에어비앤비에서 개발한 **워크플로우 스케줄링, 모니터링 플랫폼**

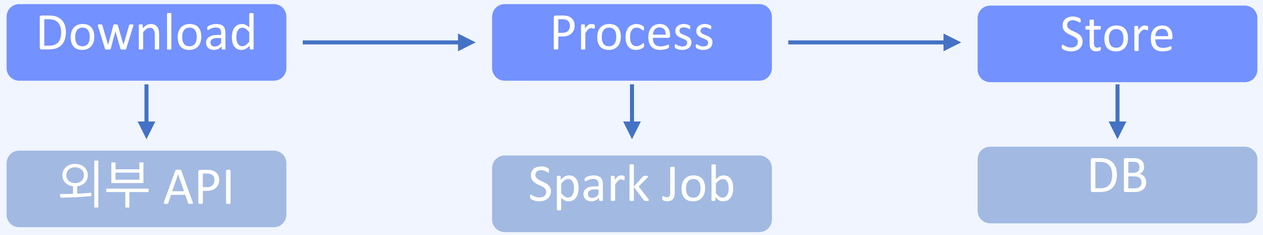
에어비앤비에서 개발한 **워크플로우 스케줄링, 모니터링 플랫폼**

실제 데이터의 **처리**가 이루어지는 곳은 아님

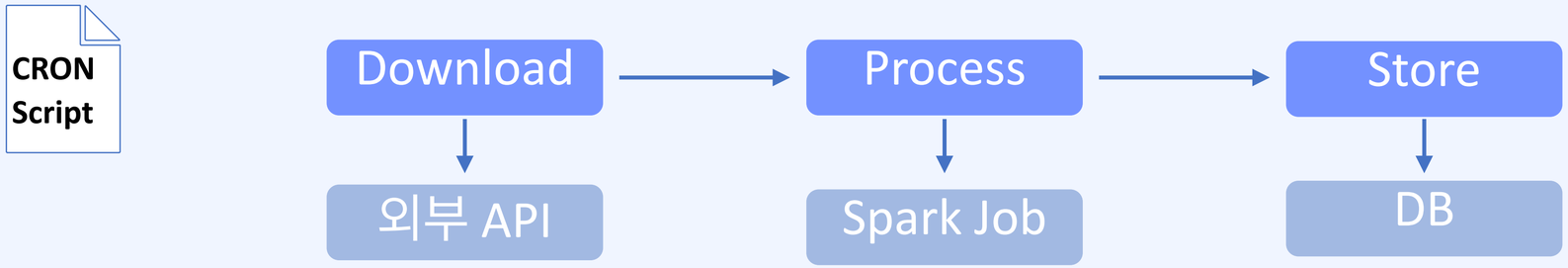
Apache Airflow

- Airbnb 개발
- 2016년 아파치 재단 incubator program
- 현재 아파치 탑레벨 프로젝트
- Airbnb, Yahoo, Paypal, Intel, Stripe

매일 10시에 주기적으로 돌아가는 데이터 파이프라인을 만드려면?



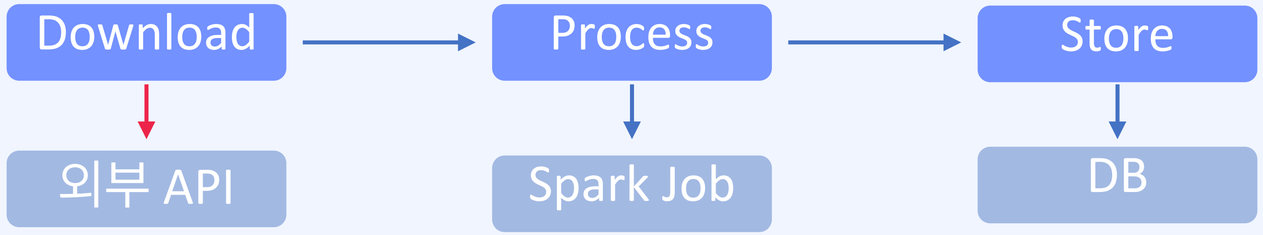
매일 10시에 주기적으로 돌아가는 데이터 파이프라인을 만드려면?



