모던 데이터 엔지니어링 아키텍쳐

2022.1

데이터 웨어하우스

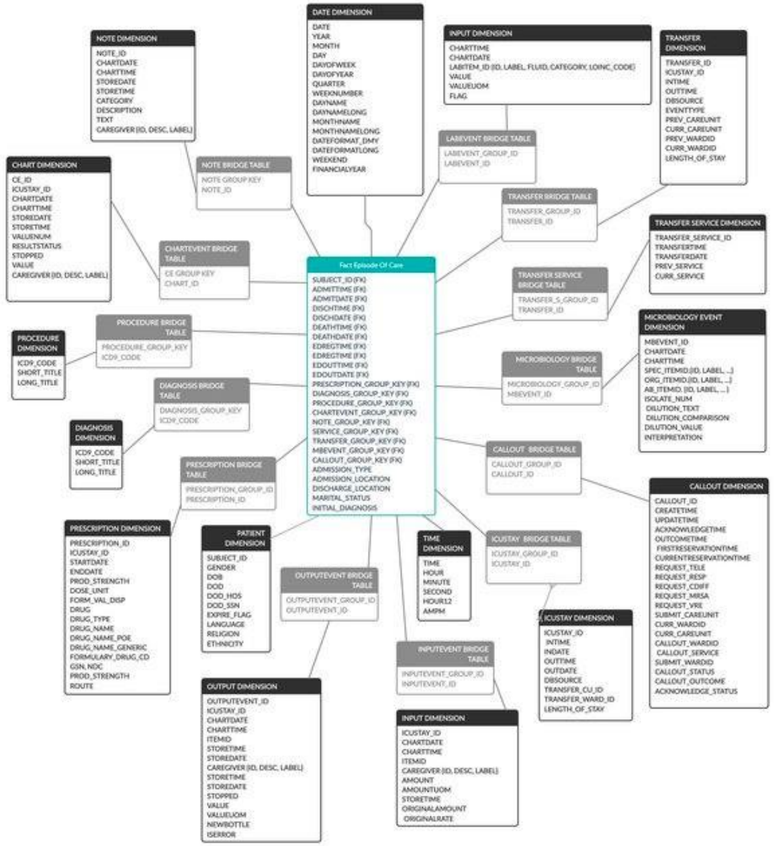
과거

1. 컴퓨팅 파워와 용량이 비싸다
2. 용도가 정해져 있다
3. 데이터가 나올 곳도 정해져있다

데이터 웨어하우스

과거의 데이터 관리 방식

1. 데이터의 형식, 즉 스키마를 미리 만들어야 합니다
2. 데이터의 변동이 별로 없습니다
3. 효율적인 데이터베이스 모델링이 중요합니다



ETL

1. 추출(Extract)
2. 스키마에 맞게 변환(Transform)
3. 데이터베이스에 적재(Load)

다양해지는 데이터 형식

- 데이터로 할 수 있는 일이 다양해지고 형태를 예측하기 불가능해지면서 스키마를 정의하기 힘들어졌습니다.
    1. 실시간성을 요구하는 기능들
    2. 빨라지는 기능 추가
    3. 실시간 로그
    4. 비정형 데이터
    5. 서드 파티 데이터

저렴해지는 컴퓨터 파워

컴퓨팅 파워도 많이 저렴해졌습니다.

최대한 많은 데이터를 미리 저장해두고
많은 양의 프로세싱을 더할 수 있게 되었습니다.

일반적인 회사에선 이제
컴퓨팅 파워에 대한 비용 최적화보다
비즈니스와 속도를 최적화하는 쪽이 이득이 큽니다.

현재 데이터를 운영하는 방식

기존의 ETL 방식에서 ELT 방식의 아키텍쳐로 변환하고 있습니다.

현재 데이터를 운영하는 방식(예)

시스템의 복잡도에 따라 데이터 추출과 적재를 한번에 하기도 합니다.

데이터 인프라 트렌드

- 클라우드 웨어하우스 - Snowflake, Google Big Query
- Hadoop에서 Databricks, Presto같은 다음 세대로
- 실시간 빅데이터 처리 (Stream Processing)
- ETL → ELT
- Dataflow 자동화 (Airflow)
- 데이터 분석 팀을 두기 보단 누구나 분석할 수 있도록
- 중앙화 되는 데이터 플랫폼 관리 (access control, data book)

모던 데이터 아키텍처 해부

데이터 아키텍쳐 분야를 크게 6가지로 나누어보았을때

모던 데이터 아키텍처 해부

데이터 아키텍쳐 분야를 크게 6가지로 나누어보았을때

소스	수집 및 변환	저장	과거	예측	출력
비즈니스와 운영 데이터 생성	운영 시스템에서 데이터 추출 (E) 추출된 데이터를 저장하고 스키마 관리 (L) 데이터를 분석할 수 있도록 변환 (T)	데이터를 쿼리와 처리 시스템이 쓸 수 있도록 저장 비용과 확장성 면으로 최적화	데이터 분석을 위한 인사이트 만들기 (Query) 저장된 데이터를 이용해 쿼리를 실행하고 필요시 분산처리 (Processing) 과거에 무슨 일이 일어났는지 혹은 미래에 무슨일이 일어날지 (ML)		데이터 분석을 내부와 외부 유저에게 제공 데이터 모델을 운영 시스템에 적용

