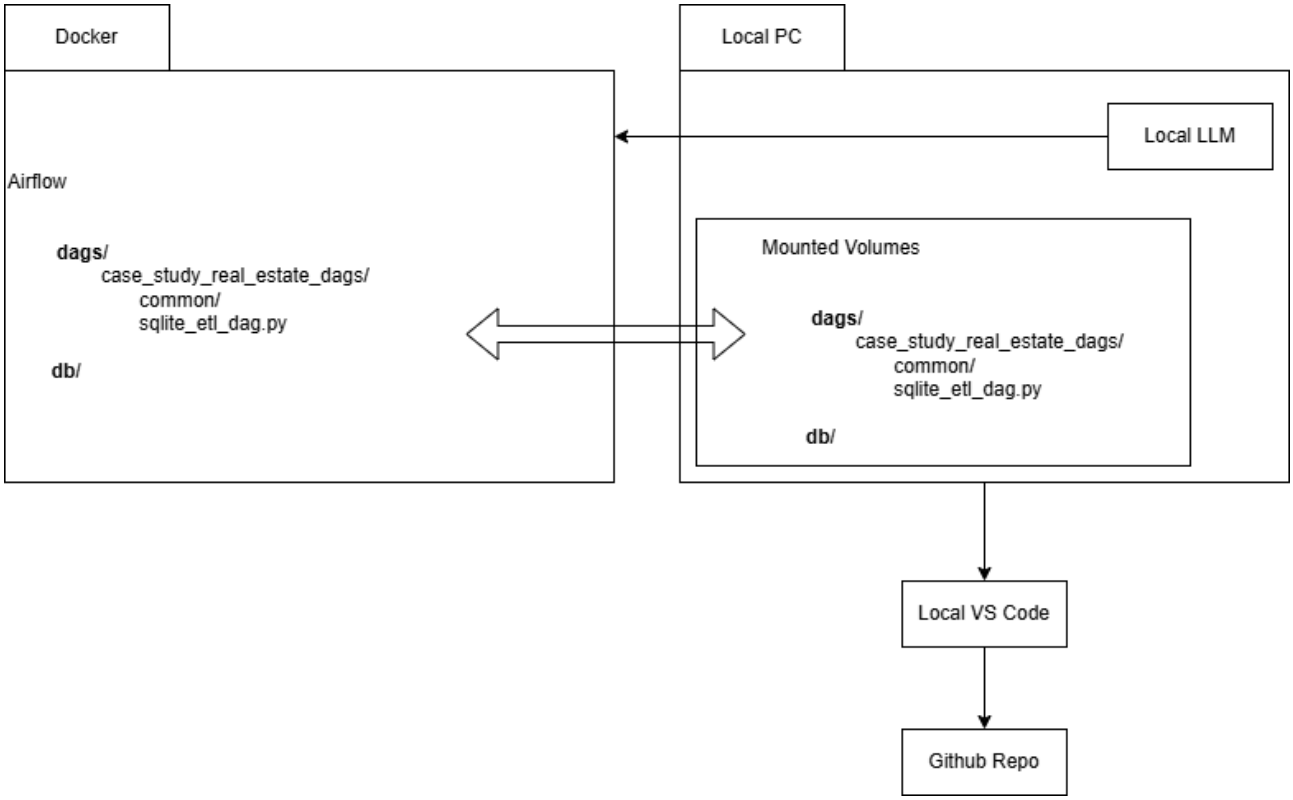
We this report, I’m provinding:

