

데이터마이닝 팀과제 자료 조사

2025-03-22

Table of contents

1	Airflow 소개 (1분 30초)	2
1.1	Airflow란?	2
1.2	핵심 개념	2
1.3	데이터 분석에서의 중요성	2
2	현업 데이터 분석 문제점과 Airflow 해결책 (2분)	3
2.1	데이터 분석 현업의 문제점	3
2.2	Airflow의 해결책	3
3	주요 활용 사례 (3분 30초)	4
3.1	ETL 프로세스 자동화	4
3.2	매출 데이터 ETL DAG 예시	4
3.3	데이터 품질 관리	5
3.4	ML 모델 파이프라인	5
4	채용 시장에서의 Airflow 수요 (2분 30초)	7
4.1	채용 공고 분석	7
4.2	직무별 Airflow 수요	7
4.3	주요 기업 JD 발췌 예시	9
5	실습 데모와 기술적 차별점 (1분 30초)	12
5.1	간단한 DAG 구조 데모	12
5.2	경쟁 도구와의 차이점	12
5.3	Airflow의 기술적 장단점	12
6	결론 및 미래 전망 (30초)	13
6.1	요약	13
6.2	미래 전망	13
6.3	마무리	13

1. Airflow 소개 (1분 30초)

1.1 Airflow란?

- **Apache Airflow**: 워크플로우 작성, 스케줄링 및 모니터링을 위한 오픈소스 플랫폼
- 2014년 Airbnb에서 개발, 2016년 Apache 재단으로 이관
- Python으로 작성된 데이터 파이프라인 오케스트레이션 도구

1.2 핵심 개념

- **DAG(Directed Acyclic Graph)**: 작업 흐름을 표현하는 방향성 비순환 그래프
- **Task**: 개별 작업 단위 (데이터 추출, 변환, 적재 등)
- **Operator**: 작업 실행 방법을 정의 (PythonOperator, BashOperator 등)
- **Scheduler**: 작업 실행 시점 관리
- **Web Server**: 대시보드를 통한 모니터링 인터페이스

1.3 데이터 분석에서의 중요성

- 복잡한 데이터 워크플로우의 자동화 및 오케스트레이션
- 작업 간 의존성 관리 및 모니터링
- 실패한 작업의 자동 재시도 및 알림

2. 현업 데이터 분석 문제점과 Airflow 해결책 (2분)

2.1 데이터 분석 현업의 문제점

- 수동 프로세스의 한계: 반복 작업의 비효율성과 인적 오류
- 복잡한 의존성: 여러 시스템과 데이터 소스 간의 조율 어려움
- 스케줄링 이슈: 정기적인 데이터 처리와 의존성 관리
- 에러 처리: 실패 시 복구 및 알림 메커니즘 부재
- 모니터링 부족: 작업 진행 상황과 성능 추적 어려움

2.2 Airflow의 해결책

- 코드형 워크플로우: Python으로 DAG 정의 (GitOps 가능)
- 시각적 모니터링: 웹 UI를 통한 직관적인 작업 흐름 파악
- 스마트 스케줄링: Cron 기반 스케줄링 + 의존성 기반 실행
- 강력한 확장성: 다양한 시스템과의 통합 (AWS, GCP, Azure, Databricks 등)
- 견고한 에러 처리: 자동 재시도, 알림, 대체 경로 설정

3. 주요 활용 사례 (3분 30초)

3.1 ETL 프로세스 자동화

- 사례: 일일 매출 데이터 통합 파이프라인

3.2 매출 데이터 ETL DAG 예시

```
from airflow import DAG
from airflow.operators.python import PythonOperator
from datetime import datetime, timedelta
import warnings

warnings.filterwarnings("ignore")

default_args = {
    'owner': 'data_team',
    'depends_on_past': False,
    'retries': 3,
    'retry_delay': timedelta(minutes=5)
}

def extract_sales_data(**kwargs):
    return {"sales_data": "extracted"}

def transform_data(**kwargs):
    return {"transformed_data": "ready"}

def load_to_warehouse(**kwargs):
    print("Data loaded successfully")

with DAG(
    'daily_sales_etl',
    default_args=default_args,
    schedule_interval='0 2 * * *',
    start_date=datetime(2023, 1, 1)
) as dag:
```

```

extract_task = PythonOperator(
    task_id='extract_sales_data',
    python_callable=extract_sales_data
)

transform_task = PythonOperator(
    task_id='transform_sales_data',
    python_callable=transform_data
)

load_task = PythonOperator(
    task_id='load_to_warehouse',
    python_callable=load_to_warehouse
)

extract_task >> transform_task >> load_task