데이터는 어떻게 흘러갈까

데이터가 생성돼서 적용되기까지

Sources

데이터 생성

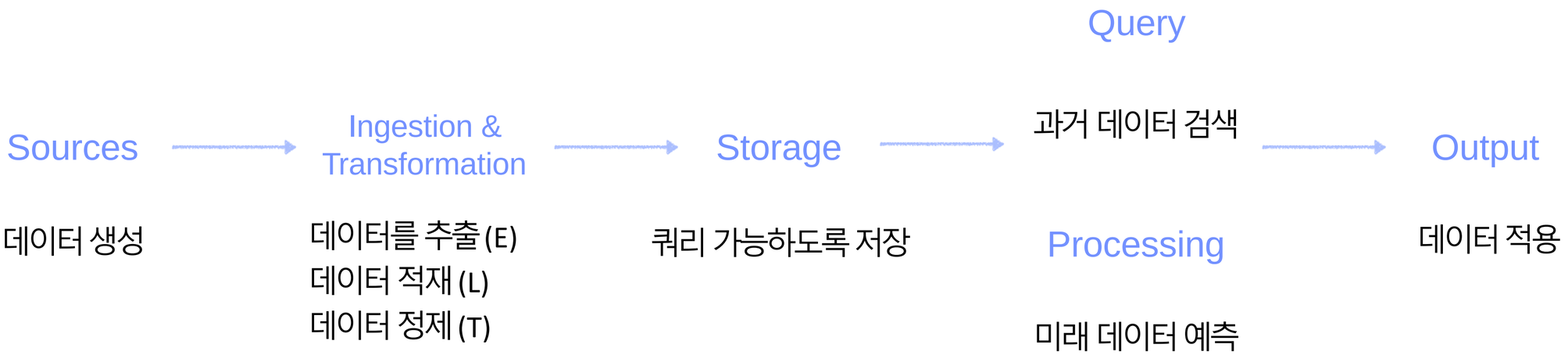
Output

데이터 적용

모던 데이터 아키텍쳐

데이터는 어떻게 흘러갈까

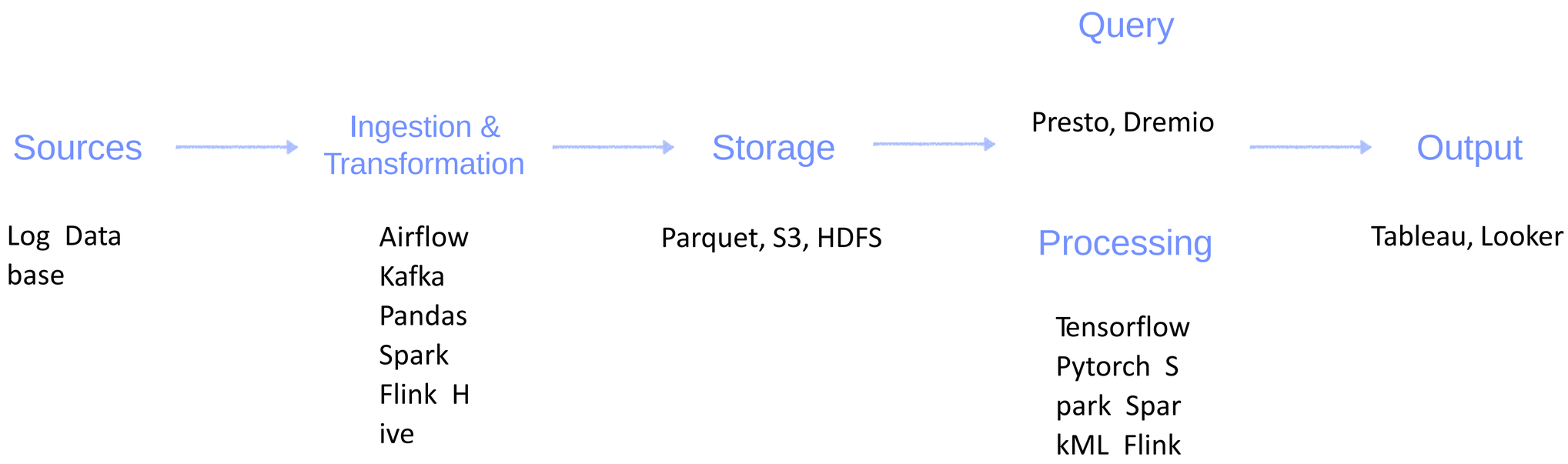
데이터가 생성돼서 적용되기까지