- dag files
- db files
- ERD as svg file
- sql file to cerate the db

Global architecure



EDA

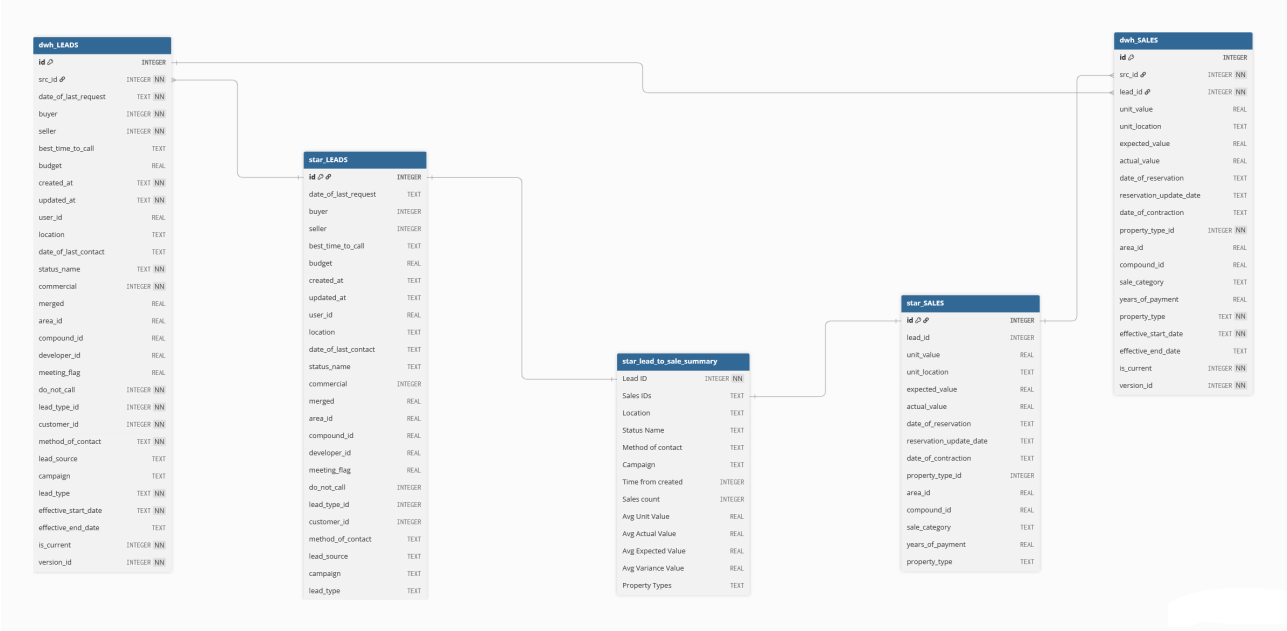
Leads

columns	data type	#missing	%missing	#unique	%unique	min	max	average	standard_devi...	first value	second value	third value
best_time_to_call	object	54396	97.27	957	1.71					All day , any day	None	None
id	Int32	0	0.00	55845	99.86	30.00	413105.00	206891.20	118102.50	9815	90225	198759
date_of_last_request	object	0	0.00	52860	94.52					2018-09-07 00...	2022-05-04 19...	2023-03-22 05...
buyer	object	0	0.00	2	0.00					True	True	True
seller	object	0	0.00	2	0.00					False	False	False
budget	Int32	55531	99.30	84	0.15	1.00	30000000.00	4935223.53	4592318.33	<NA>	<NA>	<NA>
created_at	object	0	0.00	54336	97.16					2018-09-07 15...	2022-05-04 19...	2023-03-22 05...
updated_at	object	0	0.00	26303	47.03					2024-01-24 15...	2024-01-10 13...	2024-02-07 15...
user_id	Int32	4636	8.29	709	1.27	1.00	4141.00	1891.91	1518.70	2942	1494	4043
location	object	7579	13.55	3085	5.52					Coast 82	The Waterwa...	Silversands
date_of_last_contact	object	331	0.59	36826	65.85					2024-01-24 14...	2024-01-10 13...	2024-02-07 15...
status_name	object	0	0.00	48	0.09					Reassigning	Future contact	Not interested
commercial	object	0	0.00	2	0.00					False	False	False
merged	object	44991	80.45	2	0.00					None	None	None
area_id	Int32	53851	96.29	15	0.03	1.00	17.00	4.38	3.84	3	<NA>	<NA>
compound_id	Int32	54120	96.77	124	0.22	3.00	386.00	197.02	94.11	170	<NA>	<NA>
developer_id	Int32	55835	99.84	6	0.01	8.00	87.00	23.97	17.45	<NA>	<NA>	<NA>
meeting_flag	Int32	55392	99.05	3	0.01	0.00	2.00	1.57	0.50	<NA>	<NA>	<NA>
do_not_call	object	0	0.00	2	0.00					False	False	False
lead_type_id	Int32	0	0.00	10	0.02	1.00	134.00	1.54	5.72	1	1	1
customer_id	Int32	0	0.00	55761	99.71	6.00	400150.00	200172.00	114419.26	22785	172065	192904
method_of_contact	object	0	0.00	77	0.14					form adwords	generic form	whatsapp
lead_source	object	250	0.45	95	0.17					google	google	google website
campaign	object	16368	29.27	1481	2.65					google_cpc	dev_-_equity	None
lead_type	object	0	0.00	10	0.02					Primary	Primary	Primary

