



바람개비 스터디 (airflow 초급)

리더	김성일
구분	스터디
시즌	시즌 1
확정 멤버	박시윤 김유민 윤주 송 송이유정

1 주제



바람개비가 아주 알록달록합니다.

에어플로우(Airflow)는 Python 기반의 오픈소스 스케줄러입니다.

어떠한 작업을 특정한 시간 및 특정한 간격으로 수행해야 할 때, 스케줄러를 씁니다.

에어플로우는 여러 스케줄러 중 아주 탁월한 프로그램이라고 할 수 있습니다.

에어플로우는 데이터 엔지니어링을 꿈꾸는 사람이라면 필수 기술이라고 할 수 있습니다.

데이터 엔지니어(데엔) JD를 보면 하둡, 에어플로우, 스팍, 카프카 등을 요구합니다.

그렇습니다... 데엔이 하고 싶으면 에어플로우를 다룰 줄 아는 것이 좋습니다.

물론 에어플로우가 다루기 쉽지는 않습니다.

저도 에어플로우 내공이 많이 부족해서 이 스터디를 계획했구요.

함께 뭉쳐 배워가는 것이 많은 스터디가 되었으면 좋겠습니다! 😊

우리 모두 에어플로우를 배우고

바람개비를 힘차게 돌려 클클을 친환경 풍력발전소로 만들어 봅시다.

2 대상

| 스터디 대상



airflow를 써보지 않으셨지만, 관심이 있으신 분

(airflow를 잘 다루시는

중급자 이상에게는 시간 낭비가 될 수 있으며,
커리큘럼은 입문자/초급자에게 맞춰서 설계할 예정입니다.)



데이터 엔지니어를 커리어로 염두하고 계시는 분



데이터 ETL 파이프라인 설계 및 데이터 엔지니어링에 관심이 있으신 분



(중요) 스터디원 서로의 의견과 사고 방식을 **존중**하고, 스터디원 모두가 저마다의 **감정**을 가지고 있는 사람임을 인지하며,
어려운 내용을 함께 고민하고
다같이 성장하실 분



(중요) 스터디 끝까지 함께 완주하실 의지를 가지고 계신 분

선행 기술(필수) - 꼭 알아야 됨

- Python 중급

선행 기술(권장) - 몰라도 됨

- SQL 쿼리 기초 지식 (SELECT가 뭔지, DML이 뭔지 아시는 분)
- 데이터베이스 기초 지식
- Docker 기초 (최소 Docker Container를 띄워보신 분)

airflow를 Docker에 띄워서 돌릴 예정이므로, Docker에 대한 기본 이해가 요구됩니다.

3 커리큘럼



커리큘럼은 추후 변동될 수 있으며,
첫 주차 오리엔테이션 시간 때 다함께 방향성을 논의할 예정입니다.

주 교재 : <Apache Airflow 기반의 데이터 파이프라인>

Apache Airflow 기반의 데이터 파이프라인 - 예스24

Airflow 설치부터 파이프라인 작성, 테스트, 분석, 백필 그리고 배포 및 관리까지를 한 권으로 해결! 이 책은 효과적인 데이터 파이프라인을 만들고 유지하는 방법을 설명하고 있으며, 이를 통해 여러분은 다양한

<https://www.yes24.com/Product/Goods/107878326>

Apache Airflow 기반의 데이터 파이프라인

에어플로우 중심의 워크플로우 구축에서
커스텀 컴포넌트 개발 및 배포, 관리까지

비즈 이앤솔루션, 윌리엄 다라워터즈 지음
김정민, 문선홍 옮김




에어플로우 책은 사실상 이게 가장 유명합니다. 이 책의 진도를 거의 끝까지 나갈 계획입니다.

참고하면 좋은 자료 (스터디 때 따로 사용하지는 않을거예요.)

1. Airflow 마스터 클래스

Airflow 마스터 클래스 강의 - 인프런

데이터 파이프라인을 효율적으로 만들고 관리하기 위한
Orchestration 도구인 Airflow에 대해 배우는 강의입니다. 초보자도
차근차근 배울 수 있는 Airflow 마스터 클래스, 환영합니다!, 데이터

 <https://www.inflearn.com/course/airflow-마스터-클래스>



좋은 강의이나, 가격이 조금 비쌉니다. 따로 깊게 배우고 싶으신 분은 추천드립니다.

