

Taller Final - Big Data con Apache Airflow

1. Nombre del Proyecto: Flujo de datos del IDEAM - Precipitación mensual 2020-2025. Departamento del Meta, estación SENA (Villavicencio)

2. Descripción general del flujo: este proyecto implementa un pipeline de Big Data orquestado con Apache Airflow, que automatiza la captura, transformación, visualización y análisis de datos de precipitación. Se simula una fuente tipo Kafka, se transforman los datos con Dask, se almacenan en PostgreSQL, se generan gráficos y se entrena un modelo de regresión lineal.

También, se ha incorporado un módulo de predicción para simular los valores futuros de precipitación en los próximos meses, extendiendo el modelo a una visión prospectiva automatizada.


Fases del flujo:

1. **Ingesta de datos:** Simulación de Kafka para cargar un archivo CSV con datos originales.
2. **Procesamiento:** Uso de Dask para transformar los datos (duplicación de valores para simular pérdidas o errores).
3. **Persistencia:** Almacenamiento de los datos procesados en una base de datos PostgreSQL.
4. **Visualización:** Generación de gráficos (línea e histograma) usando Matplotlib.
5. **Modelo de Machine Learning:** Entrenamiento de una regresión lineal con scikit-learn para predecir la precipitación.
6. **Exportación:** Guardado del modelo en formato .joblib y de las predicciones en un archivo .csv.

2. Diagrama del flujo en Airflow

plaintext

 Copiar

 Editar

```
Kafka (simulado)
  ↓
  Dask
  ↓
PostgreSQL
  ↓
Visualización
  ↓
Machine Learning
  ↓
Predicción futura (nuevo)
```

3. Fragmento destacado del DAG

```
with DAG(
    dag_id='proyecto_bigdata',
    default_args=default_args,
    start_date=datetime(2025, 4, 10),
    schedule_interval='@daily',
    catchup=False,
    description='Flujo de procesamiento big data con ML',
) as dag:
    ...
    entrenar_modelo_ml_task = PythonOperator(
        task_id='entrenar_modelo_ml',
        python_callable=entrenar_modelo_ml
    )
    procesar_datos_dask >> entrenar_modelo_ml_task
```

```
python

t1 = PythonOperator(task_id='leer_kafka', python_callable=leer_datos_kafka)
t2 = PythonOperator(task_id='procesar_datos_dask', python_callable=procesar_datos)
t3 = PythonOperator(task_id='guardar_postgresql', python_callable=guardar_en_sql)
t4a = PythonOperator(task_id='crear_grafico_linea', python_callable=graficar_linea)
t4b = PythonOperator(task_id='crear_grafico_outliers', python_callable=crear_grafico_outliers)
t4c = PythonOperator(task_id='crear_histograma', python_callable=graficar_histograma)
t5 = PythonOperator(task_id='modelo_machine_learning', python_callable=modelo_ml)
t6 = PythonOperator(task_id='entrenar_modelo_ml', python_callable=entrenar_modelo_ml)
t7 = PythonOperator(task_id='prediccion_futuro', python_callable=predecir_futuro)

t1 >> t2 >> t3 >> [t4a, t4b, t4c] >> t5 >> t6 >> t7
```

4. Justificación técnica y logros

- Se integraron herramientas reales del ecosistema Big Data: Kafka (simulado), Dask, PostgreSQL y Airflow.
- Se generaron visualizaciones automáticas dentro del flujo.
- Se incluyó un modelo de Machine Learning simple (regresión lineal), entrenado automáticamente.
- Extensión del flujo para predecir valores futuros a partir del modelo entrenado.
- Todos los productos (CSV, PNG, modelo .joblib) se almacenan en /opt/airflow/salidas, permitiendo auditar el flujo.
- Airflow permite observar ejecuciones, tiempos, y logs por cada tarea.

