# 1: Data Ingestion & Storage

**Task 1: Download a Real-world Dataset**

🔹 Dataset: **New York Taxi Trips Data**  
📥 Download: [NYC Taxi Data (Parquet format)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page)  
🔹 Alternative: **Kaggle Datasets** ([Download CSV datasets](https://www.kaggle.com/datasets))

**Task 2: Load Data into a Local Database**

* Install and Use **PostgreSQL** (or SQLite) as a database.
* Write a **Python script to load data** into the database.

**Resources:**

* [PostgreSQL Installation Guide](https://www.postgresql.org/download/)
* [Pandas to PostgreSQL (Tutorial)](https://towardsdatascience.com/from-pandas-to-postgresql-3f24c7f9f4f4)
* [SQLite Quickstart](https://www.sqlitetutorial.net/)

📌 **Practice Steps:**  
✅ Install **PostgreSQL** or **SQLite**.  
✅ Use **Pandas** to **read the dataset**.  
✅ Write a **Python script** to insert data into the database.

A computer screen with white text

AI-generated content may be incorrect.

# 2: Data Processing & Transformation

**Task 3: Transform Data Using Pandas & SQL**

* **Filter out invalid data** (e.g., negative trip distances).
* **Convert datetime columns** into proper formats.
* **Aggregate data** (e.g., average fare per trip).

**Resources:**

* [SQL Basics (W3Schools)](https://www.w3schools.com/sql/)
* [Pandas Data Transformations](https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html)

📌 **Practice Steps:**  
✅ Write **SQL queries** to clean the data.  
✅ Perform **aggregations** using Pandas.

A screenshot of a computer program

AI-generated content may be incorrect.

# 3: Data Orchestration with Apache Airflow

**Task 4: Automate Data Processing with Airflow**

* Install **Apache Airflow** (pip install apache-airflow).
* Create an **Airflow DAG** (Directed Acyclic Graph) to automate:
  + **Ingesting data** from the dataset.
  + **Transforming data** using SQL.
  + **Storing results** in a database.

**Resources:**

* [Airflow Quickstart Guide](https://airflow.apache.org/docs/apache-airflow/stable/start/index.html)
* [Airflow DAGs Tutorial](https://towardsdatascience.com/apache-airflow-tutorial-a-complete-pipeline-7fa83e23ffdc)

📌 **Practice Steps:**  
✅ Install **Airflow** and configure it.  
✅ Write a **DAG to automate data ingestion & transformation**.  
✅ Schedule the DAG to run **every fixed interval e.g.: 5 minute** or every hour:

A screenshot of a computer program

AI-generated content may be incorrect.

# Additional Resources for Downloading Notebooks & Datasets

**Open Datasets**

1. **Kaggle** – <https://www.kaggle.com/datasets>
2. **Google Dataset Search** – <https://datasetsearch.research.google.com/>
3. **AWS Open Data** – <https://registry.opendata.aws/>
4. **NYC Taxi Data** – <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

**Jupyter Notebooks & Tutorials**

1. **DataTalksClub Data Engineering Zoomcamp** – <https://github.com/DataTalksClub/data-engineering-zoomcamp>
2. **Data Engineering Notebooks (GitHub)** – <https://github.com/awesomedata/awesome-public-datasets>
3. **Pandas & SQL Practice Notebooks** – <https://github.com/jakevdp/Pandas-Tutorial>
4. **Apache Airflow Examples** – <https://github.com/apache/airflow/tree/main/airflow/example_dags>

**🚀 What You Will Have Built in 3 Labs Above:**

✅ **Ingested a real dataset** into a database (PostgreSQL).  
✅ **Transformed & cleaned data** using Pandas & SQL.  
✅ **Automated data processing** with Apache Airflow.  
✅ **Created a reproducible data pipeline** for ML.

# 📌 What's Next?

If you have more time, try these:  
🔥 **Deploy your pipeline on the cloud** (AWS/GCP/Azure).  
🔥 **Use Kafka for real-time data ingestion**.  
🔥 **Implement a Feature Store with Feast**.