What is Airflow?

A very powerful orchestrator

Definition

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

Core Components

Web server	Scheduler	Metadata Database	Executor	Worker

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

What Airflow brings you

- Pipelines developed in Python.
- Graphical representation of your DAGs..
- Airflow is scalable.
- Catchup and Backfill
- Very extensible.
- And much more ...

What Airflow is NOT

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