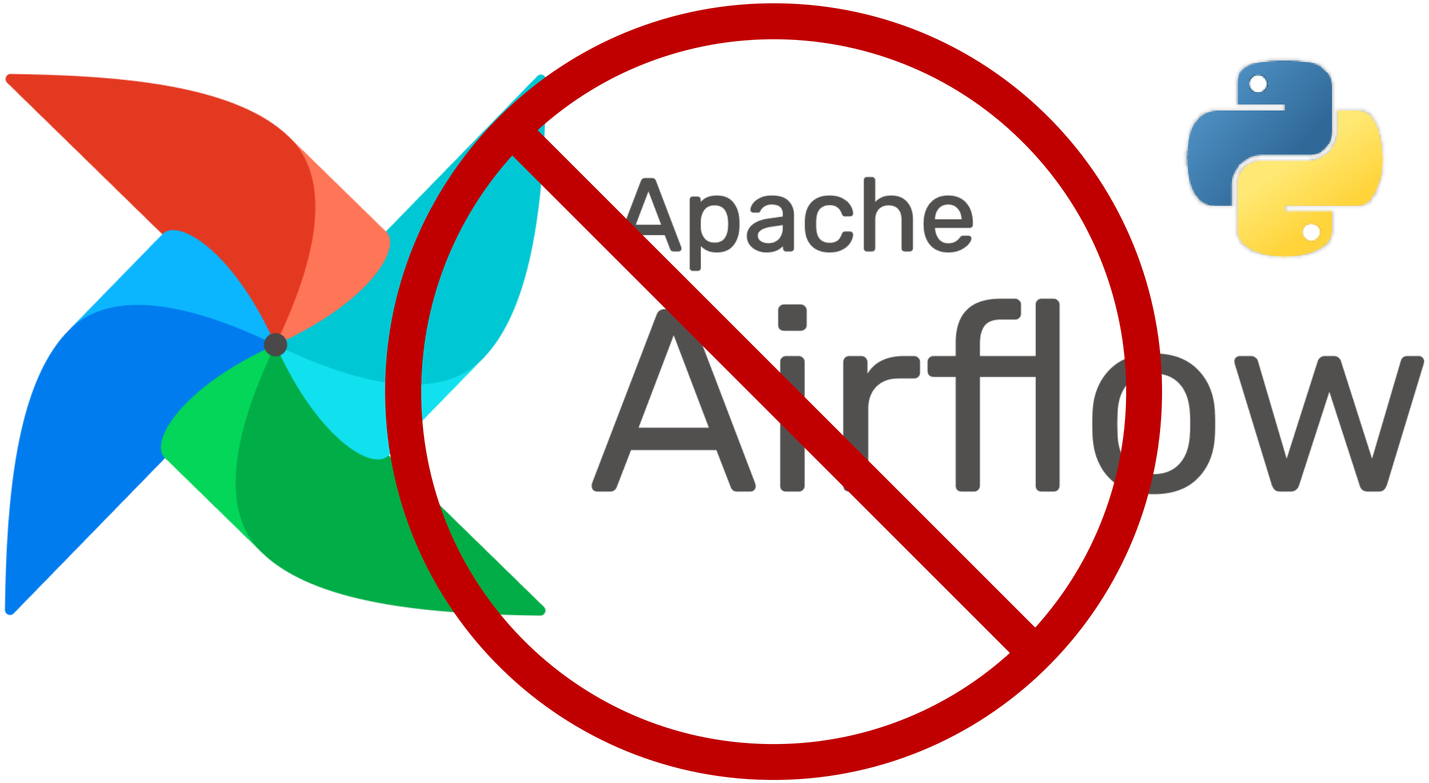
Don't Use



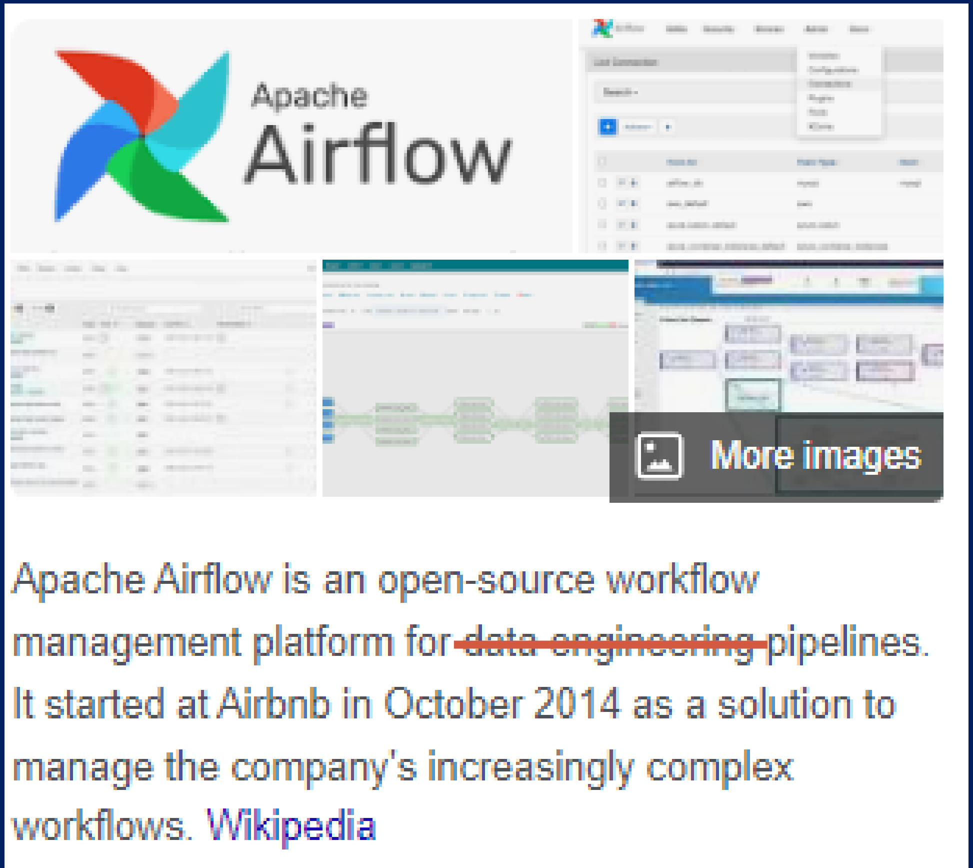
Presentation by Bryan Cafferky, All Rights Reserved

Apache Airflow