Anexos:

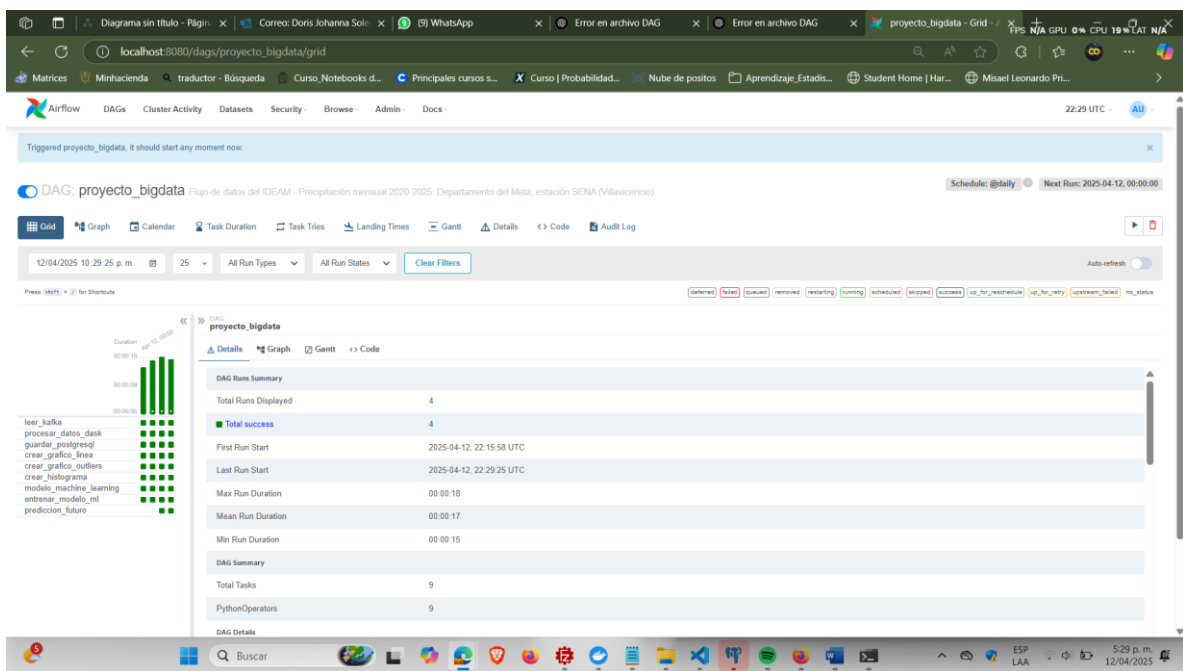
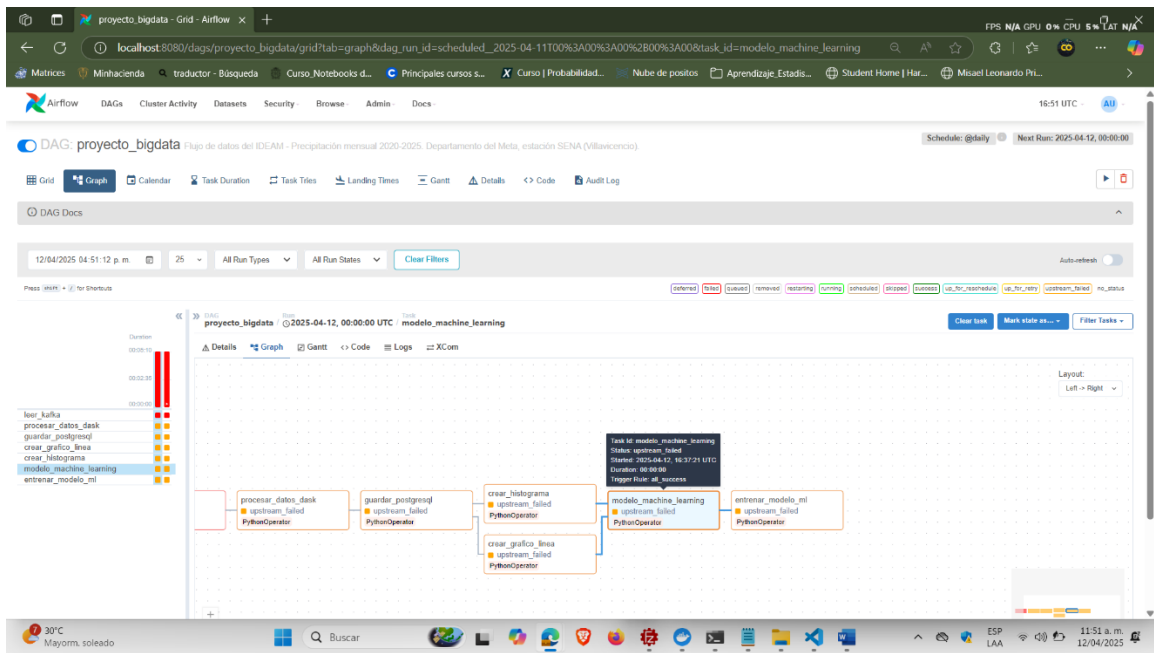
- Capturas de la interfaz de Airflow funcionando
- Archivos: proyecto_bigdata_dag.py, modelo_ml.py, datos, gráficos y predicciones.