```

- **효과:**

- 여러 데이터 소스(POS, 온라인 스토어, 외부 판매 채널)를 자동 통합
- 일관된 데이터 처리 및 변환 보장
- 작업 의존성 자동 관리

3.3 데이터 품질 관리

- **사례:** 데이터 검증 및 알림 시스템

- 데이터 완전성, 정확성, 일관성 검증
- 임계값 초과 시 자동 알림

- **구현 방식:**

- Great Expectations과 같은 도구와 통합
- 검증 실패 시 Slack, Email로 알림
- 데이터 품질 측정 및 대시보드 제공

3.4 ML 모델 파이프라인

- **사례:** 추천 모델 정기 학습 및 배포

- 새로운 사용자 행동 데이터 수집
- 정기적인 모델 재학습
- 성능 검증 후 자동 배포
- 모델 버전 관리 및 롤백

- **효과:**

- 모델 신선도 유지
- 일관된 학습 및 평가 파이프라인
- 버전 관리 및 추적 용이성

4. 채용 시장에서의 Airflow 수요 (2분 30초)

4.1 채용 공고 분석

- 데이터 관련 직무에서 Airflow 언급 비율: **약 68%**
- 연도별 Airflow 언급 추이: 2020년부터 꾸준한 증가

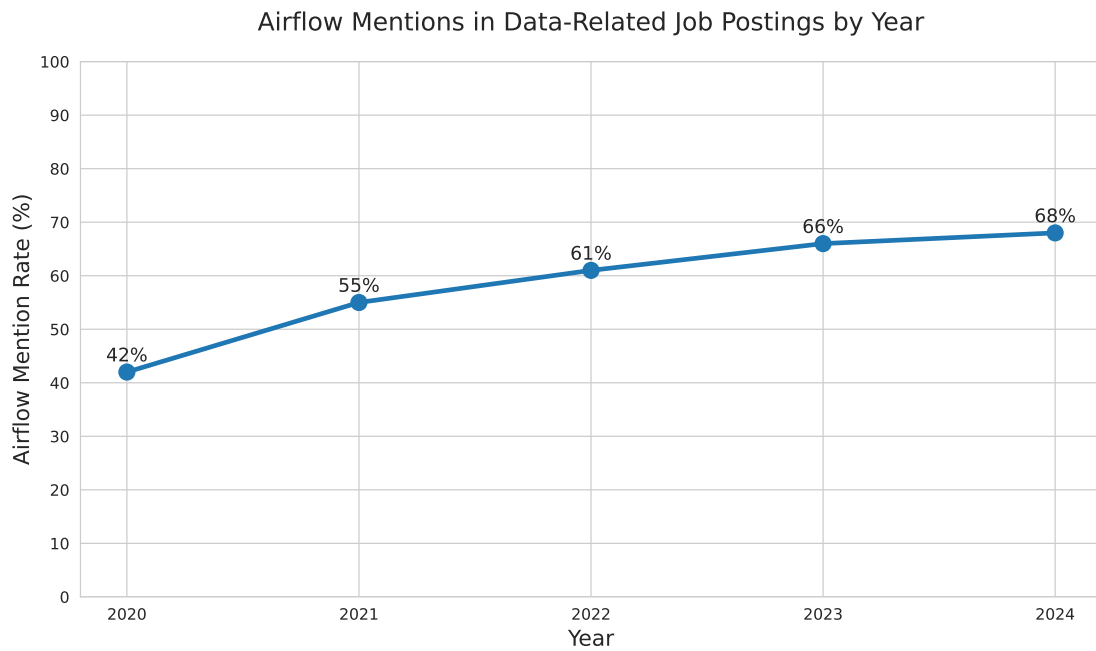


Figure 4.1: 연도별 Airflow 언급 추이 (2020-2024)

4.2 직무별 Airflow 수요

- 데이터 엔지니어: 85% 이상의 JD에서 요구
- 데이터 사이언티스트: 약 50%에서 우대사항으로 명시
- ML 엔지니어: 65%에서 필수 또는 우대 스킬로 언급
- BI 애널리스트: 30%에서 플러스 스킬로 평가

