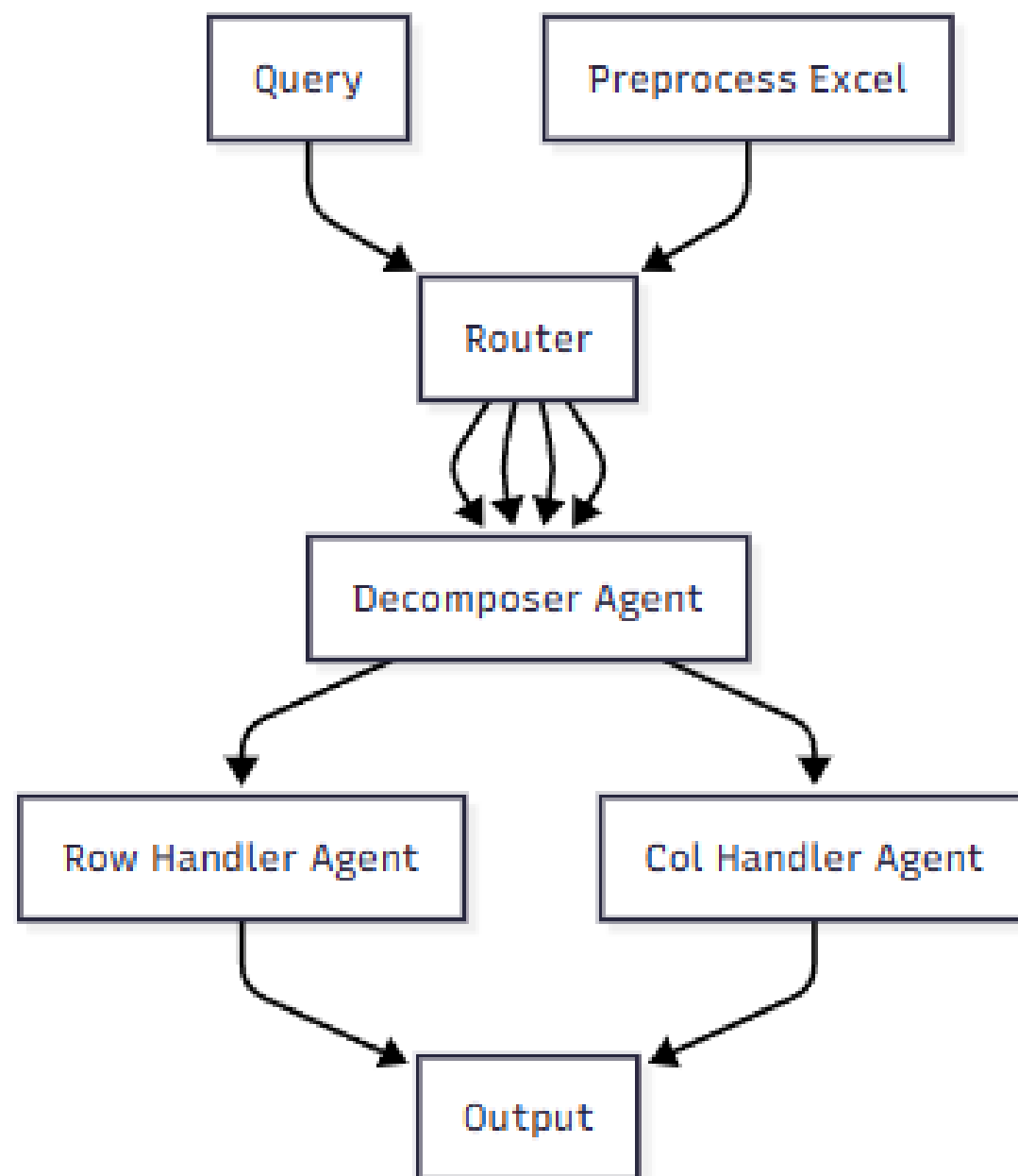
# Method – Output



- Pandas

```
df[feature_row = row_identifier][feature_col]
```

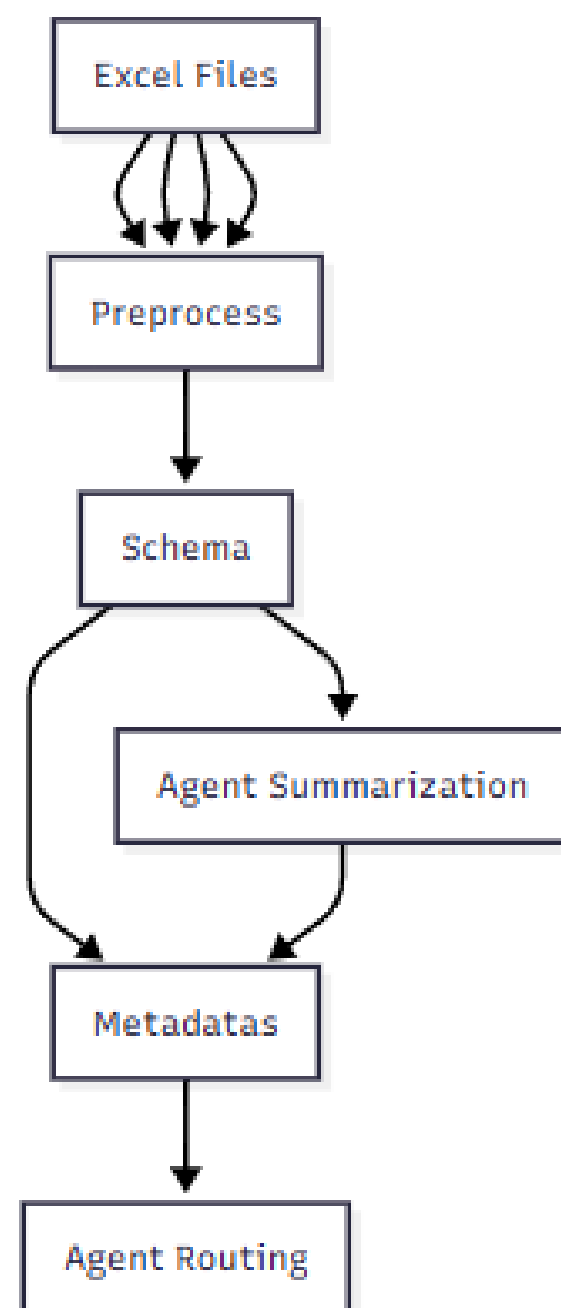
# Method – Security concern



- Only work with Schema also secures privacy for internal data!!!