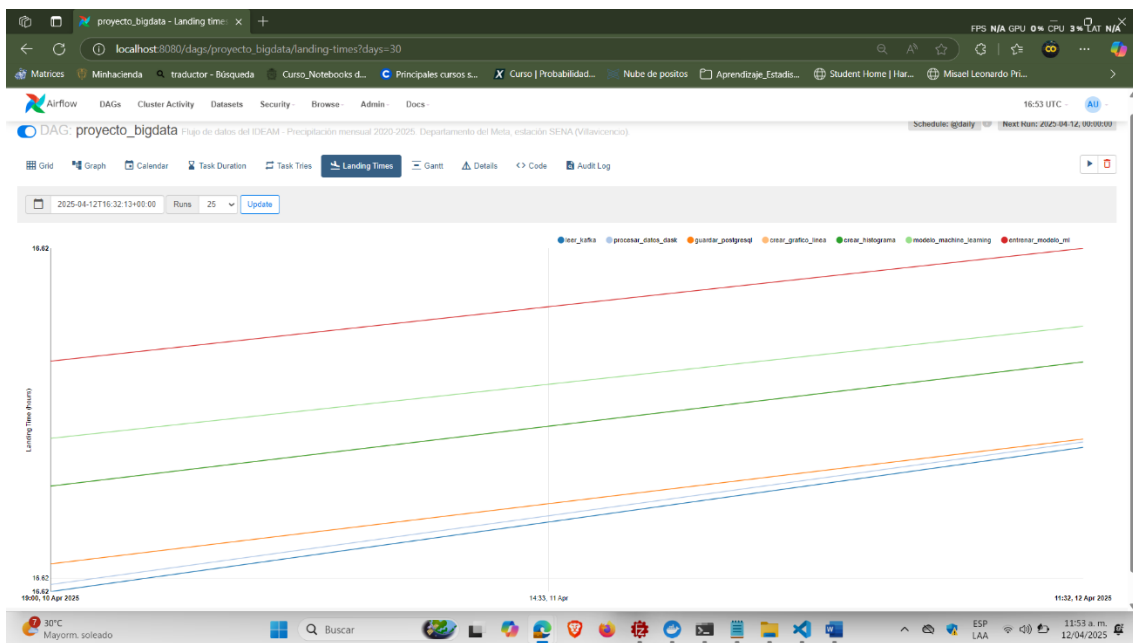
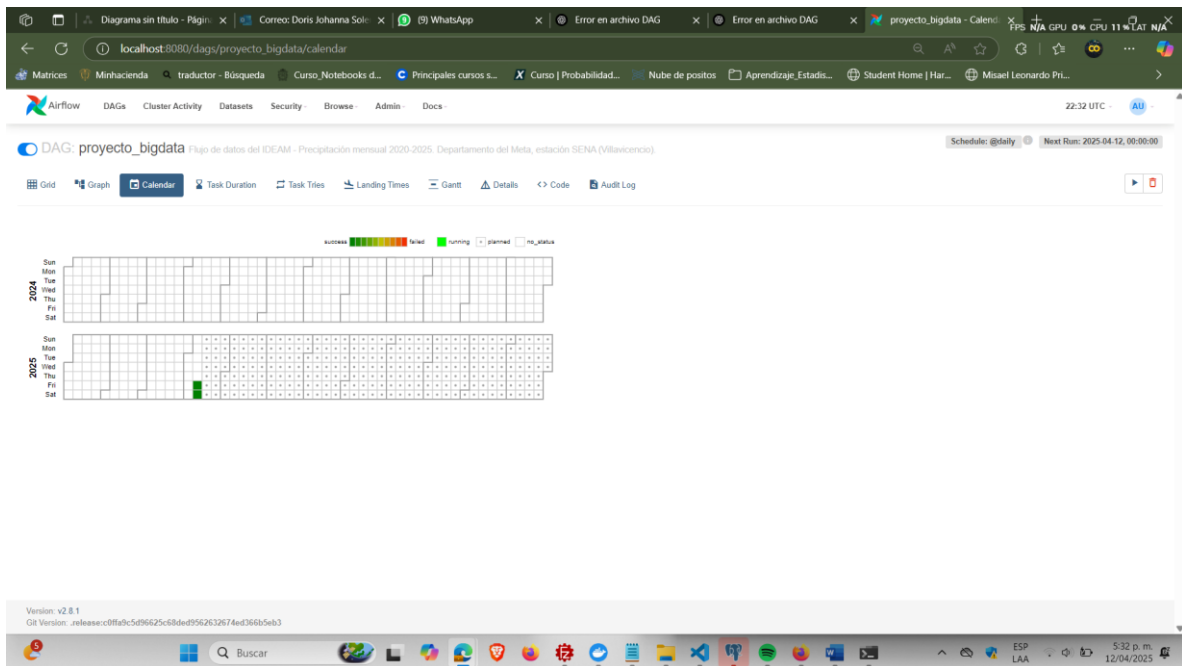
Estudiante: Doris Johanna Soler Castro CC 40330258 MINE III Semestre