Sales

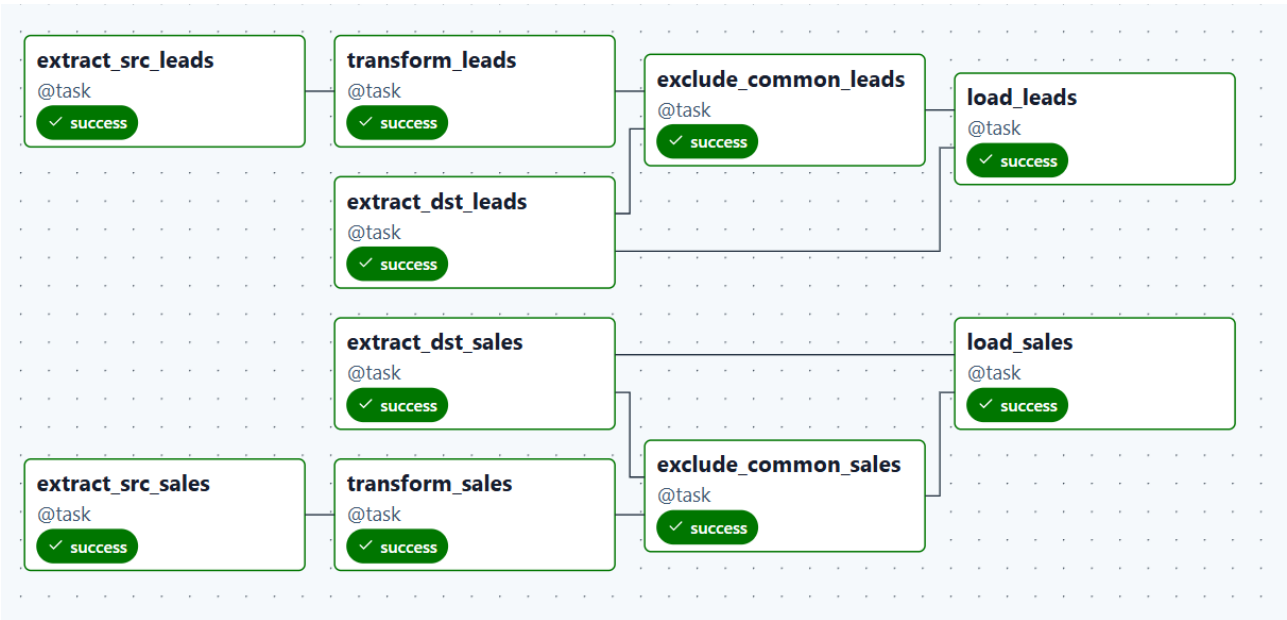
columns	data type	#missing	%missing	#unique	%unique	min	max	average	standard_devi...	first value	second value	third value
expected_value	Int32	988	63.05	277	17.68	4.00	60000000.00	7653028.71	7638707.59	<NA>	<NA>	<NA>
id	Int32	0	0.00	1567	100.00	319.00	11908.00	6179.72	3755.10	11218	776	3016
lead_id	Int32	0	0.00	1029	65.67	30.00	410425.00	158401.35	117226.22	130316	32565	20428
unit_value	float64	135	8.62	1249	79.71	1.00	67800000.00	6789495.95	6654969.14	17770000.0	648000.0	8912000.0
unit_location	object	110	7.02	31	1.98					New Cairo	New Capital C...	6th October
actual_value	Int32	770	49.14	736	46.97	225000.00	67800000.00	8037369.47	7897281.76	17777000	<NA>	8912000
date_of_reservation	object	209	13.34	610	38.93					2024-01-10 00...	2021-04-11 00...	2022-09-04 00...
reservation_update_date	object	1089	69.50	360	22.97					2024-01-10 20...	None	2022-09-05 00...
date_of_contraction	object	239	15.25	646	41.23					2024-02-08 00...	None	2022-10-23 00...
property_type_id	Int32	0	0.00	14	0.89	1.00	21.00	13.19	8.02	21	8	1
area_id	Int32	57	3.64	16	1.02	1.00	29.00	6.02	6.60	2	9	1
compound_id	Int32	77	4.91	306	19.53	5.00	1133.00	519.95	318.43	345	464	402
sale_category	object	292	18.63	6	0.38					Primary	None	Primary
years_of_payment	float64	677	43.20	14	0.89	0.00	12.00	7.19	2.06	8.0	nan	nan
property_type	object	0	0.00	14	0.89					Apartment	Office	Villa