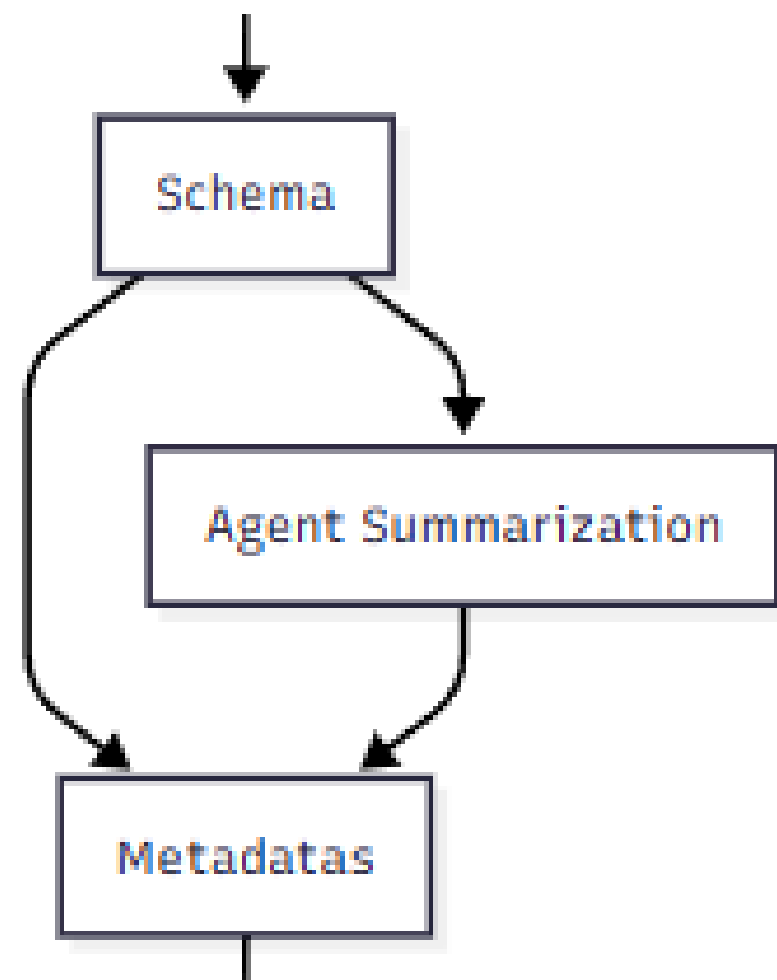


# Method – Multi-file Architecture



- Preprocess and extract schema from excel files
- Schema: Row and Col identifier
- Go through an Agent to summarize the semantic meaning of the files
- Feed metadatas to Agent Routings

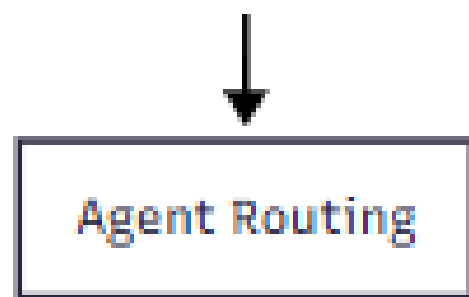
# Method – Agent Summarization



- Preprocess and extract schema from excel files
- Schema: Row and Col identifier
- Go through an Agent to summarize the semantic meaning of the files
- Feed metadatas to Agent Routings

1. Overall Purpose/Domain
2. Time Aspect/Coverage
3. Primary Row Entities & Breakdown
4. Primary Column Metrics & Dimensions
5. Inferred Data Focus

# Method – Agent Routing



```
### User Query
Cho tôi biết có bao nhiêu nữ và nam trong quận 1, và có bao nhiêu nam trong quận 2 phường 14

