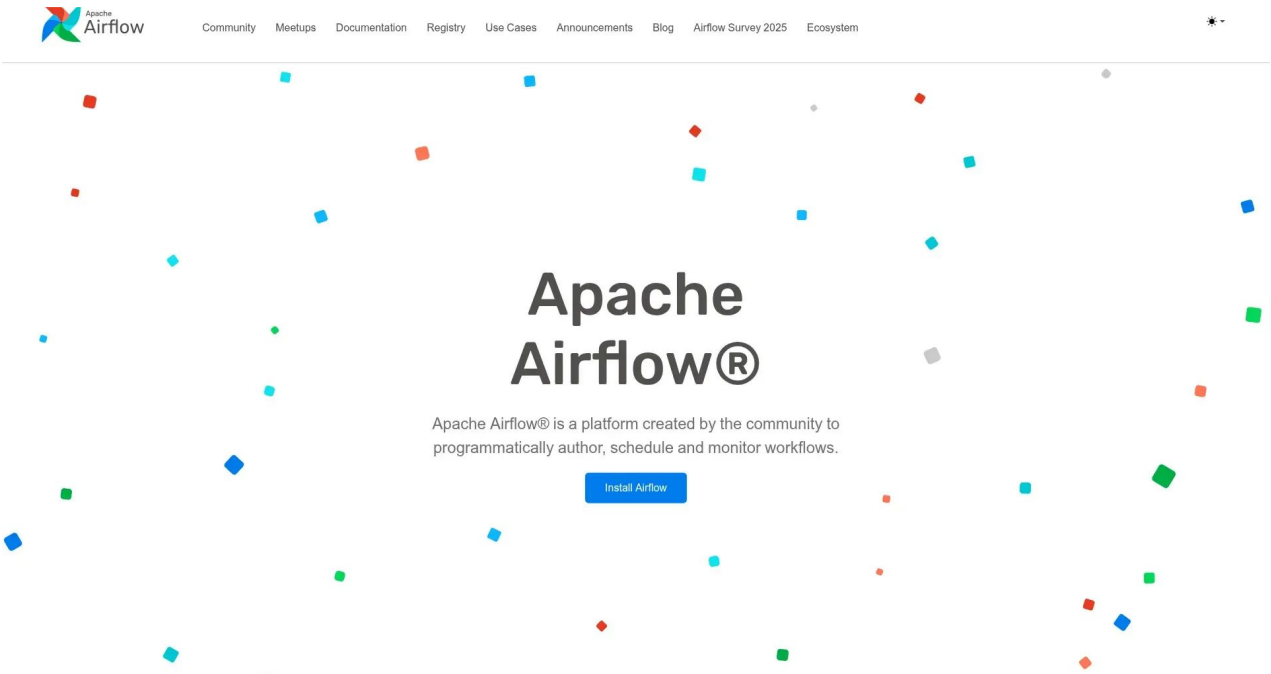


# Apache Airflow

*Plataforma de código abierto diseñada para ingenieros de datos, científicos de datos y arquitectos que necesitan orquestar flujos de trabajo complejos. Permite definir procesos mediante scripts de Python bajo el concepto de Configuration as Code, facilitando la gestión de dependencias, la automatización de procesos ETL/ELT y la monitorización de tuberías de información. Es ideal para equipos técnicos que buscan aplicar prácticas de ingeniería de software como control de versiones y CI/CD en sus datos.*

Visitar Sitio Oficial Preguntar a ChatGPT Preguntar a Claude Preguntar a Grok

## Contenido del Dossier

## INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es
Apache Airflow es una plataforma de código abierto diseñada para programar, orquestar y monitorizar flujos de trabajo (workflows) complejos. Su núcleo se basa en el concepto de "Configuration as Code", lo que permite definir procesos mediante scripts de Python. Está dirigida a ingenieros de datos, científicos de datos y arquitectos de soluciones que necesitan gestionar dependencias entre tareas, automatizar procesos de ingesta y transformación de datos (ETL/ELT) y asegurar la fiabilidad de sus tuberías de información.

Principal ventaja profesional
La capacidad de definir flujos de trabajo como código puro en Python (DAGs). Esto permite aplicar las mejores prácticas de ingeniería de software a los datos: control de versiones (Git), pruebas unitarias, revisiones de código e integración continua (CI/CD), ofreciendo una flexibilidad y escalabilidad casi ilimitadas frente a herramientas visuales rígidas.