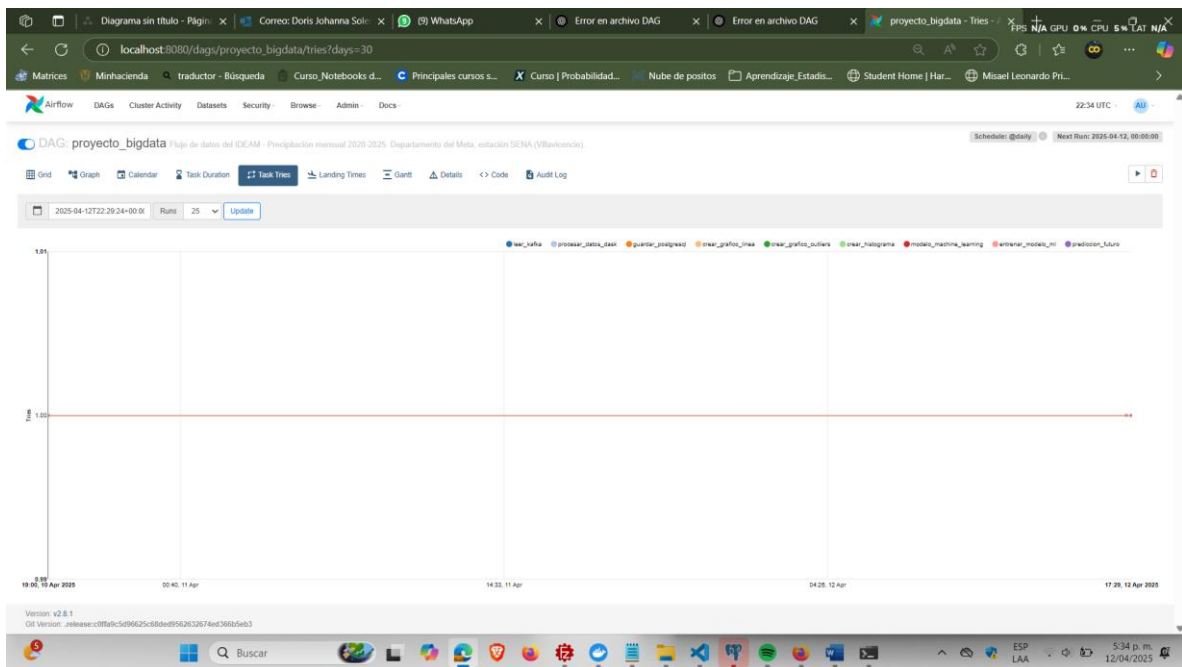
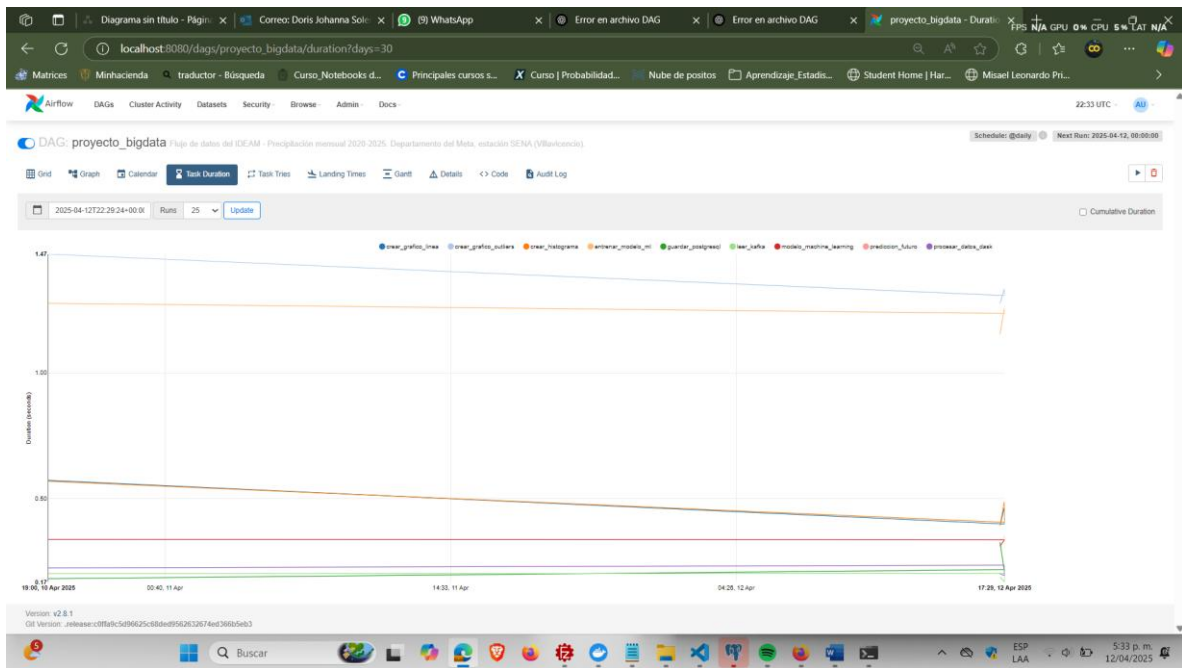
CAPTURAS DE LA INTERFAZ DE AIRFLOW FUNCIONANDO