기존 방식의 문제점:

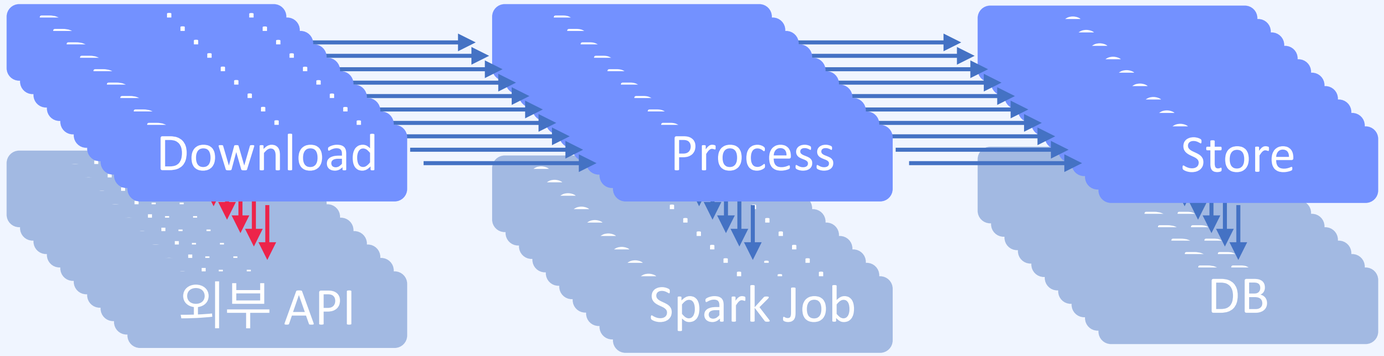
- **실패 복구** - 언제 어떻게 다시 실행할 것인가? Backfill?
- **모니터링** - 잘 돌아가고 있는지 어떻게 확인하기 힘들다
- **의존성 관리** - 데이터 파이프라인간 의존성이 있는 경우 상위 데이터 파이프라인이 잘 돌아가고 있는지 파악이 힘들다
- **확장성** - 중앙화 해서 관리하는 툴이 없기 때문에 분산된 환경에서 파이프라인들을 관리하기 힘들다
- **배포** - 새로운 워크플로우를 배포하기 힘들다

매일 10시에 주기적으로 돌아가는 데이터 파이프라인을 만드려면?



매일 10시에 주기적으로 돌아가는 데이터 파이프라인을 만드려면?

이런 파이프라인이 수십개가 있다면?



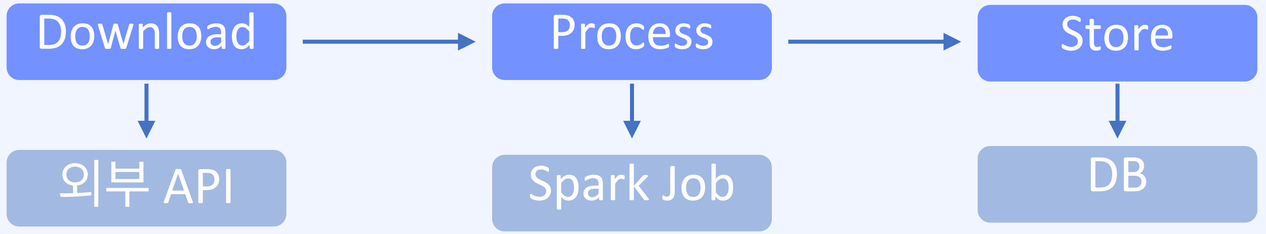
Airflow란?

