**Chapter 1. AWS 기반 데이터 과학 소개**

* Amazon.com Recommendations: Items-to-Item Collaborative Filtering논문
* 클라우드는 데이터과학 프로젝트를 개발하는 것은 물리적으로 자체 서버 인프라를 구축하지 않고서도 프로토타입에서 프로덕션으로 모델을 원활하게 전화할 수 있게 해준다.
  + 1. 큐브플로우 파이프라인: 쿠버네티스 환경에 빌드된 Kubeflow pipeline이라는 오케스트레이션 하위 시스템을 포함하는 비교적 새로운 오픈 소스 머신러닝 플랫폼이다.
    2. Apache Airflow는 주로 데이터 엔지니어링과 추출extract\_변환 Transformaion\_로드 load(ETL) 파이프라인을 조정하는 인기있는 오픈소스 워크플로우 플랫폼이다.

A **pipeline orchestrator** is a tool that helps to automate these workflows. An orchestrator can schedule jobs, execute workflows, and coordinate dependencies among tasks.

1.3.3. 신뢰성: 우리는 훈련 데이터에 대한 변경 사항 트래킹 및 버전 관리를 자동화해야 한다. 이를 통해 오류가 발생해도 정확하게 이전 버전의 모델로 다시 되돌릴 수 있다.

1.5.1 아마존 S3와 AWS 레이크 포메이션을 활용한 데이터 수집과 데이터 레이크

- 데이터에 엑세스하고 분석하는 팀과 애플리케이션이 점점 더 많아지면서 많은 기업에서는 확장성이 뛰어나고 가용성이 높으며 안전하고 유연한 데이터 저장소인 데이터 레이크 data lake로 이동하고 있다.