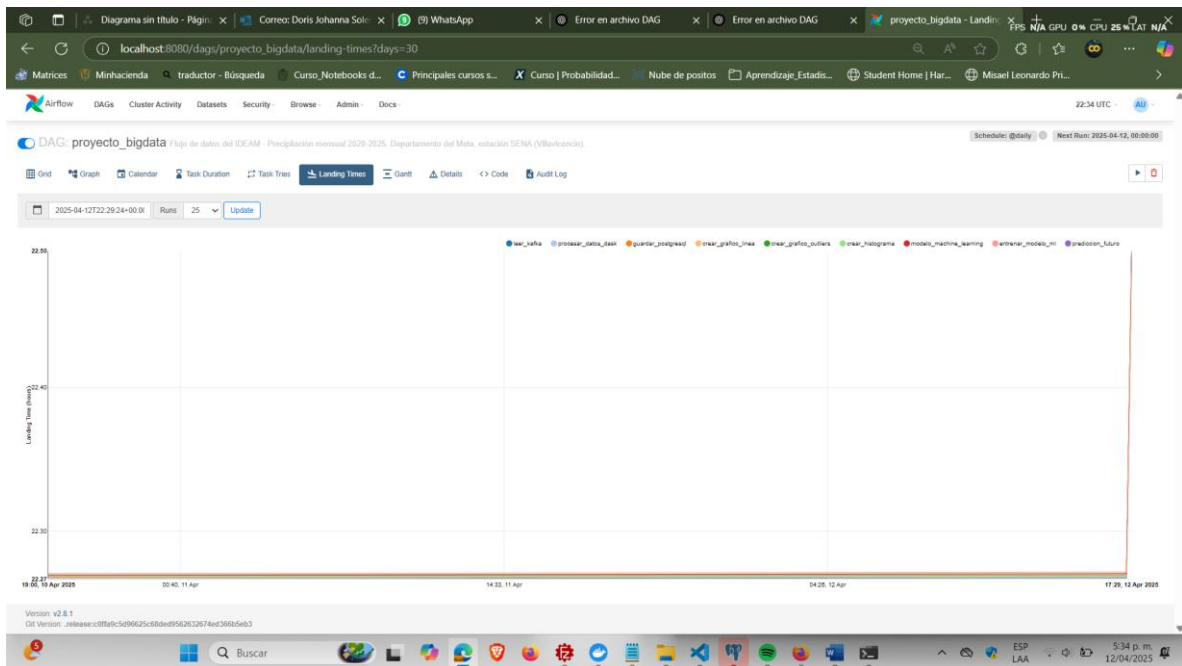
This screenshot shows the main 'DAGs' view of the Apache Airflow web interface. The browser address bar indicates the URL is `localhost:8080/home`. The interface includes a top navigation bar with links to 'DAGs', 'Cluster Activity', 'Datasets', 'Security', 'Browse', 'Admin', and 'Docs'. Below the navigation bar, there are filters for 'All', 'Active', and 'Paused' DAGs, along with a 'Filter DAGs by tag' input and a 'Search DAGs' search bar. The main table lists DAGs with columns for 'DAG', 'Owner', 'Runs', 'Schedule', 'Last Run', 'Next Run', 'Recent Tasks', 'Actions', and 'Links'. A single DAG, 'proyecto_bigdata', is listed with a status of 'Running' and a 'Daily' schedule. The bottom of the interface shows 'Showing 1.1 of 1 DAGs'.