Para quién no es
No es una herramienta adecuada para perfiles puramente de negocio o analistas que no posean conocimientos de programación en Python. Tampoco es recomendable para empresas con flujos de trabajo extremadamente simples o lineales que puedan resolverse con scripts básicos de cron, ni para organizaciones que busquen una solución "zero-code" sin capacidad de mantenimiento de infraestructura o código.

Funcionalidades clave
- **DAGs (Directed Acyclic Graphs):** Representación de flujos de trabajo donde se definen las dependencias y el orden de ejecución de las tareas.
- **Planificador (Scheduler):** Orquestador que dispara las tareas en el momento preciso según las dependencias y el calendario definido.
- **Interfaz de Usuario Web:** Panel de monitorización para visualizar el estado de los flujos, revisar logs, reintentar tareas fallidas y gestionar variables.
- **Escalabilidad: Sistema basado en ejecutores (Celery, Kubernetes, Local) que permite distribuir tareas en múltiples nodos o clústeres.**
- **Manejo de errores y reintentos:** Configuración automática de reintentos con políticas de espera (backoff) y alertas integradas.
- **Backfilling:** Capacidad de ejecutar flujos de trabajo retroactivamente sobre datos históricos de forma sencilla.

Precios
- **Versión gratuita (Open Source):** Apache Airflow es un proyecto de la Apache Software Foundation totalmente gratuito bajo licencia Apache 2.0. El coste es de infraestructura y mantenimiento por parte de la empresa.
- **Managed Services (SaaS/PaaS):** Existen versiones gestionadas por proveedores de nube que eliminan la carga de administración.
- **Astronomer:** Plataforma comercial con soporte empresarial y optimizaciones sobre Airflow (precios bajo presupuesto o pago por uso).
- **Amazon MWAA / Google Cloud Composer:** Servicios gestionados en la nube con costes variables basados en el tamaño de la instancia y horas de uso (ej. desde ~0.49€/hora para entornos pequeños en AWS).

Perfil del usuario
- **Empresas con grandes volúmenes de datos:** Sectores como el financiero, e-commerce, telco y tecnología.
- **Departamentos de Data Engineering y Machine Learning Operations (MLOps).**
- **Sectores con alta dependencia de reporting:** Business Intelligence y Analytics.
- **Profesionales:** Data Engineers, Data Scientists, Backend Developers y Ops/SREs.

Nivel técnico requerido
- **Para su uso:** Alto. Requiere dominio de Python para definir los DAGs y lógica de negocio.
- **Para su instalación/configuración:** Alto. Exige conocimientos de administración de sistemas, Docker, bases de datos (PostgreSQL/MySQL) y, preferiblemente, Kubernetes para despliegues a escala.
- **Soporte necesario:** Departamentos de IT/DevOps para la gestión de la infraestructura y seguridad.

- **Competencias necesarias:** Python, SQL, gestión de APIs y conceptos de computación distribuida.

Ejemplos de uso profesional
- **ETL/ELT Automatizado:** Extracción de datos de múltiples APIs (SaaS), transformación en un Data Warehouse (Snowflake, BigQuery, Redshift) y carga de resultados.
- **Entrenamiento de Modelos de ML:** Orquestación de la limpieza de datos, entrenamiento del modelo en clústeres externos y posterior despliegue en producción.
- **Auditoría y Compliance:** Ejecución programada de procesos de validación de calidad de datos y generación de reportes regulatorios.
- **Gestión de Infraestructura:** Disparo de tareas de mantenimiento en la nube, como creación de snapshots o limpieza de recursos temporales.

Uso y distribución
- **Versión web:** Interfaz de control accesible vía navegador una vez desplegado.
- **Versión escritorio:** No dispone de app nativa (uso vía navegador).
- **Versión móvil:** No dispone de app oficial, aunque la web es responsive.
- **CLI:** Potente línea de comandos para gestión de tareas, usuarios y configuración del sistema.

Open Source
Apache Airflow es un proyecto Apache Top-Level, lo que garantiza transparencia, una comunidad masiva y ausencia de "vendor lock-in".

