Technical Architecture Document

DVF Data Lake & Data Warehouse

1. Purpose of This Document

This Technical Architecture Document describes the design and implementation of a local data platform for dvf Data using open-source tools.

The document explains:

- how data flows through the system
- how each component works
- why specific technologies were chosen

This document is used to communicate the architecture clearly to technical and nontechnical stakeholders.

2. Project Overview

Project Name

Real Estate Market Analysis in France

Objective

The objective of this project is to design a simple but realistic data architecture that follows modern data engineering principles:

- ingest open data
- store raw data safely
- clean and transform data
- load data into a Data Warehouse
- enable SQL-based analytics

3. Global Architecture Overview

The architecture is based on a layered approach, which separates responsibilities and improves clarity.

The system is composed of four main layers:

1. Data Source
2. Data Lake
3. Data Warehouse

4. Analytics and BI

Each layer has a specific role and responsibility.

4. Data Source Layer

Description:

The data source is an open dvf dataset provided as a CSV file.

Characteristics

- public and free
- static dataset
- Real estate market data

Role:

The data source represents the entry point of the data pipeline.

5. **Data Lake Layer**

Purpose:

The Data Lake stores data as files and preserves data flexibility and history.

Technology

- Local file system
- CSV files
- Python Scripts for processing

Data Lake Zones

RAW Zone

- stores original data
- no transformation
- acts as the source of truth

STAGING Zone

- cleaned and standardized data
- duplicates removed
- missing values handled
- date formats normalized

CURATED Zone

- BI-ready data

- selected columns only
- aggregated datasets
- optimized for analytics

This zone is used as the input for the Data Warehouse.

6. Data Warehouse Layer

Purpose

The Data Warehouse stores clean, structured data optimized for SQL queries and analytics.

Technology Choice: DuckDB

DuckDB was chosen because:

- it runs locally
- it requires no server
- it requires no license
- it supports standard SQL
- it is optimized for analytical workloads

Storage

The Data Warehouse is stored as a local file: warehouse/dvf_market.db

Tables

- fact_monthly_indicators
- dim_top_departments_volume
- dim_top_departments_price

These tables are used for Business Intelligence queries.

7. Analytics and BI Layer

Purpose

This layer enables data analysis and business insights.

Capabilities

- SQL queries
- aggregations
- metrics and indicators

Example Analyses

- Is data available for January 2026?

- If not, what is the latest available month?
- What is the median price per square meter
- Which are the top 10 departments

Architecture Flow Summary

The data flow follows this logic:

- Data is ingested from an open source
- Raw data is stored in the Data Lake
- Data is cleaned and transformed
- Clean data is loaded into the Data Warehouse
- SQL queries generate insights and reports

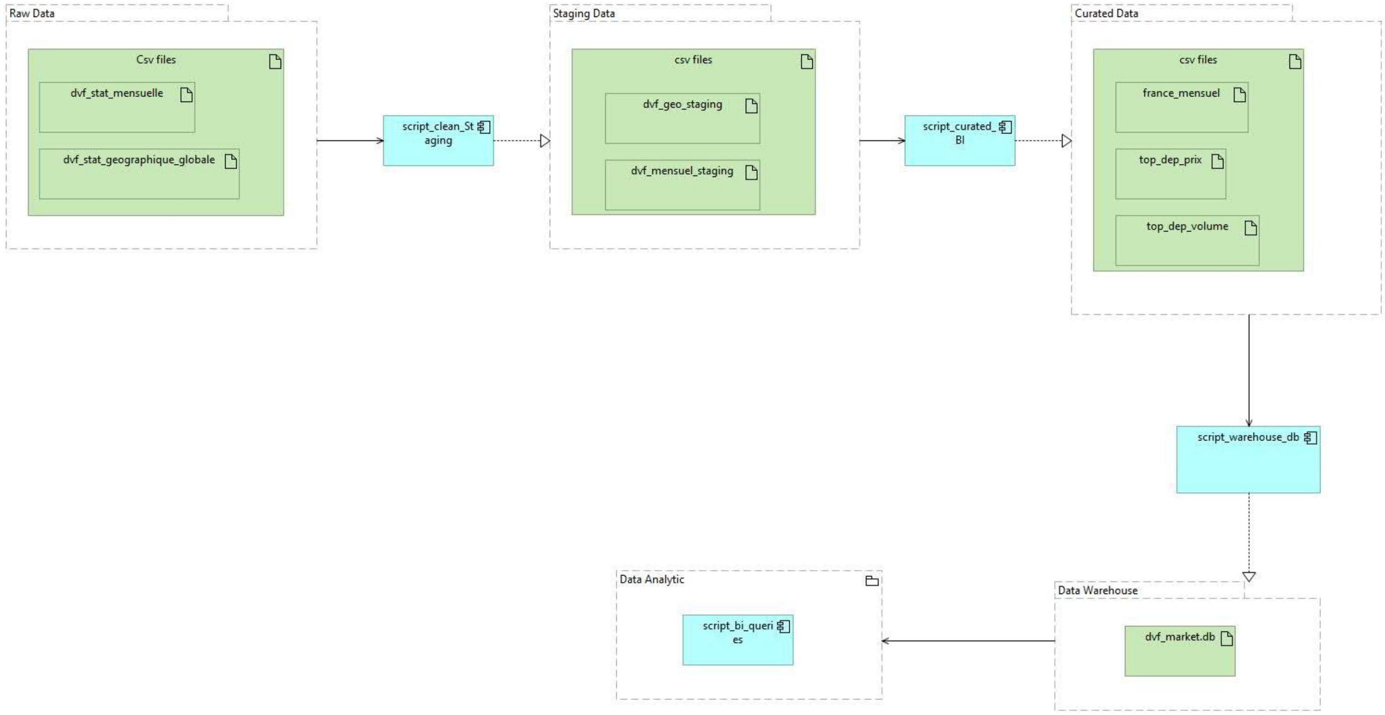

Figure: The Architecture of the dvf data pipeline