This screenshot shows the detailed view of the 'proyecto_bigdata' DAG in the Airflow web interface. The browser address bar shows the URL `localhost:8080/dags/proyecto_bigdata?tab=graph&dag_run_id=manual_2025-04-12T16%3A32%3A13.783188%2B00%3A00`. The interface includes a top navigation bar with links to 'DAGs', 'Cluster Activity', 'Datasets', 'Security', 'Browse', 'Admin', and 'Docs'. Below the navigation bar, there are filters for 'All', 'Active', and 'Paused' DAGs, along with a 'Filter DAGs by tag' input and a 'Search DAGs' search bar. The main table lists DAGs with columns for 'DAG', 'Owner', 'Runs', 'Schedule', 'Last Run', 'Next Run', 'Recent Tasks', 'Actions', and 'Links'. A single DAG, 'proyecto_bigdata', is listed with a status of 'Running' and a 'Daily' schedule. The bottom of the interface shows 'Showing 1.1 of 1 DAGs'.









proyecto_bigdata - DAG Details | FPS N/A GPU 0% CPU 13% LAT N/A

localhost:8080/dags/proyecto_bigdata/details

Matrices | Minihacienda | traductor - Búsqueda | Curso_Notebooks d... | Principales cursos s... | Curso | Probabilidad... | Nube de positos | Aprendizaje_Estadis... | Student Home | Har... | Misael Leonardo PH...

Airflow | DAGs | Cluster Activity | Datasets | Security | Browse | Admin | Docs

16:53 UTC

DAG: proyecto_bigdata | Flujo de datos del IDEAM - Precipitación mensual 2020-2025, Departamento del Meta, estación SENA (Villavicencio)

Schedule: @daily | Next Run: 2025-04-12, 00:00:00

Grid | Graph | Calendar | Task Duration | Task Titles | Landing Times | Gantt | Details | <> Code | Audit Log

