"I'm extracting data now... when does ETL happen?"

```

### \*\*Reality:\*\*

```

You're BUILDING the ETL components now (Week 1-2)

You'll AUTOMATE them with Airflow later (Week 3-4)

Then ETL RUNS automatically (Week 5+)

```

---

## ETL Timeline Breakdown

### \*\*WEEK 1-2: BUILD ETL Components (What you're doing NOW)\*\*

```

This is DEVELOPMENT phase - building the pieces

EXTRACT (E):

├─ Write: scripts/collect\_crime.py

├─ Write: scripts/collect\_properties.py

├─ Write: scripts/collect\_demographics.py

├─ Run manually: python scripts/collect\_crime.py

├─ Save to: data/raw/\*.csv

└─ Status: Building extract logic ✅

TRANSFORM (T):

├─ Write: scripts/process\_crime\_data.py

├─ Write: scripts/clean\_properties.py

├─ Write: scripts/calculate\_trends.py

├─ Run manually: python scripts/process\_crime\_data.py

├─ Save to: data/processed/\*.csv

└─ Status: Building transform logic ✅

LOAD (L):

├─ Write: scripts/load\_to\_postgres.py

├─ Write: scripts/load\_to\_vectordb.py

├─ Run manually: python scripts/load\_to\_postgres.py

├─ Load to: PostgreSQL + Chroma

└─ Status: Building load logic ✅

What you're doing: Creating the ETL SCRIPTS

NOT running automated pipeline yet!

```

---

### \*\*WEEK 3-4: AUTOMATE ETL with Airflow (Put pieces together)\*\*

```

This is AUTOMATION phase - orchestrate the pipeline

Take your scripts from Week 1-2 and convert to Airflow DAG:

# airflow/dags/propbot\_etl\_pipeline.py

from airflow import DAG

from airflow.operators.python import PythonOperator

# Import your functions from Week 1-2 scripts

from scripts.collect\_crime import collect\_boston\_crime\_data

from scripts.process\_crime\_data import process\_crime

from scripts.load\_to\_postgres import load\_crime\_to\_db

dag = DAG('propbot\_etl\_pipeline', schedule\_interval='@daily')

# EXTRACT task (uses your Week 1 script!)

extract\_crime = PythonOperator(

task\_id='extract\_crime\_from\_api',

python\_callable=collect\_boston\_crime\_data,

dag=dag

)

# TRANSFORM task (uses your Week 1-2 script!)

transform\_crime = PythonOperator(

task\_id='transform\_clean\_crime\_data',

python\_callable=process\_crime,

dag=dag

)

# LOAD task (uses your Week 2 script!)

load\_crime = PythonOperator(

task\_id='load\_crime\_to\_postgresql',

python\_callable=load\_crime\_to\_db,

dag=dag

)

# ETL Flow

extract\_crime >> transform\_crime >> load\_crime

What you're doing: Connecting ETL scripts into automated DAG

Now it runs on schedule!

```

---

### \*\*WEEK 5+: RUN ETL Automatically (Production mode)\*\*

```

This is PRODUCTION phase - pipeline runs itself

Airflow scheduler runs your DAG:

├─ Every day at 2am (scheduled)

├─ Automatically runs: Extract → Transform → Load

├─ Updates databases with fresh data

├─ Logs everything

└─ Alerts if errors occur

You don't run scripts manually anymore!

Airflow runs them automatically! ✅

```

---

## Visual Timeline

```

┌─────────────────────────────────────────────────────────┐

│ WEEK 1-2: DEVELOPMENT │

│ "Building the ETL components" │

├─────────────────────────────────────────────────────────┤

│ │

│ YOU manually run: │

│ └─ python scripts/collect\_crime.py (EXTRACT) │

│ └─ python scripts/process\_crime.py (TRANSFORM) │

│ └─ python scripts/load\_to\_postgres.py (LOAD) │

│ │

│ Data flow: │

│ API → CSV → Cleaned CSV → PostgreSQL │

│ │

│ ETL happens: Manually, one step at a time │

│ Purpose: Build and test each component │

│ │

└─────────────────────────────────────────────────────────┘

↓

┌─────────────────────────────────────────────────────────┐

│ WEEK 3-4: AUTOMATION │

│ "Converting scripts to Airflow DAG" │

├─────────────────────────────────────────────────────────┤

│ │

│ Create Airflow DAG that calls your scripts: │

│ │

│ extract\_task >> transform\_task >> load\_task │

│ │

│ Test DAG runs successfully │

│ Add: Logging, error handling, alerts │

│ │

│ ETL happens: In Airflow, but triggered manually │

│ Purpose: Orchestrate components into pipeline │

│ │

└─────────────────────────────────────────────────────────┘

↓

┌─────────────────────────────────────────────────────────┐

│ WEEK 5+: PRODUCTION │

│ "Automated pipeline runs on schedule" │

├─────────────────────────────────────────────────────────┤

│ │

│ Airflow runs automatically: │

│ └─ Every day at 2am │

│ └─ Extract → Transform → Load │

│ └─ Updates databases │

│ └─ Sends alerts if failures │

│ │

│ You DON'T run manually anymore! │

│ Airflow handles everything! │

│ │

│ ETL happens: Automatically, on schedule │

│ Purpose: Production-ready data pipeline │

│ │

└─────────────────────────────────────────────────────────┘

```

---

## Current Status: Where You Are NOW

```

YOU ARE HERE (Week 1):

│

├─ ✅ Project structure created

├─ ✅ Virtual environment set up

├─ ✅ Dependencies installed

├─ 🔄 About to run first EXTRACT script

│

├─ ❌ NOT using Airflow yet (Week 3)

├─ ❌ NOT loading to databases yet (Week 2)

├─ ❌ NOT running automated pipeline yet (Week 5)

│

└─ FOCUS: Manually collect data, save to CSV files

This is CORRECT! You're following the right order! ✅