### Separated Query
example1.xlsx - Cho tôi biết có bao nhiêu nữ và nam trong quận 1
example1.xlsx - Cho tôi biết có bao nhiêu nam trong quận 2 phường 14

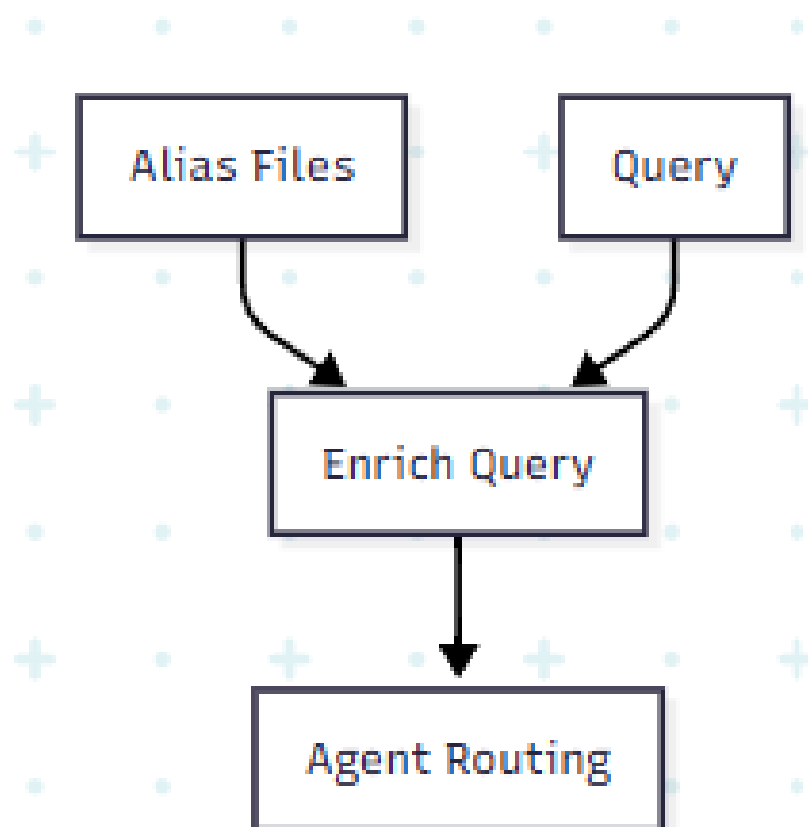
### User Query
Cà phê đen có giá sẽ khoảng bao nhiêu

### Separated Query
example2.xlsx - Cà phê đen có giá sẽ khoảng bao nhiêu
example4.xlsx - Cà phê đen có giá sẽ khoảng bao nhiêu

### User Query
số lượng nam ở hà nội và đà nẵng, số lượng cà phê mỹ nhập khẩu từ brazil và
cà phê loại tốt thường thu hút bao nhiêu người thu nhập cao

### Separated Query
example1.xlsx - số lượng nam ở hà nội và đà nẵng
example3.xlsx - số lượng cà phê mỹ nhập khẩu từ brazil
example3.xlsx - cà phê loại tốt thường thu hút bao nhiêu người thu nhập cao
```