Integraciones
- **Facilidad de integración:** Full code. Se integra mediante "Providers" y "Hooks".
- **API propia:** Dispone de una API REST completa para interactuar con los DAGs de forma externa.
- **Ecosistema nativo:** Cuenta con más de 100 proveedores oficiales para conectar con AWS, Google Cloud, Azure, Slack, Salesforce, Snowflake, dbt, Spark, etc.
- **Ejemplos concretos:** Integración con Kubernetes para ejecutar tareas en contenedores aislados o con Slack para recibir notificaciones inmediatas de fallos en producción.

Notas finales
Información legal, licencias, contratos
Se distribuye bajo la **Licencia Apache 2.0**, que permite el uso comercial, modificación y distribución sin coste de royalties. Los usuarios son dueños de sus DAGs y la propiedad intelectual de sus desarrollos.
Otros
Actualmente, Airflow es el estándar de la industria en orquestación de datos. La versión 3.0 (recientemente anunciada) introduce mejoras críticas en rendimiento y usabilidad.

Para más información:
- Sitio web oficial: https://airflow.apache.org
- Documentación técnica: https://airflow.apache.org/docs
- Github: https://github.com/apache/airflow
- Ecosistema de integraciones: https://airflow.apache.org/docs/apache-airflow-providers
- Slack de la comunidad: https://s.apache.org/airflow-slack

## CONSEJOS DE IMPLANTACIÓN

Aplicación profesional
Apache Airflow se aplica en empresas tecnológicas, fintech, retail y sectores con arquitecturas de datos complejas. Es el estándar para la gestión de tuberías de datos (pipelines) que requieren alta disponibilidad y trazabilidad. El presupuesto varía desde el coste cero de licencia (Open Source) hasta servicios gestionados que oscilan entre 400€ y varios miles de euros mensuales según el volumen de tareas. Los puntos clave de su aplicación son la automatización de procesos ETL/ELT, la orquestación de modelos de Machine Learning y la sincronización de datos entre nubes heterogéneas.

Madurez digital requerida
- **Usuarios y equipo**: Se requiere un equipo con dominio avanzado de Python y principios de ingeniería de software (control de versiones con Git, entornos virtuales). Los usuarios deben entender la arquitectura de microservicios y contenedores.
- **Empresa y departamentos**: La organización debe poseer una estructura de datos clara y la necesidad de escalar procesos que los scripts manuales o las herramientas visuales ya no pueden gestionar eficientemente. Es fundamental contar con una cultura de CI/CD para el despliegue de flujos de trabajo.

Plan orientativo de implantación
Pasos necesarios y estimaciones
- **Evaluación y diseño (1-2 semanas)**: Análisis de los flujos de trabajo actuales e identificación de dependencias. Definición de la arquitectura (Local, Celery o Kubernetes Executor).
- **Prueba de Concepto (2-3 semanas)**: Instalación de un entorno de desarrollo para validar la conectividad con las fuentes de datos y el almacenamiento de metadatos (PostgreSQL).
- **Configuración y Despliegue (3-4 semanas)**: Configuración del entorno de producción, integración con sistemas de autenticación (LDAP/OAuth) y establecimiento de políticas de seguridad y logs.
- **Migración y formación (4-8 semanas)**: Codificación de los primeros DAGs, migración de procesos antiguos y capacitación técnica del equipo de datos en las librerías específicas de Airflow.
- **Estabilización (Continuo)**: Monitorización de rendimiento, ajuste de recursos en el Scheduler y Workers, y optimización de tiempos de ejecución mediante paralelismo.

Necesidades de formación del equipo
El equipo debe recibir formación específica en la arquitectura de Airflow (DAGs, Operators, Hooks y Sensors). Es crítico el aprendizaje sobre la gestión del estado de las tareas y el manejo de XComs para el paso de mensajes. Se recomienda formación en Docker y Kubernetes si se opta por despliegues escalables.

Perfiles necesarios
- **Perfiles técnicos necesarios**: Data Engineers (desarrollo de DAGs), DevOps/SRE (mantenimiento de la infraestructura y escalabilidad), Cloud Architects (configuración de red y permisos).
- **Personal externo recomendado**: Consultores expertos en arquitectura de datos para la configuración inicial y optimización del Scheduler en entornos de alta carga.

Retorno de la inversión
- **Tiempos**: Reducción de hasta un 70% en el tiempo dedicado a la resolución de fallos manuales y un 40% en el tiempo de puesta en producción de nuevos procesos de datos.
- **KPIs**: Tasa de éxito de DAGs, tiempo medio de recuperación ante fallos (MTTR), coste de infraestructura por tarea ejecutada y reducción de la deuda técnica en integraciones.

