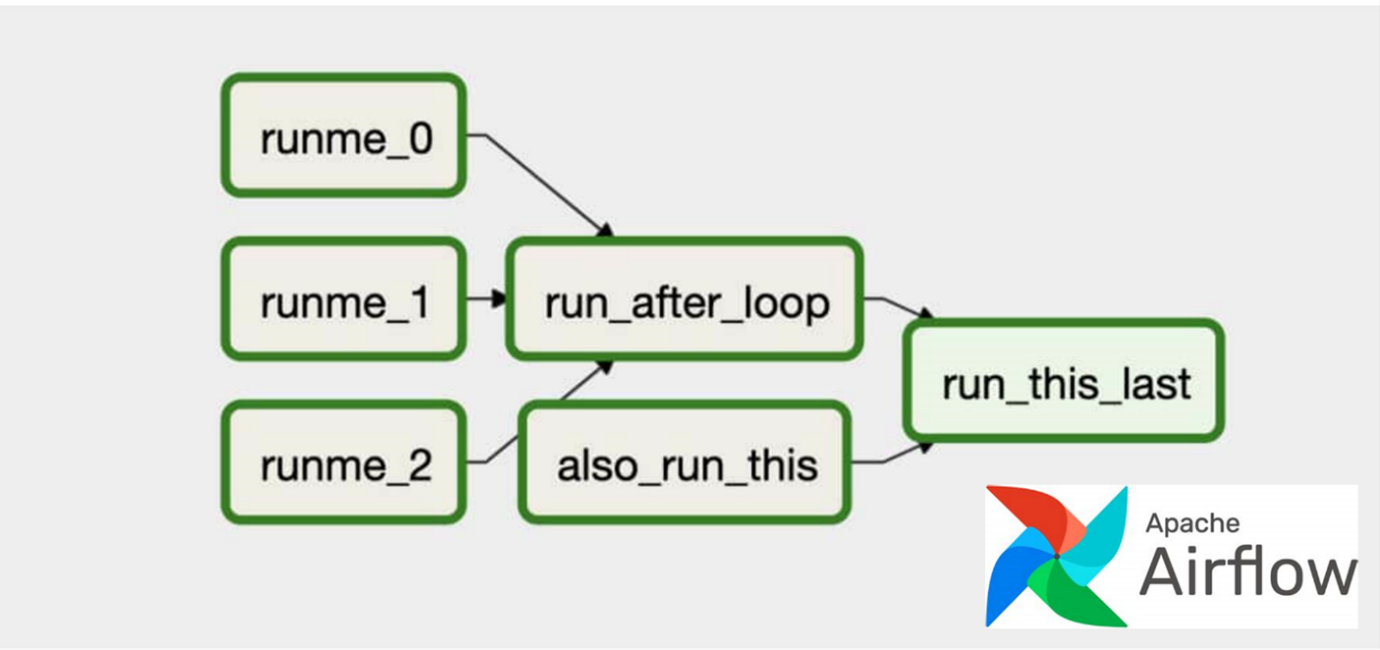
**ETL Workflows as DAGs**

ETL workflows can involve considerable complexity. By breaking down the details of the workflow into individual tasks and dependencies between those tasks, one can gain better control over that complexity. Workflow orchestration tools such as Apache Airflow do just that.

Airflow represents your workflow as a directed acyclic graph (DAG). A simple example of an Airflow DAG is illustrated in Figure 4. Airflow tasks can be expressed using predefined templates, called operators. Popular operators include Bash operators, for running Bash code, and Python operators for running Python code, which makes them extremely versatile for deploying ETL pipelines and many other kinds of workflows into production.



*Fig 4. An Apache Airflow DAG representing a workflow. The green boxes represent individual tasks, while the arrows show dependencies between tasks. The three tasks on the left, ‘runme\_j’ are jobs that run simultaneously along with the ‘also\_run\_this’ task. Once the ‘runme\_j’ tasks complete, the ‘run\_after\_loop’ task starts. Finally, ‘run\_this\_last’ engages once all tasks have finished successfully.*

**Popular ETL tools**

There are many ETL tools available today. Modern enterprise grade ETL tools will typically include the following features:

* Automation: Fully automated pipelines
* Ease of use: ETL rule recommendations
* Drag-and-drop interface: “o-code” rules and data flows
* Transformation support: Assistance with complex calculations
* Security and Compliance: Data encryption and HIPAA, GDPR compliance

Some well-known ETL tools are listed below, along with some of their key features. Both commercial and open-source tools are included in the list.

**Talend Open Studio**

* Supports big data, data warehousing, and profiling
* Includes collaboration, monitoring, and scheduling
* Drag-and-drop GUI for ETL pipeline creation
* Automatically generates Java code
* Integrates with many data warehouses
* Open-source

**AWS Glue**

* ETL service that simplifies data prep for analytics
* Suggests schemas for storing your data
* Create ETL jobs from the AWS Console

**IBM InfoSphere DataStage**

* A data integration tool for designing, developing, and running ETL and ELT jobs
* The data integration component of IBM InfoSphere Information Server
* Drag-and-drop graphical interface
* Uses parallel processing and enterprise connectivity in a highly scalable platform

**Alteryx**

* Self-service data analytics platform
* Drag-and-drop accessibility to ETL tools
* No SQL or coding required to create pipelines

**Apache Airflow and Python**

* Versatile “configuration” as code data pipeline platform
* Open-sourced by Airbnb
* Programmatically author, schedule, and monitor workflows
* Scales to Big Data
* Integrates with cloud platforms

**The Pandas Python library**

* Versatile and popular open-source programming tool
* Based on data frames – table-like structures
* Great for ETL, data exploration, and prototyping
* Doesn’t readily scale to Big Data