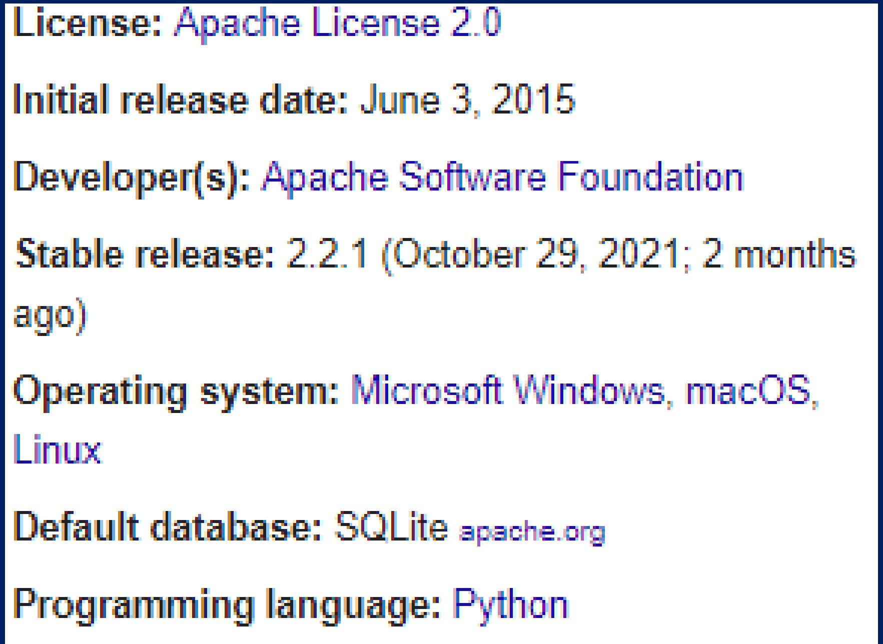
Where Are We Going?

- What is Airflow?
- Airflow's Identity Crisis!
- Limitations
- What Airflow is Good For?
- Better Options

What is Airflow?