Otros
- **Seguridad**: Soporta integración con secretos de AWS, Google Secrets Manager y Hashicorp Vault, evitando el almacenamiento de credenciales en el código.
- **Versión 3.0**: Se enfoca en la modernización de la interfaz de usuario y una mejora significativa en la latencia del Scheduler, eliminando cuellos de botella en ejecuciones de alta frecuencia.

## PREGUNTAS FRECUENTES

### ¿Qué es Apache Airflow y cuál es su función principal?

Es una plataforma de código abierto diseñada para programar, orquestar y monitorizar flujos de trabajo complejos. Su función principal es permitir la creación de procesos de datos mediante código Python, organizados en Grafos Acíclicos Dirigidos (DAGs), lo que facilita la gestión de dependencias y la automatización de tareas ETL/ELT.

### ¿Es Apache Airflow una herramienta Open Source?

Sí, es un proyecto de la Apache Software Foundation bajo la licencia Apache 2.0. Esto garantiza que el software es gratuito para uso comercial, modificación y distribución, permitiendo a las organizaciones evitar el bloqueo por parte de proveedores (vendor lock-in).

### ¿Puedo descargarlo de GitHub?

Efectivamente, el código fuente completo, el historial de versiones y la documentación técnica están disponibles en el repositorio oficial de la organización Apache en GitHub. Esto permite auditorías de seguridad, contribuciones de la comunidad y despliegues personalizados.

### ¿Cuál es el coste de implementar Airflow en un entorno profesional?

La herramienta en sí no tiene coste de licencia. Sin embargo, existen costes asociados a la infraestructura y al mantenimiento. Si se opta por servicios gestionados como Amazon MWAA o Google Cloud Composer, los costes son variables según el uso de recursos, mientras que plataformas como Astronomer ofrecen soporte empresarial bajo presupuesto.

### ¿Qué nivel técnico se requiere para manejar la plataforma?

El nivel técnico requerido es alto. Para el desarrollo de flujos se necesita dominio de Python y SQL. Para la instalación y administración de la infraestructura, es esencial tener conocimientos avanzados en Docker, bases de datos relacionales como PostgreSQL y, preferiblemente, experiencia con orquestadores de contenedores como Kubernetes.

### ¿Cumple con la normativa española y europea de privacidad?

Al ser una herramienta de orquestación autohospedada o gestionada en la nube, el cumplimiento del RGPD y otras normativas españolas depende de cómo la empresa configure el despliegue y gestione los datos que fluyen por sus tuberías. Airflow permite una gestión detallada de logs y auditorías, lo que facilita el cumplimiento normativo si se implementa correctamente.

### ¿Cómo garantiza Airflow la seguridad de los datos?

Airflow hereda las capas de seguridad de la infraestructura donde se despliega. Ofrece mecanismos de autenticación (LDAP, OAuth), control de acceso basado en roles (RBAC) para la interfaz web, gestión de secretos encriptados para conexiones externas y soporte para el aislamiento de tareas mediante contenedores, minimizando riesgos de seguridad en la ejecución.

### ¿Para qué sirve la funcionalidad de 'Backfilling'?

Es una característica que permite ejecutar flujos de trabajo de manera retroactiva sobre periodos de tiempo pasados. Es fundamental cuando se necesita procesar datos históricos tras haber modificado la lógica de un DAG o tras recuperar el sistema después de una caída prolongada.

### ¿Qué integraciones ofrece con otros servicios?

Airflow cuenta con una arquitectura extensible basada en 'Providers' que incluye más de 100 integraciones oficiales. Permite conectar de forma nativa con servicios cloud (AWS, Azure, GCP), bases de datos (Snowflake, BigQuery), herramientas de transformación (dbt), sistemas de mensajería (Slack) y plataformas de contenedores (Kubernetes).

### ¿Existe una versión para escritorio o aplicación móvil?

No dispone de aplicaciones nativas de escritorio o móviles. Todo el control se realiza a través de una interfaz de usuario web responsive accesible desde el navegador o mediante su potente interfaz de línea de comandos (CLI) para tareas de administración y automatización interna.