ERD



Name	Type	Schema
Tables (4)		
data_source_LEADS		CREATE TABLE "data_source_LEADS" ("id" INTEGER, "date_of_last_request" TEXT, "buyer" INTEGER, "seller" INTEGER, "best_time_to_call" TEXT, "budget" REAL, "created_at" TEXT, "updated_at" TE
data_source_SALES		CREATE TABLE "data_source_SALES" ("id" INTEGER, "lead_id" INTEGER, "unit_value" REAL, "unit_location" TEXT, "expected_value" REAL, "actual_value" REAL, "date_of_reservation" TEXT, "reservati
dwh_LEADS		CREATE TABLE "dwh_LEADS" ("id" INTEGER NOT NULL PRIMARY KEY, "src_id" INTEGER NOT NULL, "date_of_last_request" TEXT NOT NULL, "buyer" INTEGER NOT NULL, "seller" INTEGER NOT NULL, '
dwh_SALES		CREATE TABLE "dwh_SALES" ("id" INTEGER NOT NULL PRIMARY KEY, "src_id" INTEGER NOT NULL, "lead_id" INTEGER NOT NULL, "unit_value" REAL, "unit_location" TEXT, "expected_value" REAL, "ac
Indices (0)		
Views (3)		
star_LEADS		CREATE VIEW star_LEADS AS SELECT src_id as id, date_of_last_request, buyer,seller, best_time_to_call, budget, created_at, updated_at, user_id, location, date_of_last_contact, status_name, comm
star_SALES		CREATE VIEW star_SALES AS SELECT src_id as id, lead_id, unit_value, unit_location, expected_value, actual_value, date_of_reservation, reservation_update_date, date_of_contraction, property_type
star_lead_to_sale_summary		CREATE VIEW star_lead_to_sale_summary AS SELECT l.id as "Lead ID", json_group_array(DISTINCT s.id) as "Sales Ids", l.location as "Location", l.status_name as "Status Name" , l.method_of_conta
Triggers (0)		

Pipeline



Above is the pipeline. We start by extracting the data from the source table in the “data_source” schema, and we store it in a Parquet file. I chose to store data in a Parquet file to avoid transmitting large amounts of data between tasks using XCom.

After extracting data, we move it to transformation step (see section below) where normalize it and correct it using a local LLM.

Then we extract data from the destination table, to avoid reprocessing identical data. We process only new data or updated data.

Finally, we implement scd type 2 and load the results into the “dwh” schema.

Transform data: Leads & Sales

We remove duplicated data from the data_source. Then:

Dates columns

convert the column (when necessary) to a common format `"%m/%d/%Y %H:%M:%S"`

Boolean columns

convert the column (when necessary) to True | False
if the column contains (1 or 0 or t or n)

String columns

All string columns are normalized as followed:

- remove leading and trailing spaces (trim)
- repalce [‘-’, ‘_’, ‘/’] with a space
- replace ‘&’ with ‘ and ’

- convert the value to lower case
- replace abbreviations with their full words (based on a manually maintained business dictionary)
- replace multiple consecutive spaces with a single space
- convert empty string to NA (this is more efficient for execution time than comparing strings to ‘’)
- convert the value to title case

Then I corrected the encoding of the Arabic ones, for example:

```
Ø£ÙƒØªÙˆØ¨Ø± Ø¨ÙˆÙ„Ø§Ø² Ù…Ù† Ø³ÙˆØ¯ÙŠÙƒ
```

becomes

```
أكتوبر بلاز ا من سوديك
```

Then I applied local LLM to correct the free text.
I used the model “`qwen2.5:latest`”. Since it’s a lightweight model running on my laptop, the execution time is high (and accuracy is not high). So:

- I applied it only on “`best_time_to_call`” in LEADS
- I build a LLM cache to store the corrections. For future execution, we retrieve the corrected text from the cache. If it doesn’t exist, we call the local LLM to correct it and then store the result in the cache

Why an LLM Cache?
To reduce execution time
It also allows us to build and maintain the cache independently of the pipeline.