Apache Airflow is an open-source workflow management platform for ~~data engineering~~ pipelines. It started at Airbnb in October 2014 as a solution to manage the company's increasingly complex workflows. Wikipedia



License: Apache License 2.0

Initial release date: June 3, 2015

Developer(s): Apache Software Foundation

Stable release: 2.2.1 (October 29, 2021; 2 months ago)

Operating system: Microsoft Windows, macOS, Linux

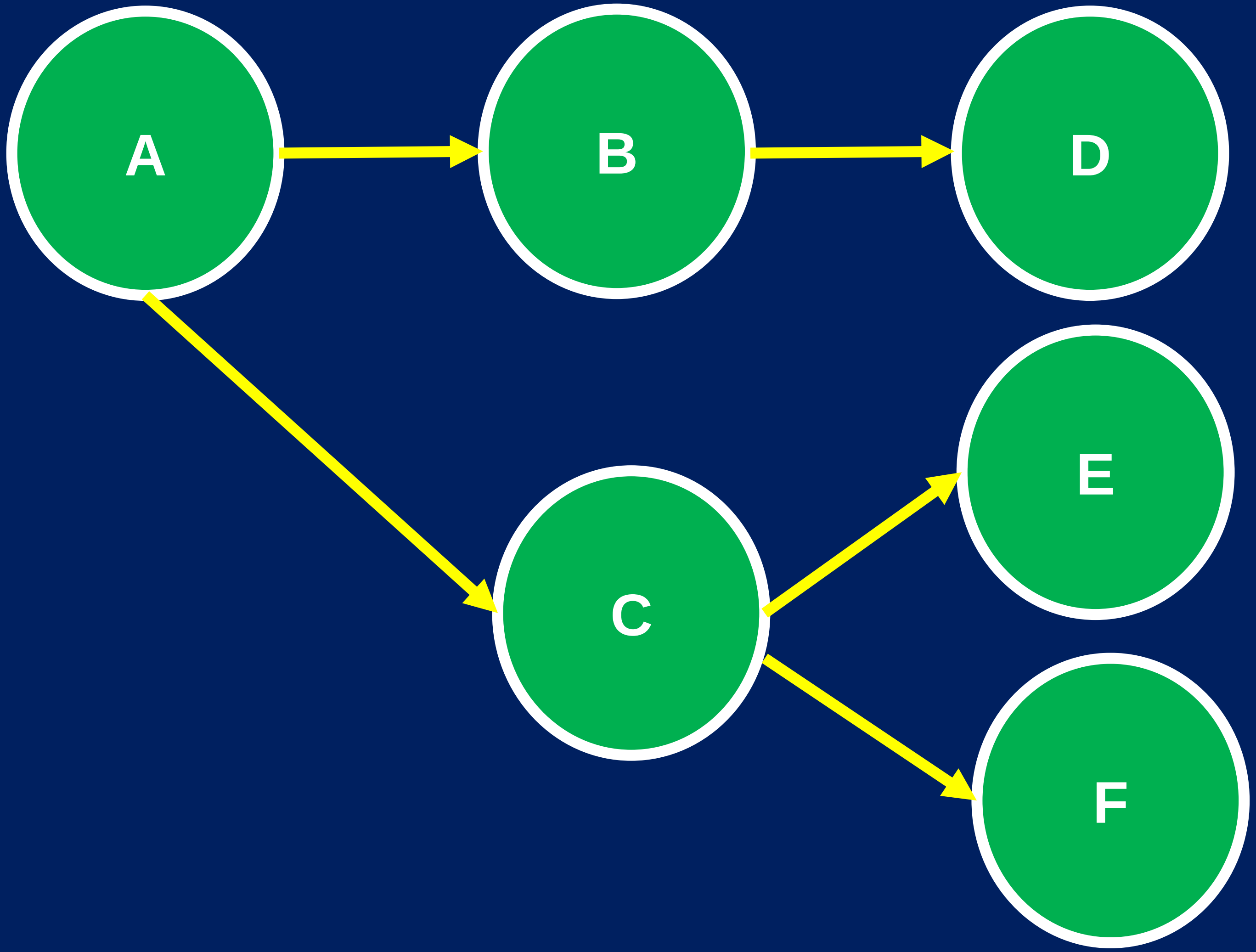
Default database: SQLite apache.org

Programming language: Python

It's a Job Scheduler

Airflow is a DAG Creator

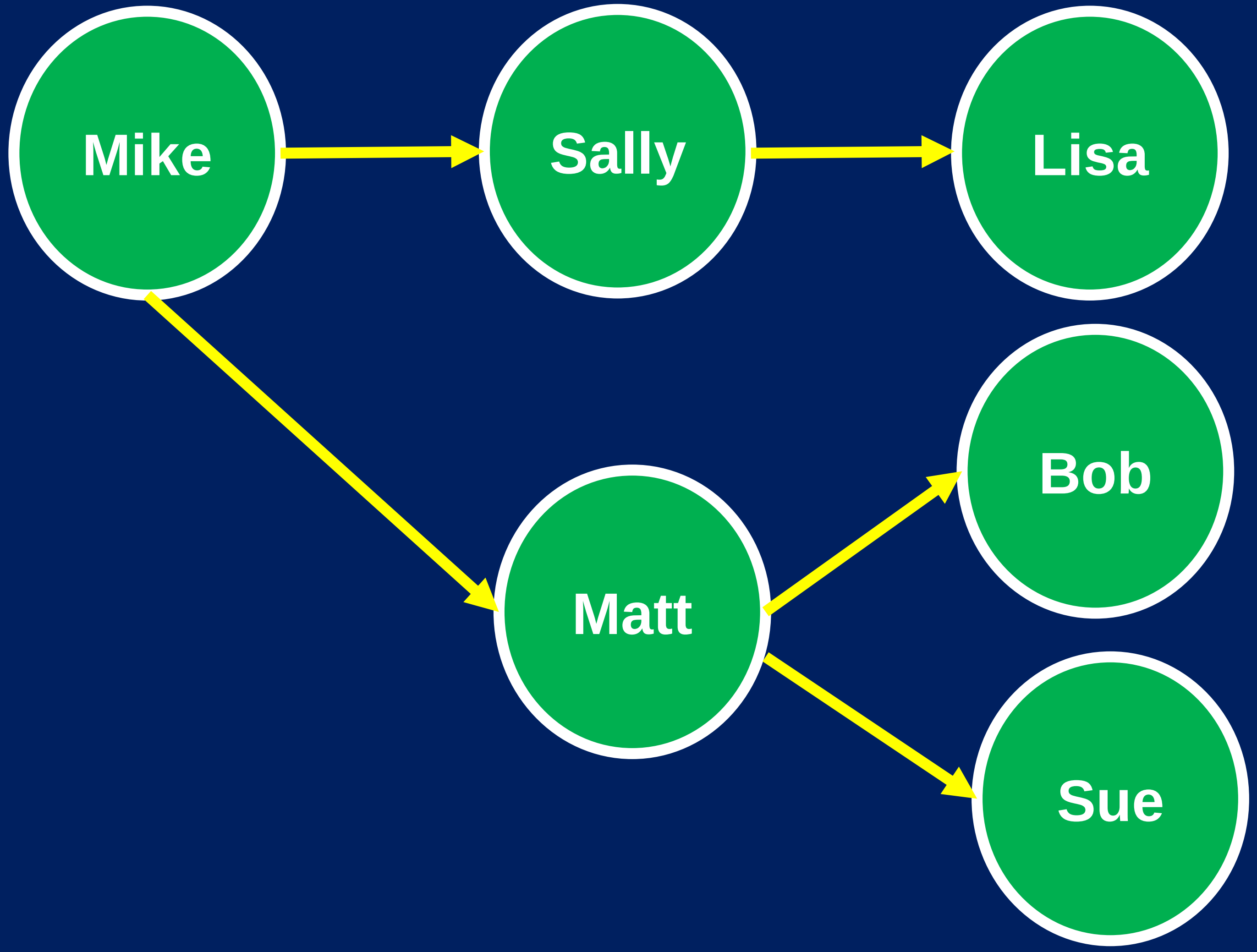
Directed Acyclic Graph



- Vertices (Objects)
- Edges (Relationships)
- Objects and Their Relationships to Other Objects
- Directional (Directed)
- No Loops!!! Acyclic

Apache Airflow

Social Media DAG



- Mike is Connected to Sally and Matt
- Sally is Connected to Lisa
- Matt is Connected to Bob and Sue

Apache Airflow

DAG Applications

- **Apache Spark**
- **Python DASK**
- **Any Graph Analysis/APIs**
- **Machine Learning**
- **Apache Airflow – Job Scheduling**

Apache Airflow

Identity Crisis

- Heavily Promoted as an ETL Service
- CRON on Steroids!
- Can Be Used for ANY Scheduled Work
- Designed to Be an Underlying Service

Solution Looking for a Problem?