## CONTRATOS Y CONDICIONES

Principales recomendaciones
- Implementar una política de clasificación de datos antes de integrarlos en los flujos de trabajo (DAGs), ya que Airflow actúa como orquestador y puede procesar información sensible.
- En caso de usar versiones gestionadas (Cloud Composer, MWAA), es imperativo firmar un Acuerdo de Encargo de Tratamiento (DPA) con el proveedor de servicios en la nube para cumplir con el RGPD.
- Configurar el sistema de telemetría (introducido en versiones recientes como la 3.0) en modo "opt-out" o desactivarlo si las políticas corporativas prohíben el envío de métricas de uso a la Apache Software Foundation.
- Asegurar que las "Connections" y "Variables" que contengan secretos (passwords, tokens de API) estén encriptadas en la base de datos de metadatos utilizando una clave de encriptación fuerte (Fernet key).
- Establecer controles de acceso basados en roles (RBAC) para limitar quién puede ver los logs de las tareas, ya que estos pueden contener accidentalmente datos personales o sensibles durante el proceso de depuración.

Ley de Inteligencia Artificial (AI Act)
- Airflow se clasifica generalmente como una herramienta de infraestructura o "sistema de propósito general" sin riesgo inherente.
- Si Airflow se utiliza para orquestar el entrenamiento o despliegue de modelos de IA en sectores críticos (salud, infraestructuras, recursos humanos/reclutamiento), el flujo de trabajo completo podría ser considerado de **Alto Riesgo**.
- En casos de alto riesgo, los registros (logs) y el historial de ejecución de Airflow son fundamentales para cumplir con las obligaciones de trazabilidad y documentación técnica exigidas por la Ley de IA.

Privacidad y protección de datos
- **Responsabilidades:** Al ser software de código abierto, la Apache Software Foundation no actúa como procesador de datos. La empresa que instala y ejecuta Airflow es la única responsable del tratamiento de los datos (Responsable del Tratamiento).
- **Ubicación de los datos:** Los datos residen donde se aloje la infraestructura (on-premise o cloud). Para empresas españolas, se recomienda el alojamiento en regiones dentro del Espacio Económico Europeo (EEE).
- **Transferencia internacional:** No existe transferencia internacional de datos por el simple uso del software, a menos que se activen servicios de telemetría o se utilicen proveedores cloud fuera del EEE sin las salvaguardas adecuadas (cláusulas contractuales tipo).
- **Derechos ARCO:** La plataforma debe configurarse para permitir la eliminación o exportación de datos si estos se almacenan temporalmente en logs o bases de datos intermedias orquestadas por Airflow.

Propiedad intelectual
- **Propiedad de los datos:** La propiedad de los datos procesados pertenece íntegramente a la empresa usuaria.
- **Propiedad del resultado:** El código desarrollado (DAGs) y los flujos de trabajo son propiedad intelectual de la empresa que los crea, bajo los términos de sus propios contratos laborales o de servicios.
- **Licencia:** Distribuido bajo la **Apache License 2.0**, que permite el uso comercial, la modificación y la distribución sin pago de regalías, siempre que se conserve el aviso de copyright original.

Usos y prohibiciones
- **Usos admitidos:** Automatización de procesos ETL, orquestación de modelos de ML, gestión de infraestructuras y tareas programadas.
- **Usos prohibidos:** El software no impone restricciones de uso específicas, pero su aplicación en sistemas de vigilancia masiva o puntuación social podría violar normativas superiores (como el AI Act en la UE).

Seguridad y certificaciones
- **Seguridad:** No ofrece seguridad "out-of-the-box" para datos sensibles en tránsito sin una configuración adecuada de TLS/SSL y encriptación de base de datos.
- **Certificaciones:** El software como tal no posee certificaciones SOC2 o ISO 27001, pero los proveedores que ofrecen Airflow gestionado (como Astronomer o AWS) sí cuentan con estas certificaciones para su infraestructura.

Otros
- **Telemetría (AIP-89):** Las versiones más recientes incluyen un sistema de telemetría para mejorar el producto. Es vital revisar la configuración telemetry.enabled en el archivo airflow.cfg para ajustarlo a la política de privacidad de la empresa.

Fuentes consultadas:
- Sitio oficial de Apache Airflow
- AIP-89: Telemetría priorizando la privacidad
- Licencia Apache 2.0
- Declaración de cumplimiento y protección de datos de ASF
- Política de privacidad de la Apache Software Foundation