Load data (SCD type 2)

In the schema dwh, we add the following columns to LEADS and SALES:

- src_id: integer
- effective_start_date: date
- effective_end_date: date
- is_current: bool
- version_id: integer

and we implement SCD type 2, before loading the data in the db.
If new data:

- we set
  - effective_start_date = current date
  - is_current = True
  - version_id = 1
- src_id is equal to the ID in the original table. We need it to keep a reference to the original ID and avoid duplicated Ids, since we are implementing scd type 2

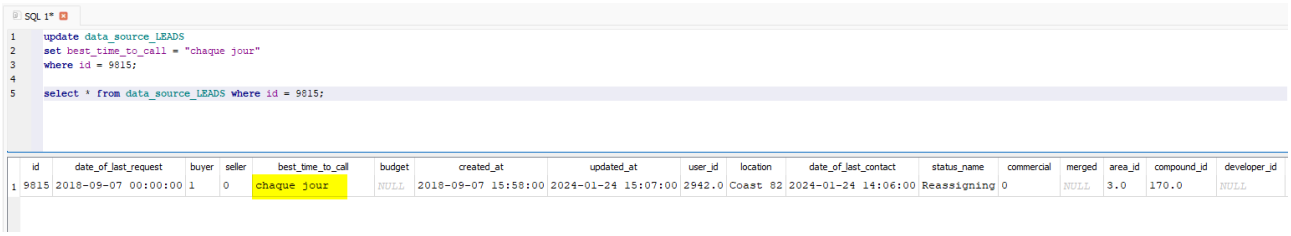
If updated data:

- last current data, we set
  - effective_end_date = current date
  - is_current = False
- new update:
  - effective_start_date = current date
  - is_current = True
  - version_id = last version + 1

Example of Execution:

We ran the pipeline the first time to extract, transform and load all data into dwh.

Let’s update “best_time_to_call” to “chaque jour” for the id = 9815 in data_source.leads:

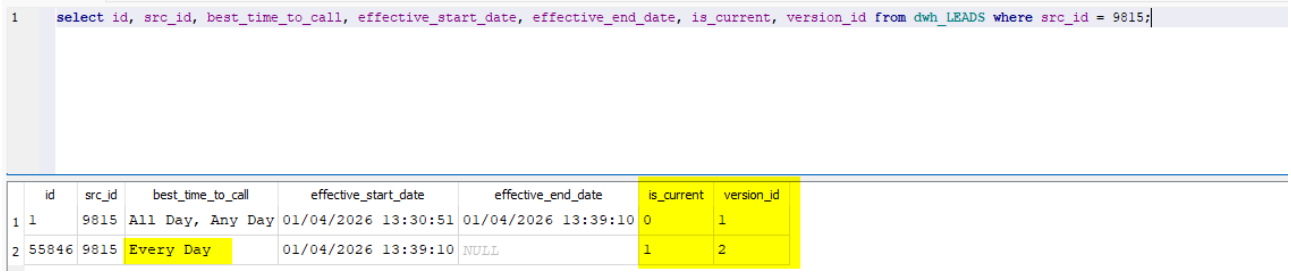


```
SQL 1*
update data_source_LEADS
set best_time_to_call = "chaque jour"
where id = 9815;

select * from data_source_LEADS where id = 9815;
```

	id	date_of_last_request	buyer	seller	best_time_to_call	budget	created_at	updated_at	user_id	location	date_of_last_contact	status_name	commercial	merged	area_id	compound_id	developer_id
1	9815	2018-09-07 00:00:00	1	0	chaque jour	NULL	2018-09-07 15:58:00	2024-01-24 15:07:00	2942.0	Coast 82	2024-01-24 14:06:00	Reassigning	0	NULL	3.0	170.0	NULL

After running the pipeline for the second time:



```
select id, src_id, best_time_to_call, effective_start_date, effective_end_date, is_current, version_id from dwh_LEADS where src_id = 9815;
```

	id	src_id	best_time_to_call	effective_start_date	effective_end_date	is_current	version_id
1	1	9815	All Day, Any Day	01/04/2026 13:30:51	01/04/2026 13:39:10	0	1
2	55846	9815	Every Day	01/04/2026 13:39:10	NULL	1	2

We can see here that the French word is automatically translated by the LLM into “Every day” and we can also see the result of the SCD Type 2 implementation.

Star table

I created three views

star_leads and star_sales

They are views pointing to the is_current=True

star_lead_to_sale_summary

It contains a summary of data that can answer following questions:

- What percentage of leads convert to at least one sale? (Sales count > 0)
- Which campaigns/contact methods have the highest conversion rate?
- Where we are losing leads?
- ....