# Method – Alias

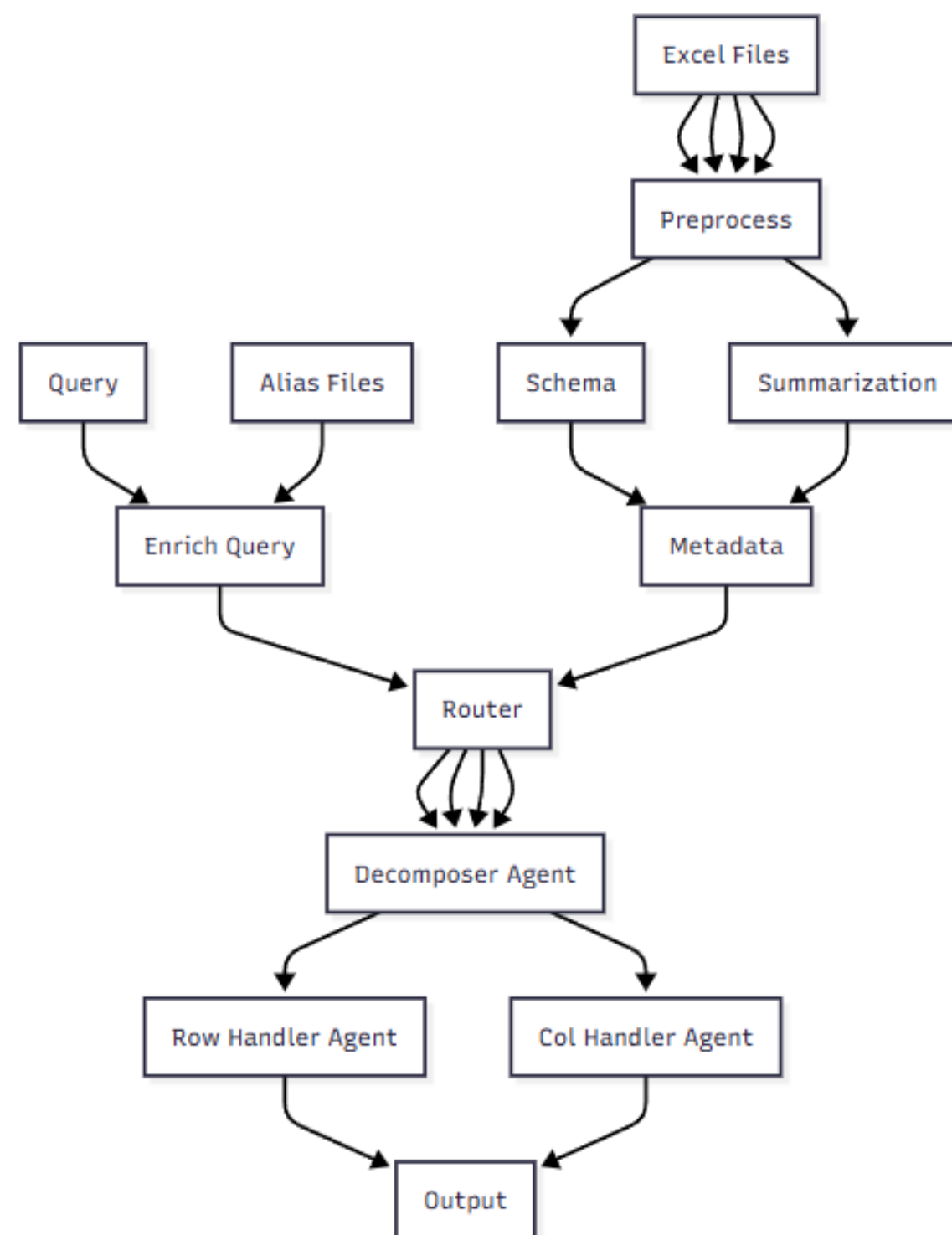


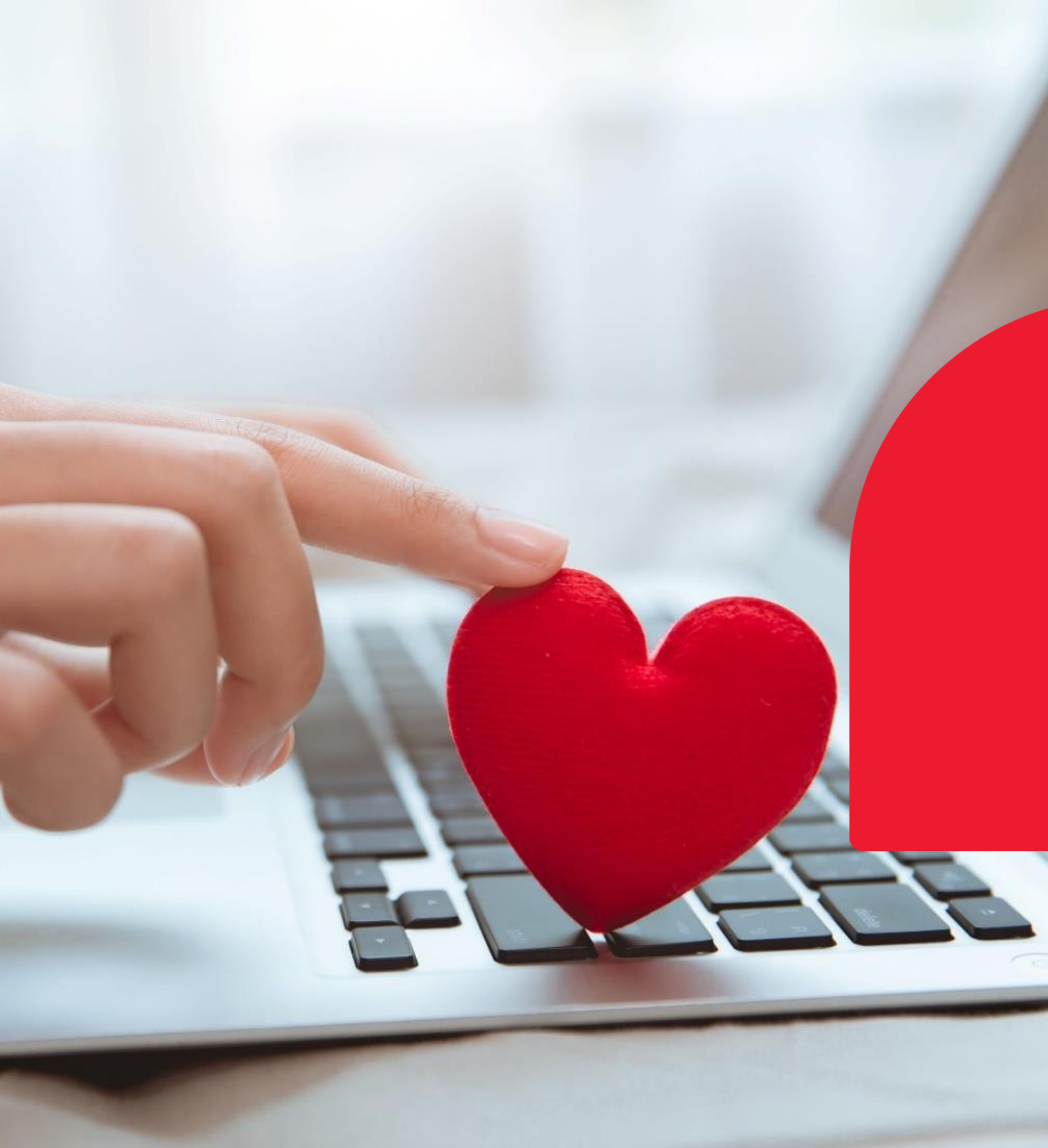
```
### Initial Query
CP cho sản xuất là bao nhiêu năm 2022?
### Enriched Query
CP(Chi phí) cho sản xuất là bao nhiêu năm 2022?
```

```
### Initial Query
So sánh Doanh Thu của SkyIQ và Viettel Timor Leste.
### Enriched Query
So sánh Doanh Thu(DT) của SkyIQ(SKYIQ PTE.LTD) và Viettel Timor Leste(Telemor-VTL).
```