Airflow 는 **워크플로우**를 **작성**하고 **스케줄링**하고 **모니터링**
하는 작업을 **프로그래밍** 할 수 있게 해주는 플랫폼

- 파이썬으로 쉬운 프로그래밍이 가능
- 분산된 환경에서 확장성이 있음
- 웹 대시보드 (UI)
- 커스터마이징이 가능

워크플로우?

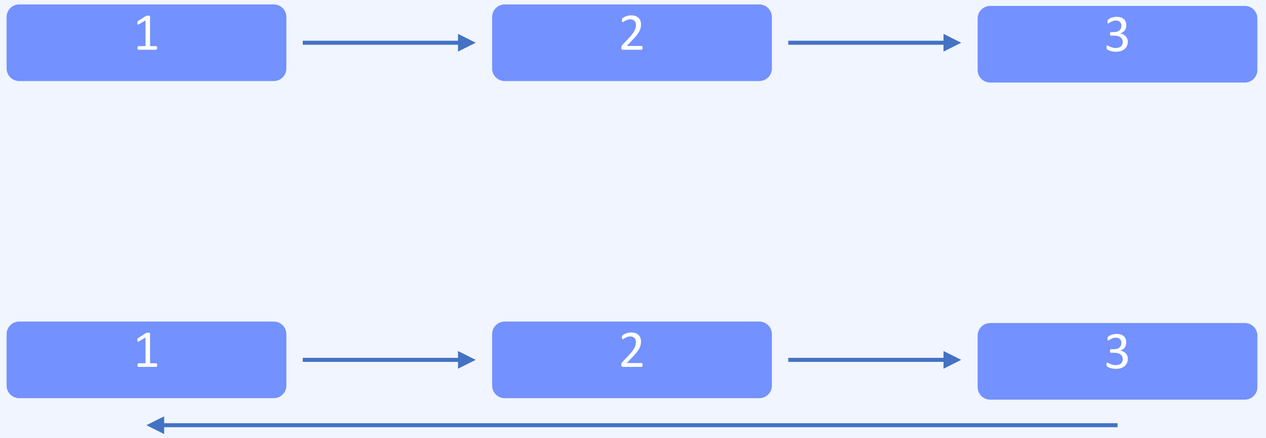
- 의존성으로 연결된 작업 (task)들의 집합
  - DAG



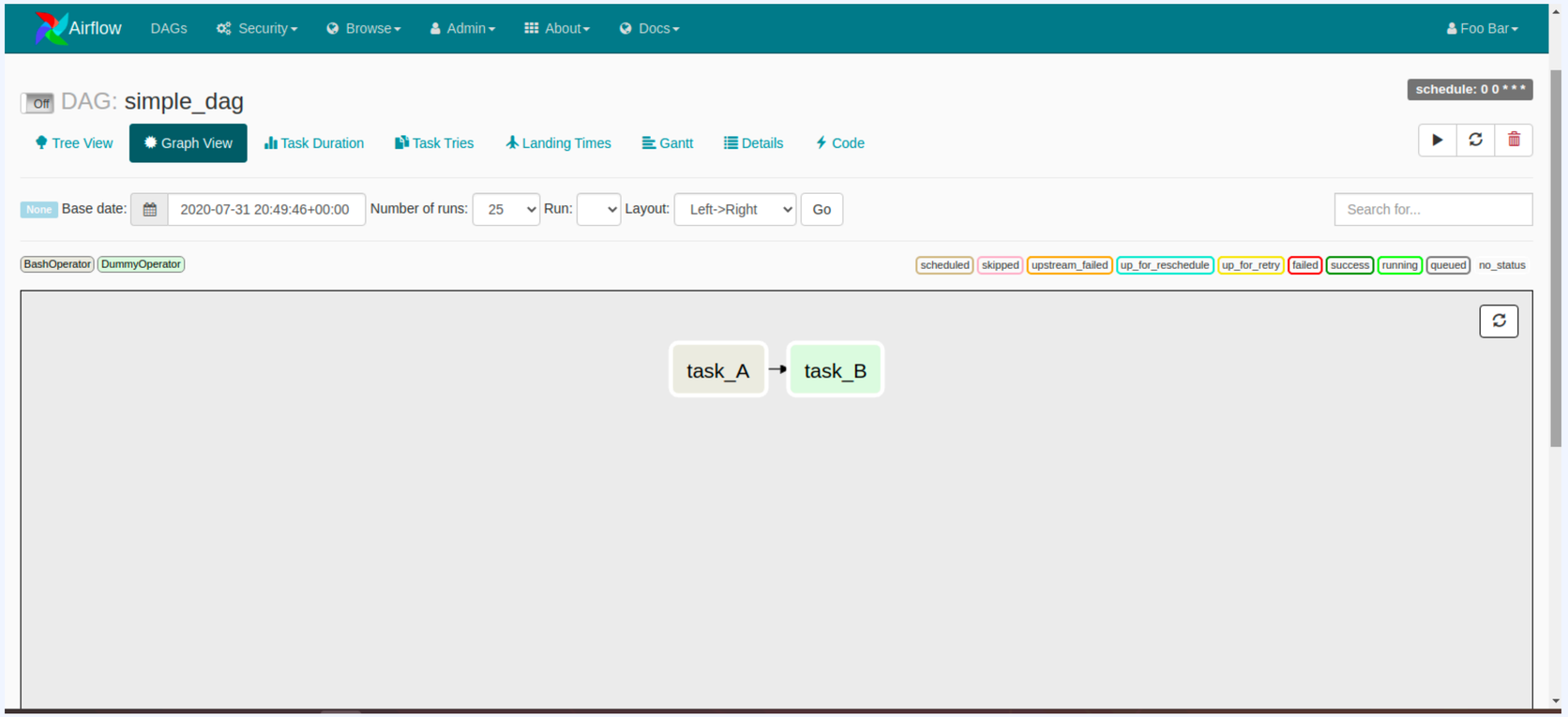
Airflow는 무엇으로 이루어져 있을까

- **웹 서버** - 웹 대시보드 UI
- **스케줄러** - 워크플로우가 **언제** 실행되는지 관리
- **Metastore** - 메타데이터 관리