Apache Airflow

What is Airflow?

- A Job Scheduler
- A Sophisticated Job Scheduler
- Uses DAGs Which Are Cool!
- A Framework Written in Python for Python

Limitations

Does Not Have Any ETL Functionality!

Requires100% Coding!

Only Supports Python

It Is Extremely Complex/Steep Learning Curve

Requires Its Own Codebase and DevOps

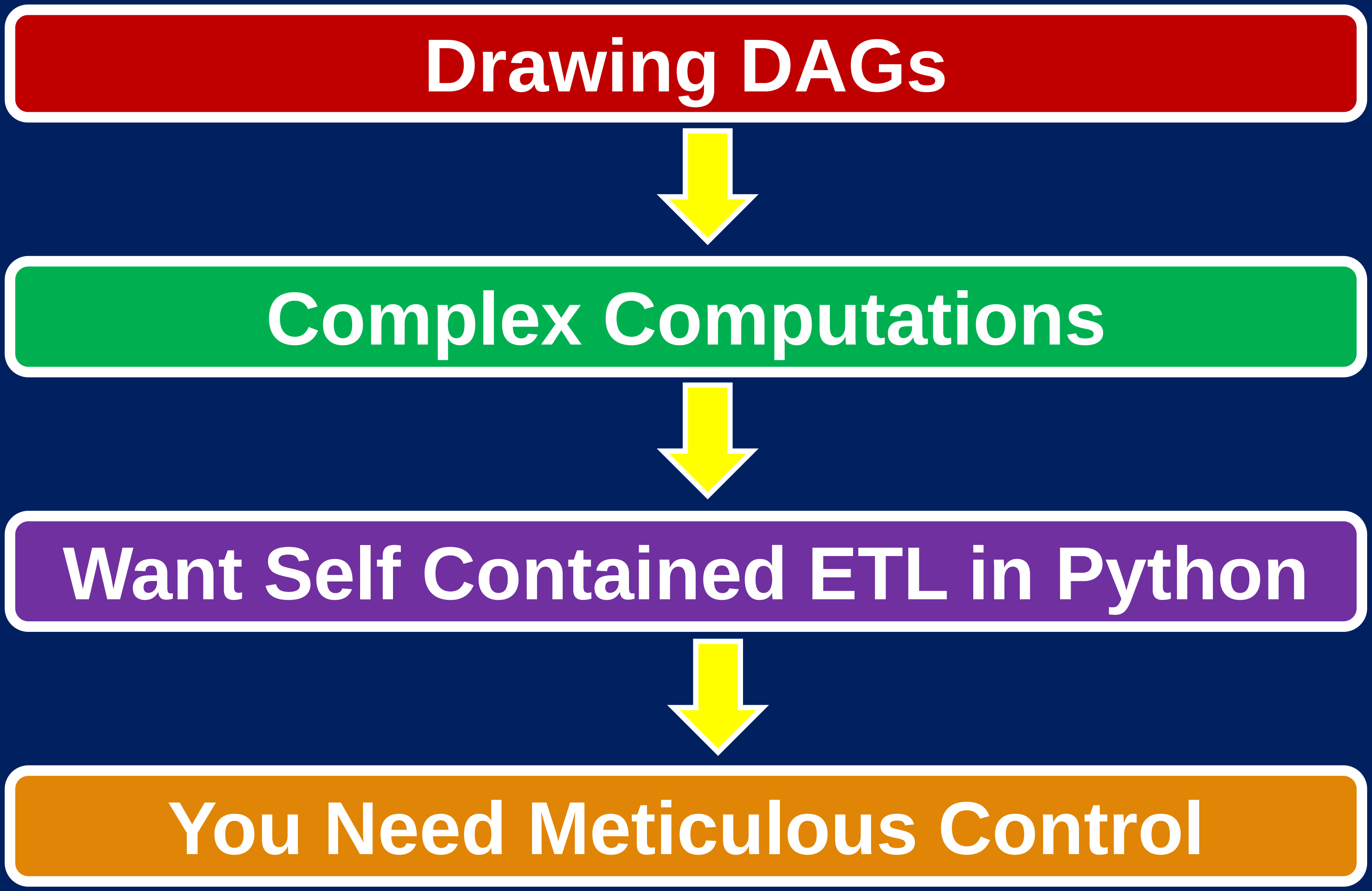
Strong Coupling of Job Scheduler with App Code

Apache Airflow

Other ETL/ELT Options

- Apache NiFi
- AWS Glue/Azure Data Factory/GCP Dataproc
- Kettle by Pentaho
- Databricks Jobs/Notebooks

What is Airflow Good For?



