Your Ultimate Guide To Landing
Top AI roles

DECODE
AiML

SQL for Pandas DataFrame

→ We want to learn how to use SQL query on a dataframe.

→ Though we have seen how to perform different operations on a dataframe using built-in methods. Why do we need SQL for operations on DF?

① SQL Popularity and Familiarity.
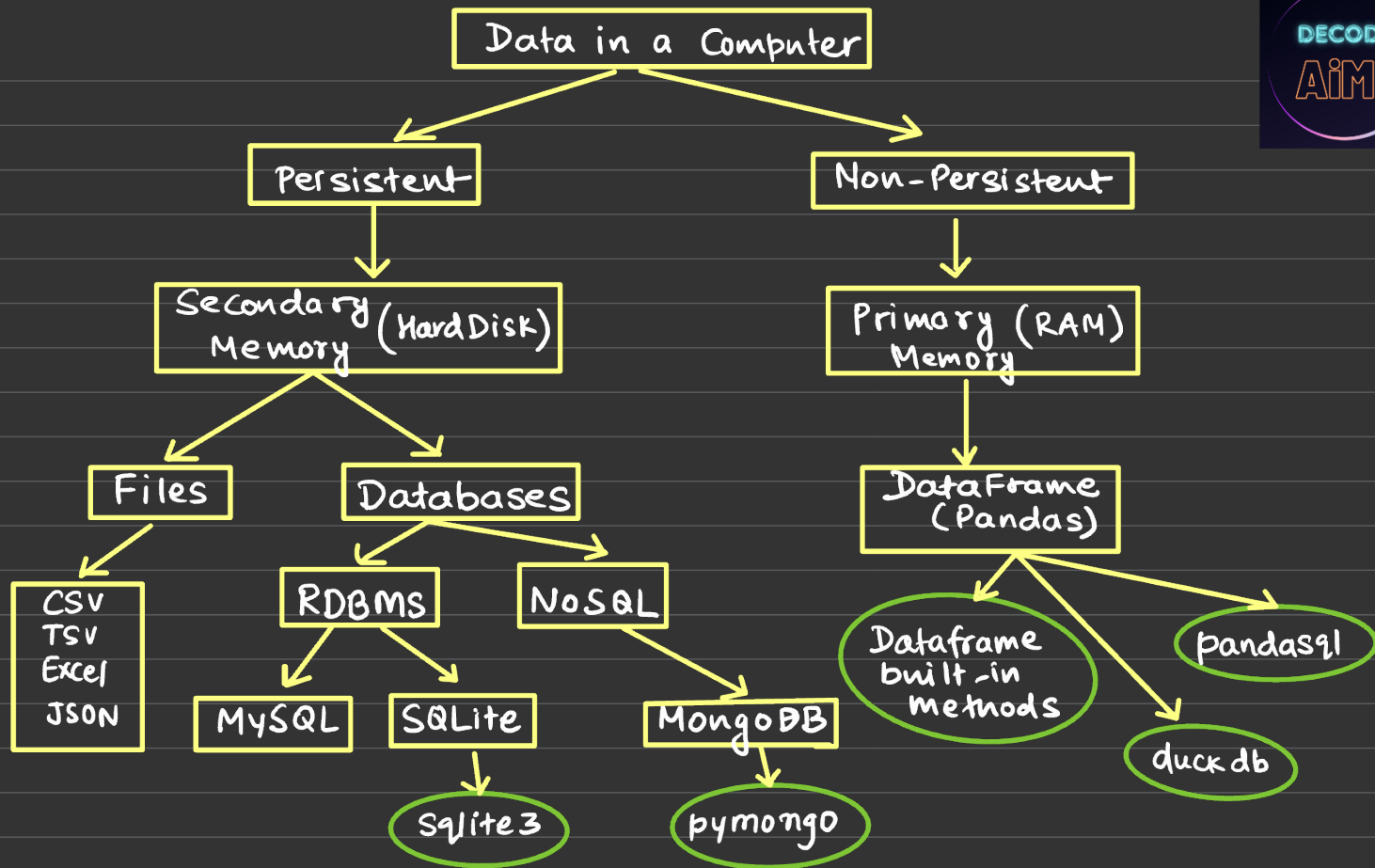② Portability of Queries
    → Code reusability across (database → pandas → Spark → etc)
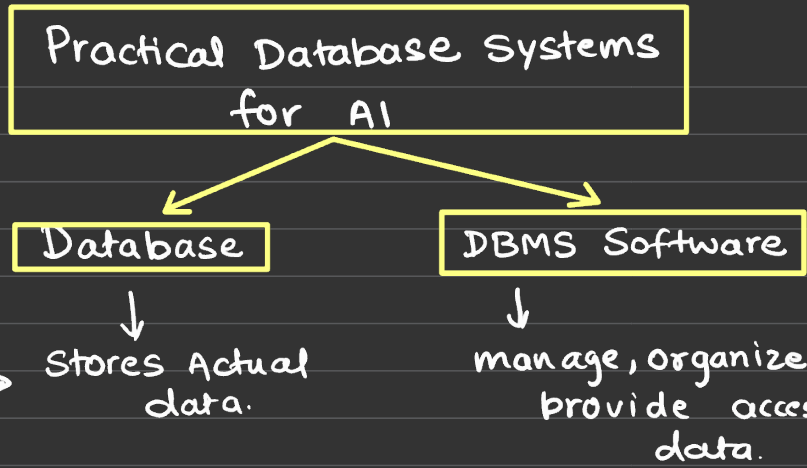③ Complex Join and nested Queries
④ Readability of SQL