- **Executor** - 테스크가 **어떻게** 실행되는지 정의
- **Worker** - 테스크를 **실행**하는 프로세스

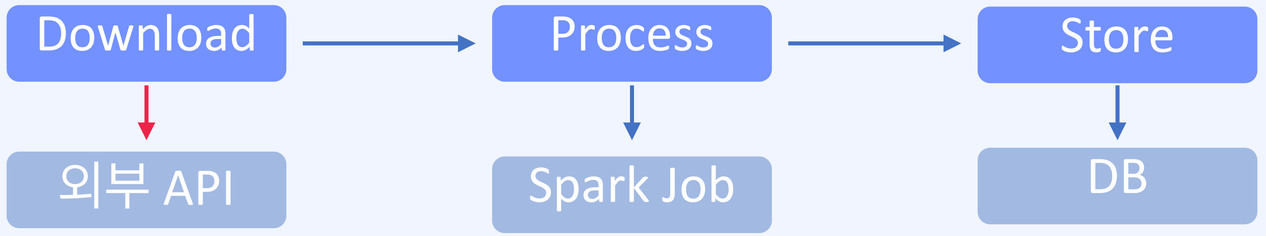
DAG



DAG



DAG



Operator: 작업(Task)를 정의하는데 사용

Action Operators

실제 연산을 수행

Transfer Operators

데이터를 옮김

Sensor Operators

테스크를 언제 실행시킬
트리거를 기다림

Operator 를 실행시키면 Task가 된다

Task = Operator Instance

여러 데이터 엔지니어링 환경에서 유용하게 쓰일 수 있다

- 데이터 웨어하우스
- 머신러닝
- 분석
- 실험
- 데이터 인프라 관리

Airflow는 어떻게 동작할까?