# Method – Final Framework





DEMO

# WEB – Tech Stack

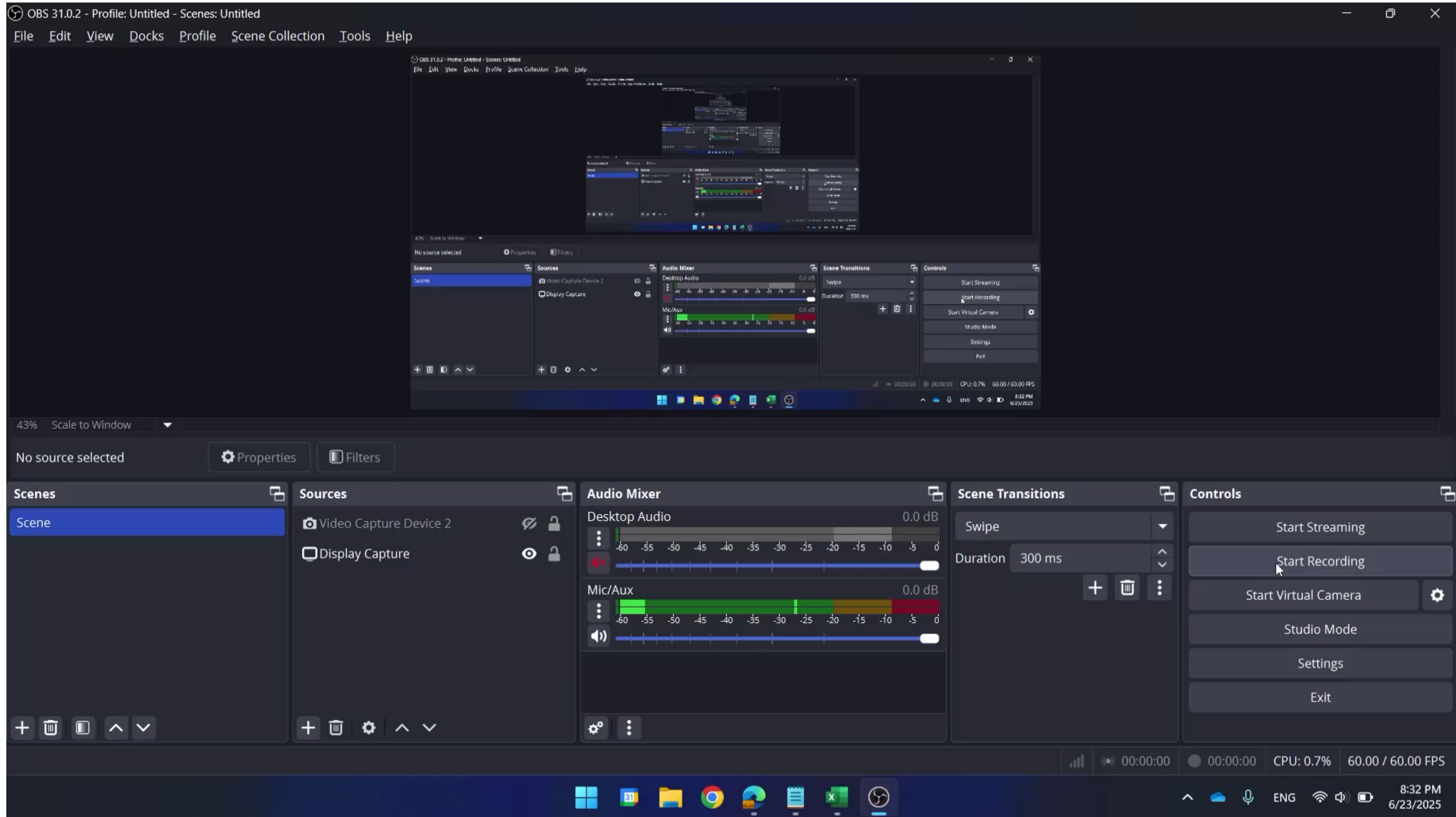
Gemini

 **LangChain**

 **FastAPI**

  
plotly





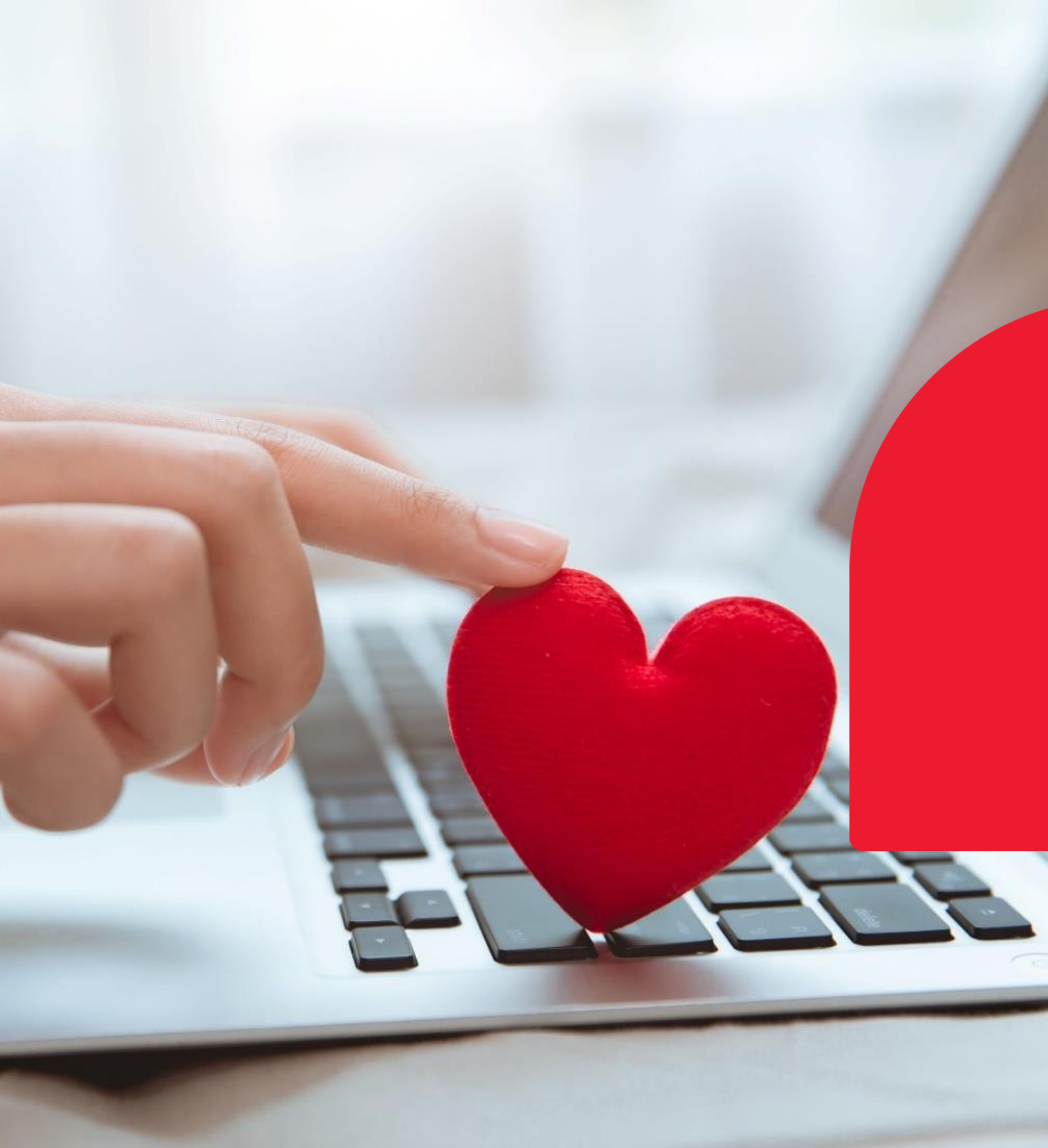


# FUTURE WORK



# Future work

- Add memory within a conversation -> Enhance UX
- Better GUI for handle many files (NotebookLM-like GUI)
- RAG to support even more files (currently under-use summarization features)
- Add login/logout
- Add store conversation into DB (sqlite or mongoDB)
- Docker and Cloud deployment
- Complex query (aggregate, compare, ...)



THE END  
Q&A