→ But, working on dataframe using builtin methods is faster compared to using SQL over a dataframe

→ Let's see step by step.

```
Data in a Computer
├── Persistent
│   └── Secondary Memory (Hard Disk)
│       ├── Files
│       │   └── CSV, TSV, Excel, JSON
│       └── Databases
│           ├── RDBMS
│           │   ├── MySQL
│           │   └── SQLite
│           │       └── Sqlite3
│           └── NoSQL
│               └── MongoDB
│                   └── pymongo
└── Non-Persistent
    └── Primary (RAM) Memory
        └── DataFrame (Pandas)
            ├── Dataframe built-in methods
            ├── pandasql
            └── duckdb
```

→ Sqlite3, pymongo, duckdb and pandasql are python modules

Practical Database Systems for AI

```
Database                          DBMS Software
   ↓                                   ↓
Stores Actual                   manage, organize and
data.                            provide access to
                                       data.
```

Content managed by DBMS software. →

→ In SQLite, the .db file is the database (data). sqlite3 is the software (DBMS) that let you access it.

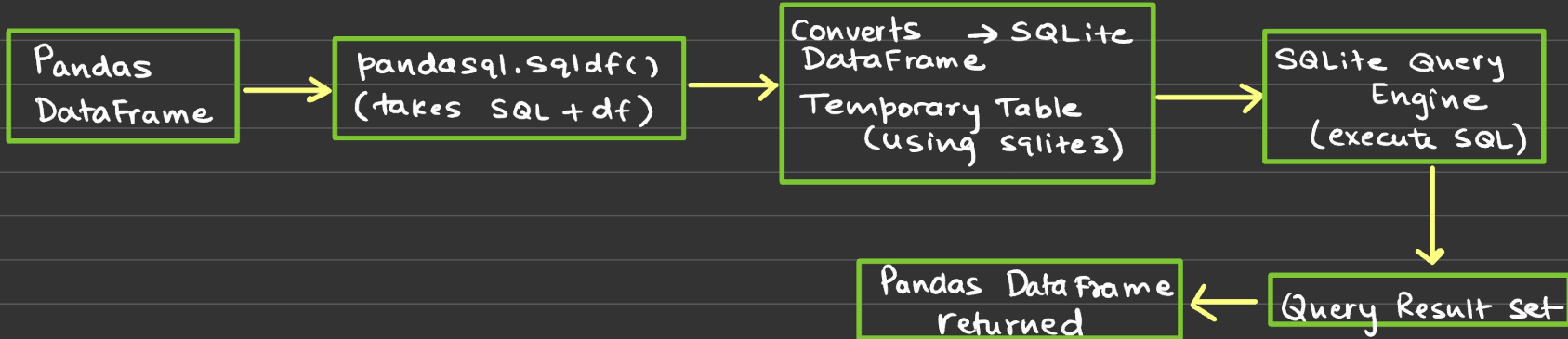→ For SQLite & sqlite3, Refer Lecture — 1.14