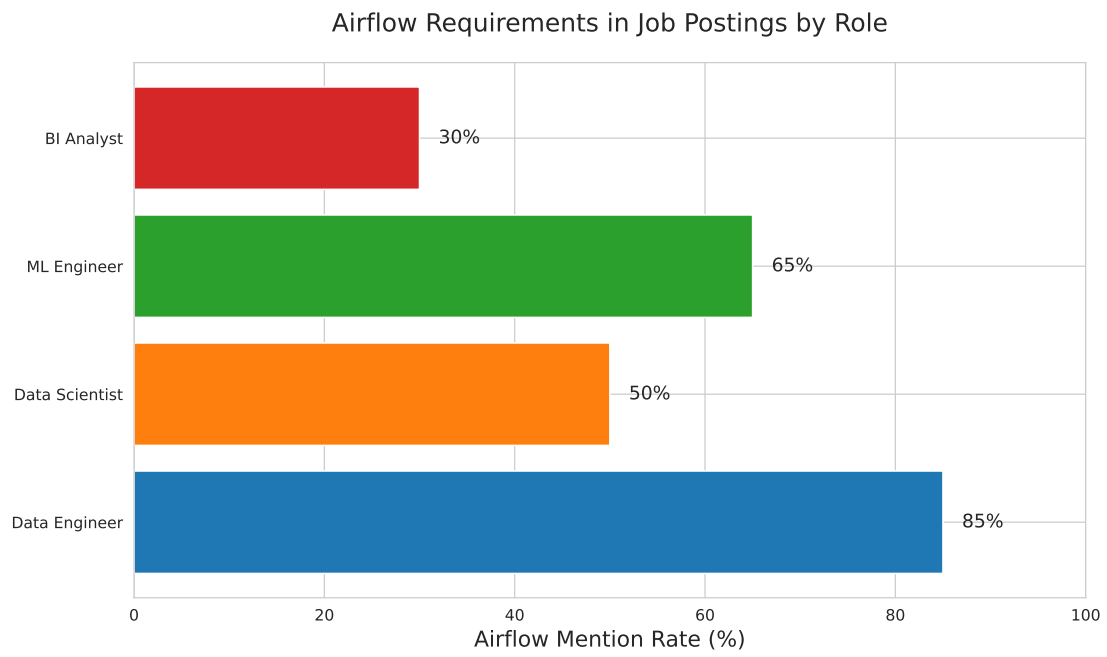


Figure 4.2: 직무별 Airflow 요구 비율

4.3 주요 기업 JD 발췌 예시

Data Engineer

[DataOps](#)[Infra](#)

토스증권 소속 | 정규직

지원하기

합류하게 될 팀에 대해 알려드려요

- 토스증권 Data Engineer(Infra)는 Data Division내에 Data Infra Team에 속해 있어요.
- Data Infra Team은 크게 Hadoop Eco 기반의 빅데이터 인프라와 로그/검색 플랫폼(Elasticsearch)을 운영하고 Data 기술조직에서 데이터 입수, 각종 데이터 작업 등 다양하게 사용하는 Kubernetes에 대한 운영과 기술지원을 하고 있어요.

Data Division을 소개합니다





- 토스증권 Data Division은 세계 최고로 데이터를 잘 다루는 증권사가 되기 위해 데이터 기술, 서비스 그리고 데이터 기반의 의사결정에 기여하고 있어요.
- 다양한 데이터 직군이 모여 밀접하게 협업하며 즐겁게 일하고 있어요.
- 또한 주기적으로 Tech Weekly를 진행하며 서로의 노하우를 공유하고 있어요. 본인의 흥미와 의지가 있다면 얼마든지 다른 직군의 업무와 노하우를 공유받을 수 있어요.



합류하면 함께 할 업무예요

- 토스증권의 Data Engineer(Platform)는 증권 서비스의 다양한 데이터를 효과적으로 저장하고 처리하기 위한 플랫폼을 운영하고 발전시키는 업무를 담당해요.
- Hadoop Ecosystem 기반의 데이터 플랫폼을 구성하여 운영중이며 사용하는 주요 컴포넌트로는 Hadoop, Spark, Impala, Hive, Kudu, Kafka, [Airflow](#), Jenkins, k8s가 있어요.
- 토스증권 서비스에서 발생하는 국내/해외 종목, 주식매매 등 다양한 데이터를 가장 효율적이고 안전한 방법으로 처리하기 위한 플랫폼 아키텍처를 설계하고 구축해요.
- 증권사에 맞는 데이터 보안과 거버넌스를 지속적으로 강화해요.
- 도전적인 문제들을 해결하기 위해 새로운 환경을 고민하고 새로운 기술을 적극적으로 도입해요.