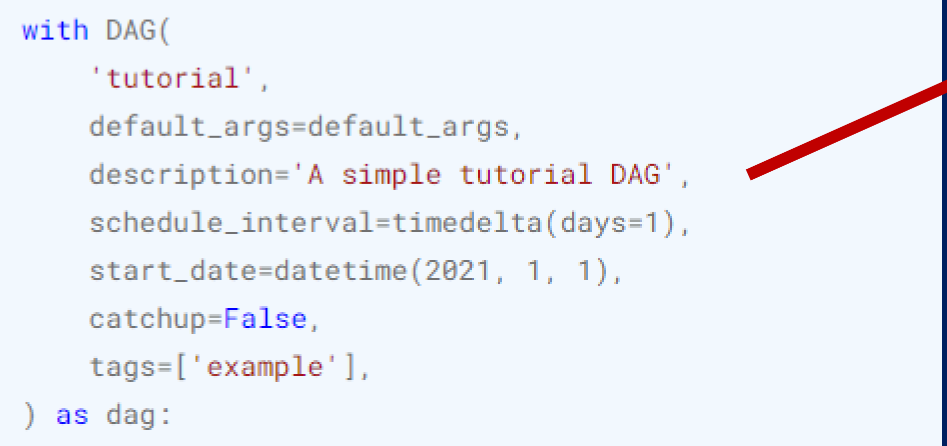
Drawing DAGs

Complex Computations

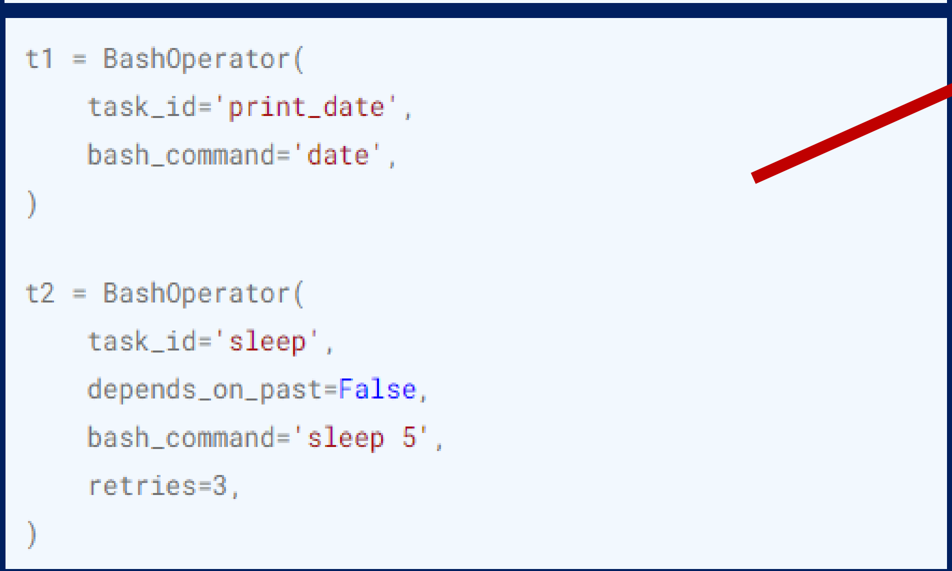
Want Self Contained ETL in Python

You Need Meticulous Control

Airflow Code

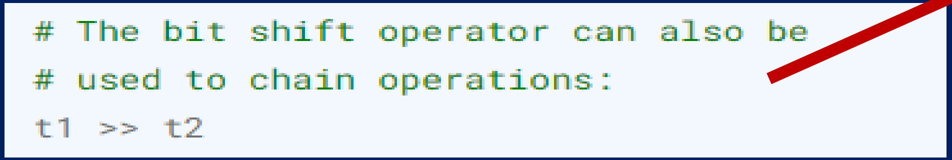


```
with DAG(
    'tutorial',
    default_args=default_args,
    description='A simple tutorial DAG',
    schedule_interval=timedelta(days=1),
    start_date=datetime(2021, 1, 1),
    catchup=False,
    tags=['example'],
) as dag:
```

