#youtube data pipeline-EC2 instance setup

Perfect! Now that you’re connected to your EC2 instance, let’s walk through the essential steps to:

✅ Install Java 8  
✅ Install Python, pip, and other dependencies  
✅ Install Apache Airflow  
✅ Set up Spark and configure it for S3  
✅ Deploy your ETL project

**✅ 1. Update your packages**

bash

sudo apt update && sudo apt upgrade -y

**✅ 2. Install Java 8 (needed for Spark)**

bash

sudo apt install openjdk-8-jdk -y

Verify installation:

bash

java -version

**✅ 3. Install Python, pip, and venv**

bash

sudo apt install python3-pip python3-venv -y

Create and activate a virtual environment:

bash

python3 -m venv airflow\_venv

source airflow\_venv/bin/activate

**✅ 4. Install Apache Airflow**

Use constraints to avoid dependency issues:

bash

AIRFLOW\_VERSION=2.8.1

PYTHON\_VERSION="$(python --version | cut -d " " -f 2 | cut -d "." -f 1,2)"

CONSTRAINT\_URL="https://raw.githubusercontent.com/apache/airflow/constraints-${AIRFLOW\_VERSION}/constraints-${PYTHON\_VERSION}.txt"

pip install "apache-airflow==${AIRFLOW\_VERSION}" --constraint "${CONSTRAINT\_URL}"

**✅ 5. Initialize Airflow DB and start webserver/scheduler**

bash

export AIRFLOW\_HOME=~/airflow

airflow db init

airflow users create \

--username admin \

--password admin \

--firstname Bhargava \

--lastname Bekkam \

--role Admin \

--email you@example.com

Start the webserver:

bash

CopyEdit

airflow webserver --port 8080

In another SSH session or use tmux/screen, start the scheduler:

bash

airflow scheduler

**✅ 6. Upload your Spark scripts and DAG**

You can:

* Use **scp** (from local machine):

bash

scp -i "bhargav1-ec2key.pem" spark\_etl.py dag\_etl.py ubuntu@<EC2\_PUBLIC\_DNS>:~/airflow/dags/

* Or use nano, vim, or Git on the EC2 instance to paste files.

**✅ 7. Install Spark**

Download and extract Spark (prebuilt with Hadoop):

bash

wget https://downloads.apache.org/spark/spark-3.4.2/spark-3.4.2-bin-hadoop3.tgz

tar -xvzf spark-3.4.2-bin-hadoop3.tgz

mv spark-3.4.2-bin-hadoop3 spark

Set environment variables (in ~/.bashrc or export manually):

bash

export SPARK\_HOME=~/spark

export PATH=$SPARK\_HOME/bin:$PATH

Install required Python packages:

bash

pip install pyspark pandas

**✅ 8. Set Up S3 Credentials**

Use .aws/credentials file or configure in Spark config like you did in your DAG file.

Once done, you’ll be ready to trigger your DAG from the Airflow UI at:

cpp

http://<EC2\_PUBLIC\_DNS>:8080