---

📌 \*\*¿Dónde guardarlo?\*\*

El archivo `README.md` ya está correctamente ubicado en tu carpeta principal del proyecto `C:\Users\JONANNA\Desktop\airflow`.

Si deseas mejorar la organización de los scripts, puedes mover todos (excepto el DAG) a la carpeta `scripts/`, pero si ya funcionan desde `dags/`, puedes dejarlos allí sin problema.

¿Deseas que generemos un video paso a paso o una plantilla final para entrega?

airflow/

│

├── dags/

│   ├── proyecto\_bigdata\_dag.py           # DAG principal que orquesta el flujo

│   ├── kafka\_ingest.py                   # Simula la ingesta de datos

│   ├── transformar\_datos.py              # Procesa los datos

│   ├── guardar\_en\_sql.py                 # Carga datos a SQL

│   ├── hacer\_grafica.py                  # Visualiza datos en una gráfica

│   ├── modelo\_ml.py                      # Entrena un modelo de ML

│   ├── airflowdagsproyecto\_bigdata\_dag.py# Pruebas iniciales del DAG (puede eliminarse)

│

├── logs/

│   ├── dag\_id=proyecto\_bigdata/          # Logs del DAG ejecutado

│   ├── dag\_processor\_manager/            # Logs de Airflow

│   ├── scheduler/                        # Logs del programador de tareas

│

├── plugins/                              # Plugins personalizados (vacío por ahora)

│

├── scripts/                              # Carpeta auxiliar (si deseas mover scripts aquí)

│

├── docker-compose.yml                    # (Opcional) si usas Docker

├── README.md                             # Este archivo

├── Taller final de Big Data (1).pdf      # Guía del taller

```

---

**## 📜 ¿Qué hace cada script?**

| Script                 | Descripción                                                                 |

|------------------------|-----------------------------------------------------------------------------|

| `proyecto\_bigdata\_dag.py` | Define el flujo (DAG) completo en Airflow.                                |

| `kafka\_ingest.py`      | Simula la ingesta de datos (lectura de CSV o generación aleatoria).        |

| `transformar\_datos.py` | Limpia, transforma y normaliza los datos.                                  |

| `guardar\_en\_sql.py`    | Inserta los datos en una base de datos (PostgreSQL o MySQL).               |

| `hacer\_grafica.py`     | Crea una gráfica PNG con los datos procesados usando Matplotlib.           |

| `modelo\_ml.py`         | Entrena un modelo de machine learning simple (por ejemplo, regresión).     |

| `airflowdagsproyecto\_bigdata\_dag.py` | Ejemplo de prueba inicial del DAG, puede eliminarse.               |

---

**## 🛠️ Cómo ejecutar Airflow**

1. Abre tu terminal y activa tu entorno virtual.

2. Inicializa la base de datos de Airflow:

```bash

airflow db init

```

3. Crea un usuario administrador:

```bash

airflow users create \

  --username admin \

  --firstname Johanna \

  --lastname Soler \

  --role Admin \

  --email johannasolermine@gmail.com

```

4. Inicia los servicios:

```bash

airflow scheduler

# En otra terminal:

airflow webserver

```

5. Abre [http://localhost:8080](http://localhost:8080) en tu navegador.

---

**## 🔁 ¿Qué hace cada tarea del DAG?**

| Tarea               | Script asociado         | Función                                                       |

|--------------------|-------------------------|----------------------------------------------------------------|

| `ingesta\_datos`     | `kafka\_ingest.py`        | Simula y produce datos que serían "leídos" de Kafka.          |

| `transformar\_datos` | `transformar\_datos.py`   | Limpia y transforma los datos crudos para su uso posterior.   |

| `guardar\_en\_sql`    | `guardar\_en\_sql.py`      | Inserta los datos ya limpios en la base de datos.             |

| `hacer\_grafica`     | `hacer\_grafica.py`       | Genera una gráfica con los datos ya cargados.                 |

| `modelo\_ml`         | `modelo\_ml.py`           | Entrena un modelo de ML básico (como regresión lineal).       |

---

**## ✅ Requisitos**

- Python 3.11

- Apache Airflow

- kafka-python

- Dask

- pandas

- matplotlib

- scikit-learn

- MySQL o PostgreSQL

Instala dependencias básicas con:

```bash

pip install kafka-python dask pandas matplotlib scikit-learn

```

---

**## 🎯 Resultado Esperado**

Cuando ejecutas el DAG desde la interfaz web de Airflow:

1. Se **\*\*ingresan datos simulados\*\***

2. Se **\*\*transforman\*\***

3. Se **\*\*guardan en una base de datos SQL\*\***

4. Se **\*\*genera una gráfica\*\***

5. Se **\*\*entrena un modelo de ML\*\***

✅ Todo esto se realiza automáticamente gracias a la orquestación de Airflow.

---

**### Autora: Johanna Soler**

Maestría en Inteligencia de Negocios – Proyecto Final Big Data

Abril 2025