→ To work with dataframe using SQL, we will be using 2

   Common tools → **pandasql** and **duckdb**

pandasql

→ A python library that let you run SQL queries on Pandas Dataframe.
→ Internally, it uses SQLite
→ Mostly used for learning SQL or small datasets.

→ How PandaSQL (pandasql) works?

| Pandas DataFrame | → | pandasql.sqldf() (takes SQL + df) | → | Converts → SQLite DataFrame Temporary Table (using sqlite3) | → | SQLite Query Engine (execute SQL) |

| Pandas DataFrame returned | ← | Query Result set |

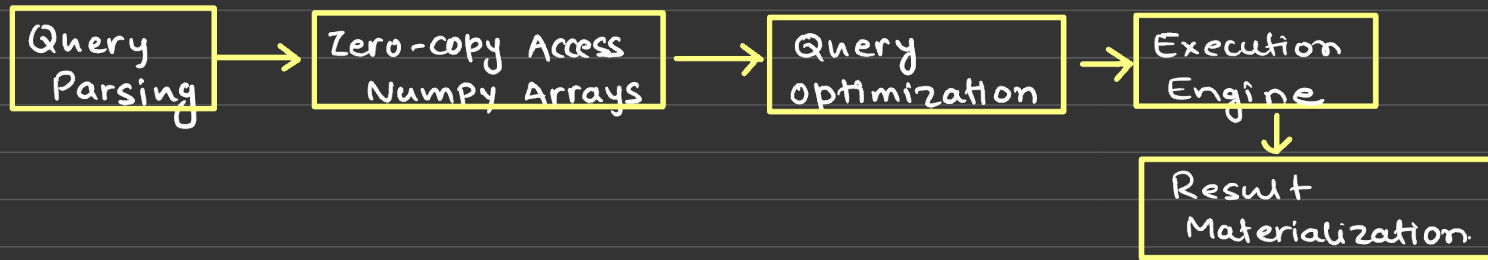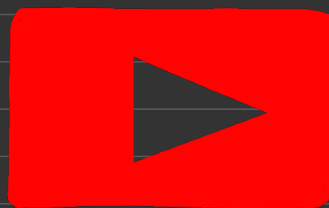→ Slow on large datasets → it copies dataframe to SQLite each time.

→ duckdb is a python module that helps run queries on a DataFrame, files or DuckDB databases.

→ Modern analytical SQL engine for Python

→ Can query Pandas dataframe directly without copying.

→ Optimized for large datasets and complex queries.

→ No overhead of copying.

→ How duckdb works?

```
┌──────────────┐    ┌──────────────────┐    ┌──────────────┐    ┌──────────────┐
│ Query        │ →  │ Zero-copy Access │ →  │ Query        │ →  │ Execution    │
│ Parsing      │    │ Numpy Arrays     │    │ Optimization │    │ Engine       │
└──────────────┘    └──────────────────┘    └──────────────┘    └──────────────┘
                                                                        │
                                                                        ↓
                                                                 ┌──────────────────┐
                                                                 │ Result           │
                                                                 │ Materialization  │
                                                                 └──────────────────┘
```

→ The result is produced as a DuckDB relation and then converted back to a dataframe using .to_df()