DAG Details

failed 1 | upstream_failed 12

Schedule Interval	@daily
Catchup	False
Started	True
End Date	None
Max Active Runs	0 / 16
Max Active Tasks	16
Default Args	{'depends_on_past': False, 'owner': 'airflow', 'retries': 1, 'retry_delay': 'datetime.timedelta(seconds=300)}
Tasks Count	7
Task IDs	['leer_kafka', 'procesar_datos_task', 'guardar_postgresql', 'crear_grafico_linea', 'crear_historgramas', 'modelo_machine_learning', 'entrenar_modelo_ml']
Relative file location	proyecto_bigdata_dag.py
Owner	airflow
Owner Links	None
DAG Run Timeout	None
Tags	bigdata

DagModel debug information

Attribute	Value
30°C	Mayorm. soleado

11:53 a. m.
12/04/2025

Diagrama sin título - Página: x Correo: Doris Johanna Sol: x WhatsApp x Error en archivo DAG x Error en archivo DAG x proyecto_bigdata - DAG 1 x FPS N/A GPU 0% CPU 21% LAT N/A

localhost:8080/dags/proyecto_bigdata/details

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 22:35 UTC

DAG: proyecto_bigdata Flujo de datos del ICEAM - Principación mensual 2020-2025, Departamento del Meta, estación SENA (Villavicencio)

Schedule: @daily Next Run: 2025-04-12, 00:00:00

Grid Graph Calendar Task Duration Task Times Landing Times Gantt Details <> Code Audit Log

DAG Details

success

Schedule Interval	@daily
Catchup	False
Started	True
End Date	None
Max Active Runs	0 / 16
Max Active Tasks	16
Default Args	{'depends_on_past': False, 'owner': 'airflow', 'ymls': 1, 'ymls_delay': datetime.timedelta(seconds=300)}
Tasks Count	9
Task IDs	['beet_hafur', 'procesar_datos_dash', 'guardar_postgresql', 'crear_grafico_lineal', 'crear_grafico_outliers', 'crear_histograma', 'modelos_machine_learning', 'entrenar_modelos_ml', 'prediccion_futuro']
Relative file location	proyecto_bigdata_dag.py
Owner	airflow
Owner Links	None
DAG Run Timeout	None
Tags	tags