Figure 4.3: 토스 증권 Data Engineer JD

MLOps Engineer Internship



 Platform & Infra  Magok, Seoul

팀 소개

Platform&Infra팀은 AI 모델의 개발부터 서비스 운영을 위한 배포에 이르기까지 AI 모델의 수명 주기를 최적화하고, 효율적으로 관리하기 위한 MLOps 파이프라인을 구축합니다. 또한 AI 서비스의 안정적인 운영 지원을 위한 보안성 강화, 인프라 관리 및 자원 최적화 업무를 수행합니다.

수행 업무

- AI 모델의 학습/추론 플랫폼을 설계하고 구축하며 운영합니다.
- AI/ML 플랫폼 운영 및 생산성 향상을 위한 다양한 서비스와 도구를 개발합니다.

지원자격

- Linux 및 CLI 환경을 다뤄본 경험이 있으신 분
- Python, Go 등 프로그래밍 언어를 활용한 웹 애플리케이션 개발을 해봤으면 좋아요.
- Docker 및 Kubernetes 같은 컨테이너 기술의 기본 개념을 이해하고 있으면 좋아요.

우대사항

- GCP, AWS, Azure 같은 Public Cloud 환경에서 개발을 해봤으면 좋아요.
- CI/CD 도구(Helm, Kustomize, ArgoCD 등)를 활용한 개발 및 운영을 해봤으면 좋아요.
- Triton, TensorRT 같은 AI 서빙 프레임워크를 사용해봤으면 좋아요.
- **Airflow** Kubeflow 같은 Workflow 툴을 사용해봤으면 좋아요.
- 기술적인 내용을 문서화하고 팀원들과 공유해봤으면 좋아요.

Figure 4.4: LG MLOps Engineer Internship

Senior Software Engineer, Enterprise Data and Engineering

Google

Hyderabad, Telangana, India

Mid

Apply

Minimum qualifications:

- Bachelor's degree or equivalent practical experience.
- 5 years of experience with software development in one or more programming languages, and with data structures/algorithms.
- 3 years of experience with ML/AI algorithms and tools, deep learning, or natural language processing or ML sub domain, including in Applied ML space.
- 3 years of experience testing, maintaining or launching software products, and 3 years of experience with software design (either distributed system design or ML design) and architecture.
- Experience in a leadership role (technical leadership or people management, supervision, or team leadership).

Preferred qualifications:

- Master's degree or PhD in Computer Science or a related technical field.
- 3 years of experience in a technical leadership or individual contributor role.
- Experience with ML frameworks, Applied ML across sub-domains.

About the job

Google's software engineers develop the next-generation technologies that change how billions of users connect, explore, and interact with information and one another. Our products need to handle information at massive scale, and extend well beyond web search. We're looking for engineers who bring fresh ideas from all areas, including information retrieval, distributed computing, large-scale system design, networking and data storage, security, artificial intelligence, natural language processing, UI design and mobile; the list goes on and is growing every day. As a software engineer, you will work on a specific project critical to Google's needs with opportunities to switch teams and projects as you and our fast-paced business grow and evolve. We need our engineers to be versatile, display leadership qualities and be enthusiastic to take on new problems across the full-stack as we continue to push technology forward.

In this role, you will deliver solutions to meet the data, reporting, and analytics needs of Googlers. You will drive high impact projects to deliver data management and investigative solutions for our partners across Google, create and maintain logical and physical database designs, and ensure the integrity of data under the purview of the projects, including establishing security procedures to protect and maintain the highest level of confidentiality and data security. In addition, you'll partner with internal teams to define and implement solutions that improve internal business processes, maintain highest levels of development practices, integrate third-party products with internal systems, and maintain highest levels of development practices, including technical design, solution development, systems configuration, test documentation/execution, issue identification and resolution, writing clean, modular and self-sustaining code.

At Corp Eng, we build world-leading business solutions that scale a more helpful Google for everyone. As Google's IT organization, we provide end-to-end solutions for organizations across Google. With an inclusive mindset, we deliver the right tools, platforms, and experiences for all Googlers as they create more helpful products and services for everyone. In the simplest terms, we are Google for Googlers.

Responsibilities

- Deliver data integration and pipeline solutions leveraging technologies such as Kafka, Spark, [Airflow](#) or similar technologies.
- Work well across Product Areas, build relationships, and deliver on shared objectives.