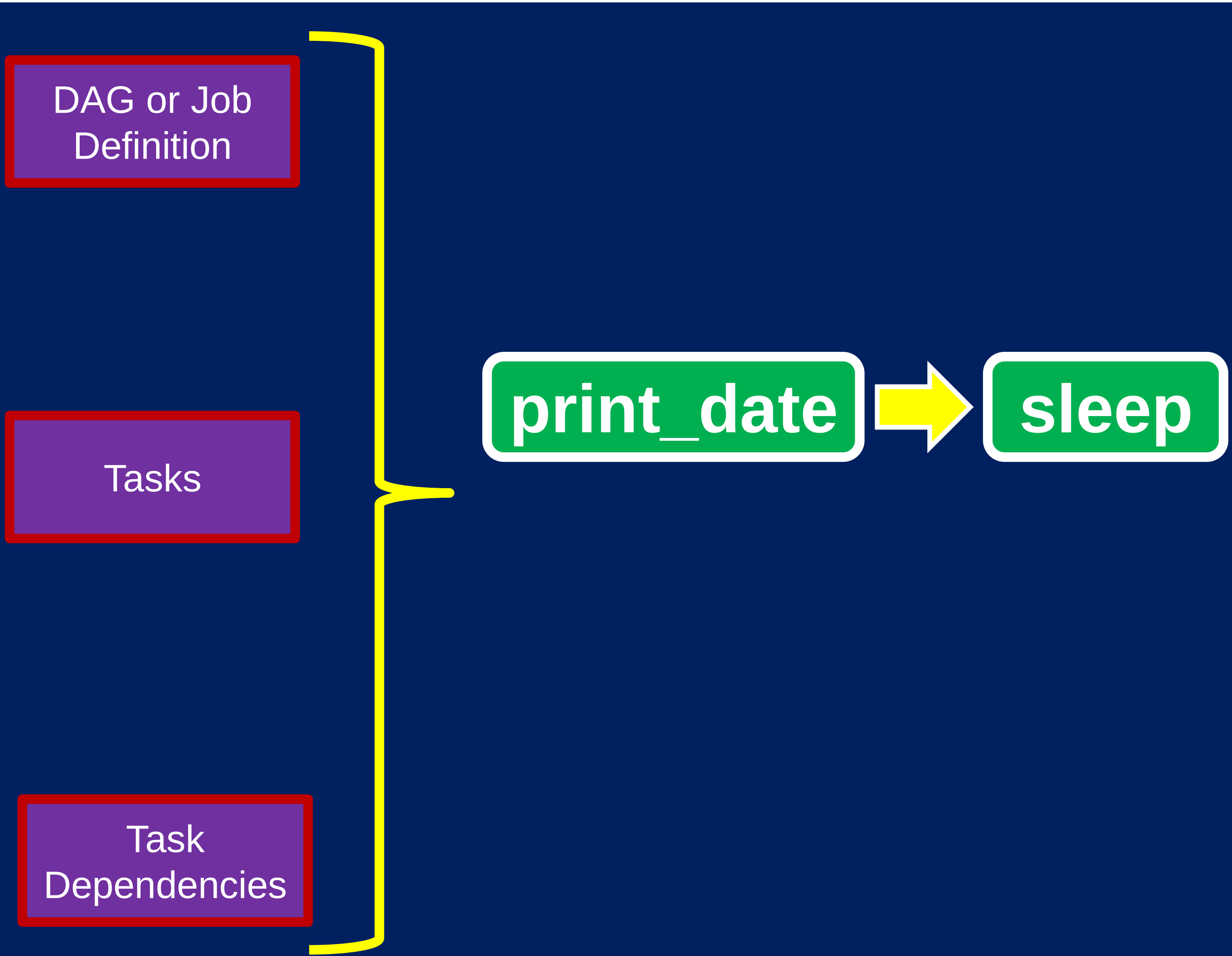


```
t1 = BashOperator(
    task_id='print_date',
    bash_command='date',
)

t2 = BashOperator(
    task_id='sleep',
    depends_on_past=False,
    bash_command='sleep 5',
    retries=3,
)
```



```
# The bit shift operator can also be
# used to chain operations:
t1 >> t2
```



DAG or Job Definition

Tasks

Task Dependencies

Apache Airflow

Wrapping Up

- What is Airflow?
- Airflow's Identit
- Limitations
- What Airflow is Good For?
- Better Options

Thank You!