2. <빅데이터를 지탱하는 기술>, <데이터 파이프라인 핵심 가이드>, <견고한 데이터 엔지니어링> 등의 데이터 엔지니어링 관련 저서

해당 저서들에 airflow 관련 내용이 언급되나, 내용이 깊지는 않습니다.

아래 주차 별 커리큘럼에서 언급되는 **CHAPTER**는 주 교재의 챕터를 의미합니다.

1주(3/5) 발표자 : 김성일

- 자기소개 및 아이스브레이킹
- 스터디 방향성 세부 논의

과제1 : 다음주(3/12) 스터디까지 Docker 설치 및 동작 확인하기

과제2 : **Apache Airflow 기반의 데이터 파이프라인** 도서 다음주(3/12)까지 준비하기

나는 2회 발표하고 싶다! 공부 너무 좋아요!! - 박시윤

2주(3/12) 발표자 : 시윤

- CHAPTER 1 Apache Airflow 살펴보기
- CHAPTER 2 Airflow DAG의 구조

3주(3/19) 발표자 : 시윤

- CHAPTER 3 Airflow의 스케줄링
- CHAPTER 4 Airflow 콘텍스트를 사용하여 태스크 템플릿 작업하기

4주(3/26) 발표자 : 윤주

- CHAPTER 5 태스크 간 의존성 정의하기
- CHAPTER 6 워크플로 트리거

5주(4/2) 발표자 : 유민

- CHAPTER 7 외부 시스템과 통신하기
- CHAPTER 8 커스텀 컴포넌트 빌드

6주(4/9) 발표자 : 성일

- CHAPTER 9 테스트하기
- CHAPTER 10 컨테이너에서 태스크 실행하기

7주(4/16) 발표자 : 유정

- CHAPTER 11 모범 사례
- CHAPTER 12 운영환경에서 Airflow 관리

8주(4/23) 발표자 :

- ~~CHAPTER 15 클라우드에서의 Airflow~~
- ~~CHAPTER 16 AWS에서의 Airflow~~

(혹은 다같이 AWS RDS나 Redshift에 데이터 파이프라인을 설계하는 **Toy Project**)

4 방식



매주 화요일 **22시-23시** 온라인 진행

1. 진행 & 발표

- [모두] 원활한 의견 교류를 위해 캠을 활성화한 상태에서 참여하시기를 **권장**합니다.
- [모두] 매주 책을 읽고, 챕터 내용을 공부 및 직접 실습 해봅니다.
- [발표자] 발표 역할을 맡은 챕터에 대한 발표를 준비합니다. 발표 내용은 **해당 챕터의 핵심 내용 요약, 어려웠던 점, 같이 보면 좋을 자료 소개, 코드 보완** 등이 있습니다.
 - 각 주차별 발표자는 1주차때 선정합니다.

- 각 챗터별 발표 시간은 15~25분을 생각하고 있으며 한 주에 2챗터 진도를 나갈 것이므로 발표는 약 30~50분 정도 진행합니다. 나머지는 의견 교류 시간입니다.
 - [스터디원] 발표 내용에 대한 의견 교류 및 질문 사항을 자유롭게 남깁니다.
2. 학습활동 **(아래 활동들은 진행하지 않을 수 있으며, 첫 주차때 다같이 논의합니다.)**
- 각 주차별 학습 내용과 관련하여 발표자분이 스터디원 분들에게 제공할 간단한 미션(과제)를 준비**(권장)**
 - 각 블로그에 공부한 내용을 업로드하고 인증합니다. **(선택)**



내용을 보시면, **매주 책을 읽고 챗터 내용을 공부한다는 방법**은 발표 주차를 제외하고는 **스터디원 자율**에 맡기는 방법입니다.
따라서 책의 내용을 꼭! 공부해주세요.

5 기록

스터디 내용을 기록 할 공간을 노션에서 따로 마련하거나, 외부 저장소 링크를 남겨주세요.

→ 각자 기록하시되(공부한 내용을), slack에 링크를 남겨주세요!

6 출석부

3회 이상 불참시 5기를 수료할 수 없습니다.

1주(3/5) - all

2주(3/12)

3주(3/19)

4주(3/26)

5주(4/2)

6주(4/9)

7주(4/16)

8주(4/23)