Google Hyderabad, Telangana, India Mid

Apply

Minimum qualifications:

- Bachelor's degree or equivalent practical experience.
- 5 years of experience with software development in one or more programming languages, and with data structures/algorithms.
- 3 years of experience with ML/AI algorithms and tools, deep learning, or natural language processing or ML sub domain, including in Applied ML space.
- 3 years of experience testing, maintaining or launching software products, and 3 years of experience with software design (either distributed system design or ML design) and architecture.
- Experience in a leadership role (technical leadership or people management, supervision, or team leadership).

Preferred qualifications:

- Master's degree or PhD in Computer Science or a related technical field.
- 3 years of experience in a technical leadership or individual contributor role.
- Experience with ML frameworks, Applied ML across sub-domains.

About the job

Google's software engineers develop the next-generation technologies that change how billions of users connect, explore, and interact with information and one another. Our products need to handle information at massive scale, and extend well beyond web search. We're looking for engineers who bring fresh ideas from all areas, including information retrieval, distributed computing, large-scale system design, networking and data storage, security, artificial intelligence, natural language processing, UI design and mobile; the list goes on and is growing every day. As a software engineer, you will work on a specific project critical to Google's needs with opportunities to switch teams and projects as you and our fast-paced business grow and evolve. We need our engineers to be versatile, display leadership qualities and be enthusiastic to take on new problems across the full-stack as we continue to push technology forward.

In this role, you will deliver solutions to meet the data, reporting, and analytics needs of Googlers. You will drive high impact projects to deliver data management and investigative solutions for our partners across Google, create and maintain logical and physical database designs, and ensure the integrity of data under the purview of the projects, including establishing security procedures to protect and maintain the highest level of confidentiality and data security. In addition, you'll partner with internal teams to define and implement solutions that improve internal business processes, maintain highest levels of development practices, integrate third-party products with internal systems, and maintain highest levels of development practices, including technical design, solution development, systems configuration, test documentation/execution, issue identification and resolution, writing clean, modular and self-sustaining code.

At Corp Eng, we build world-leading business solutions that scale a more helpful Google for everyone. As Google's IT organization, we provide end-to-end solutions for organizations across Google. With an inclusive mindset, we deliver the right tools, platforms, and experiences for all Googlers as they create more helpful products and services for everyone. In the simplest terms, we are Google for Googlers.

Responsibilities

- Deliver data integration and pipeline solutions leveraging technologies such as Kafka, Spark, **Airflow** or similar technologies.
- Work well across Product Areas, build relationships, and deliver on shared objectives.

Figure 4.5: Google Senior Software Engineer, Enterprise Data and Engineering

5. 실습 데모와 기술적 차별점 (1분 30초)

5.1 간단한 DAG 구조 데모

- 웹 UI 통한 DAG 시각화
- 태스크 상태 확인 및 관리
- 로그 및 오류 모니터링

5.2 경쟁 도구와의 차이점

- **Airflow vs Prefect:**
 - Airflow: 성숙한 생태계, 풍부한 커넥터
 - Prefect: 더 현대적인 API, 로컬 개발 편의성
- **Airflow vs Luigi:**
 - Airflow: 풍부한 UI, 스케줄링 강점
 - Luigi: 더 가벼운 구조, 더 간단한 설정

5.3 Airflow의 기술적 장단점

- **장점:**
 - 광범위한 커뮤니티와 풍부한 사용 사례
 - 다양한 시스템과의 통합 용이성
 - 강력한 모니터링 및 알림 기능
- **단점:**
 - 상대적으로 가파른 학습 곡선
 - 가벼운 워크로드에는 오버스펙일 수 있음
 - 설정 및 관리의 복잡성

6. 결론 및 미래 전망 (30초)

6.1 요약

- Airflow는 데이터 분석 현업에서 핵심 워크플로우 관리 도구로 자리매김
- ETL, 데이터 품질 관리, ML 파이프라인 등 다양한 활용 사례
- 채용 시장에서 지속적으로 수요가 증가하는 기술

6.2 미래 전망

- 클라우드 네이티브 환경으로의 발전
- Kubernetes와의 통합 강화
- 데이터 거버넌스 기능 확장
- 보다 사용자 친화적인 인터페이스로 발전

6.3 마무리

- 데이터 기반 의사결정이 중요해짐에 따라 Airflow의 중요성도 계속 증가할 전망
- 데이터 엔지니어링과 사이언스를 연결하는 핵심 도구로 발전