What is Airflow?

A very powerful tool

Definition

Apache Airflow is a way to programmatically author, schedule, and monitor data pipelines

Core Components

- Web Server
  - Airflow's UI to see the status of your jobs and a lot more informations (we will see later).
- Scheduler
  - Responsible for scheduling your jobs.
- Executor
  - Tightly bound to the Scheduler, determines the worker processes that execute each scheduled task. It runs the task.
- Worker
  - Processes that execute the tasks, determined by the executor.
- Metadatabase
  - A database where all the metadata related to your jobs are stored.

Key Concepts

- DAG
  - A graph object representing your data pipeline
- Operator
  - Describe a single task in your data pipeline
- Task
  - An instance of an operator
- TaskInstance
  - Represents a specific run of a task = DAG + TASK + POINT IN TIME
- Workflow
  - Combination of all above

Perks of Airflow

- Pipelines are configured via Python code making them dynamic.
- You have a graphical representation of your DAGs as well as metrics.
- Airflow is scalable with the right configuration that we will see later.
- Backfill: Ability to run a DAG from the past to “backfill” until a point in time
- And much more ...

What Airflow is NOT

- Airflow is not a data streaming solution
  - Airflow is not in the scope of Apache Spark or Storm.
  - Primarily built to perform scheduled batch jobs