DagModel debug information

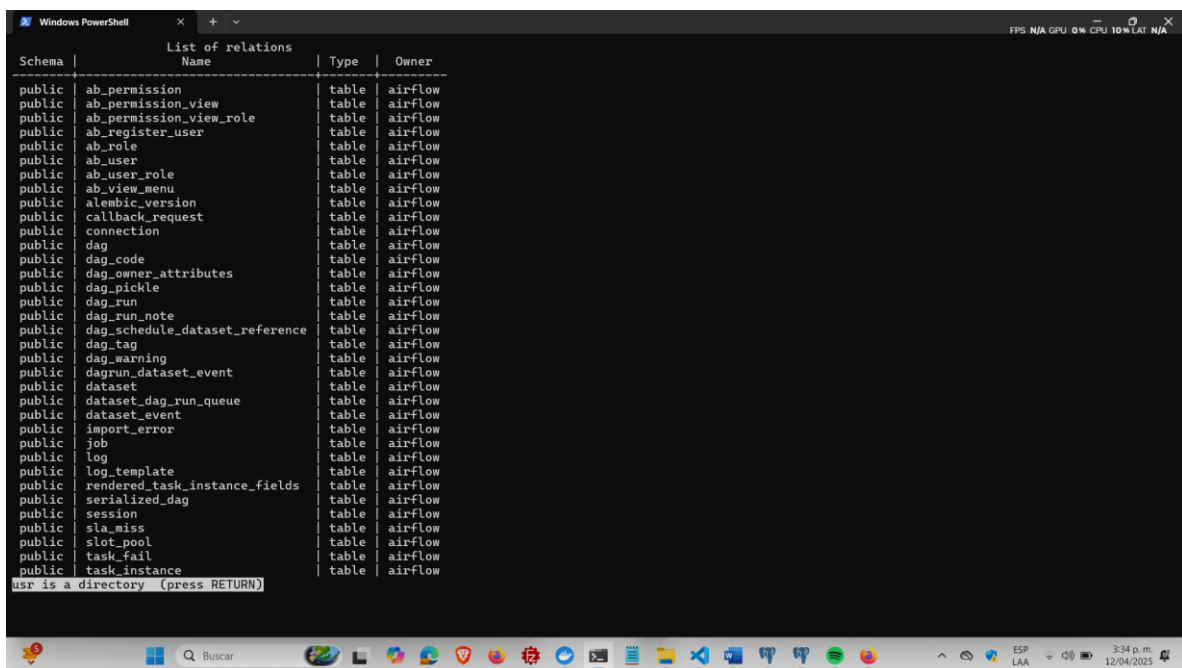
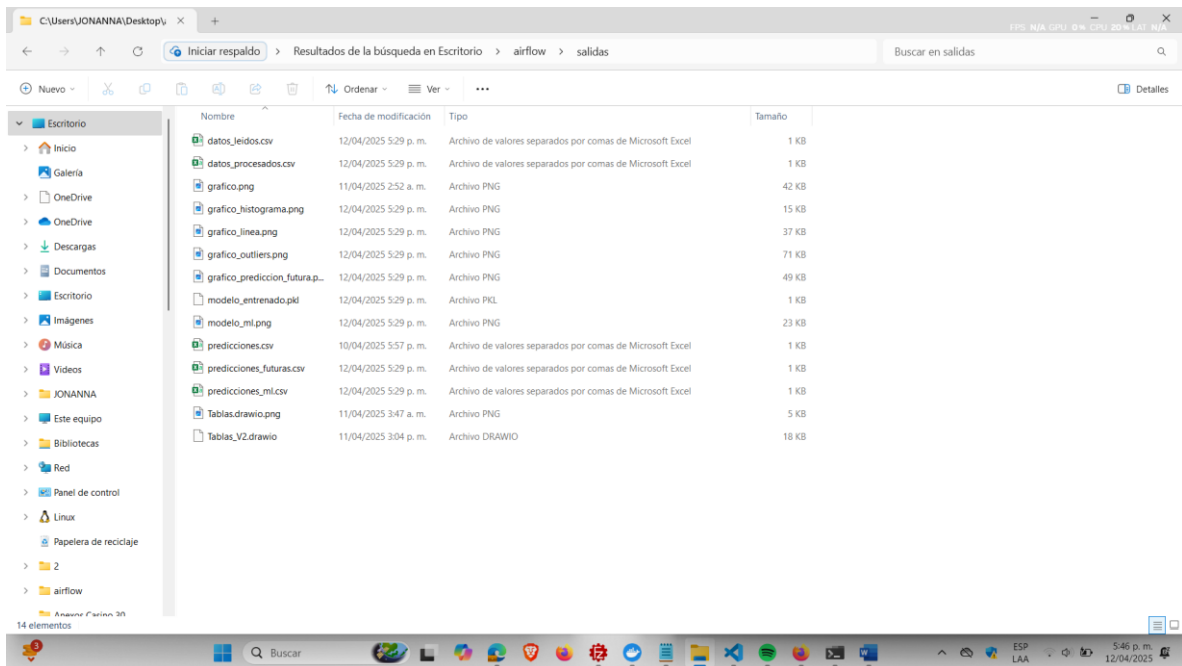
Attribute	Value
fileloc	/opt/airflow/dags/proyecto_bigdata_dag.py
has_import_errors	False
has_task_concurrency_limits	False

Inicio respaldo > Resultados de la búsqueda en Escritorio > airflow > salidas

Buscar en salidas

Ordenar Ver

Nombre	Fecha de modificación	Tipo	Tamaño
datos_leidos.csv	11/04/2025 2:52 a. m.	Archivo de valores separados por comas de Microsoft Excel	1 KB
datos_procesados.csv	11/04/2025 2:52 a. m.	Archivo de valores separados por comas de Microsoft Excel	1 KB
grafico.png	11/04/2025 2:52 a. m.	Archivo PNG	42 KB
predicciones.csv	10/04/2025 5:57 p. m.	Archivo de valores separados por comas de Microsoft Excel	1 KB
Tablas.drawio.png	11/04/2025 3:47 a. m.	Archivo PNG	5 KB
Tablas_V2.drawio	11/04/2025 3:04 p. m.	Archivo DRAWIO	18 KB



**ARCHIVOS: PROYECTO_BIGDATA_DAG.PY, MODELO_ML.PY, DATOS, GRÁFICOS Y
PREDICCIONES**