모던 데이터 아키텍쳐

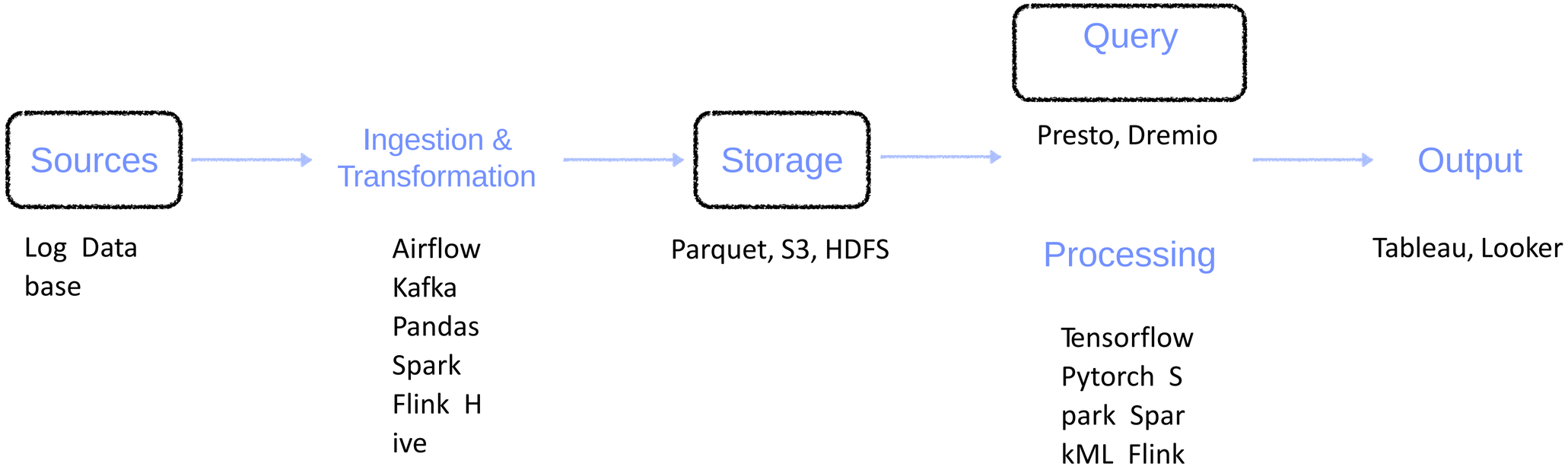
데이터는 어떻게 흘러갈까

데이터 엔지니어링 도구들



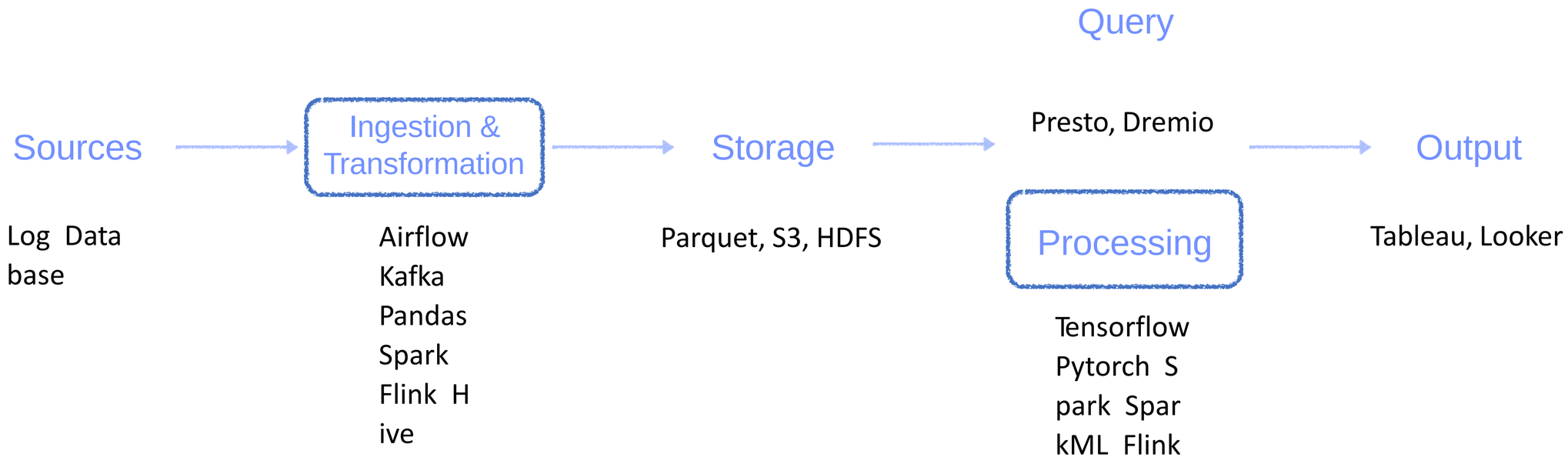
데이터는 어떻게 흘러갈까 - 도구들

서비스 레벨 보다는 로우레벨 문제들을 푸는 분야



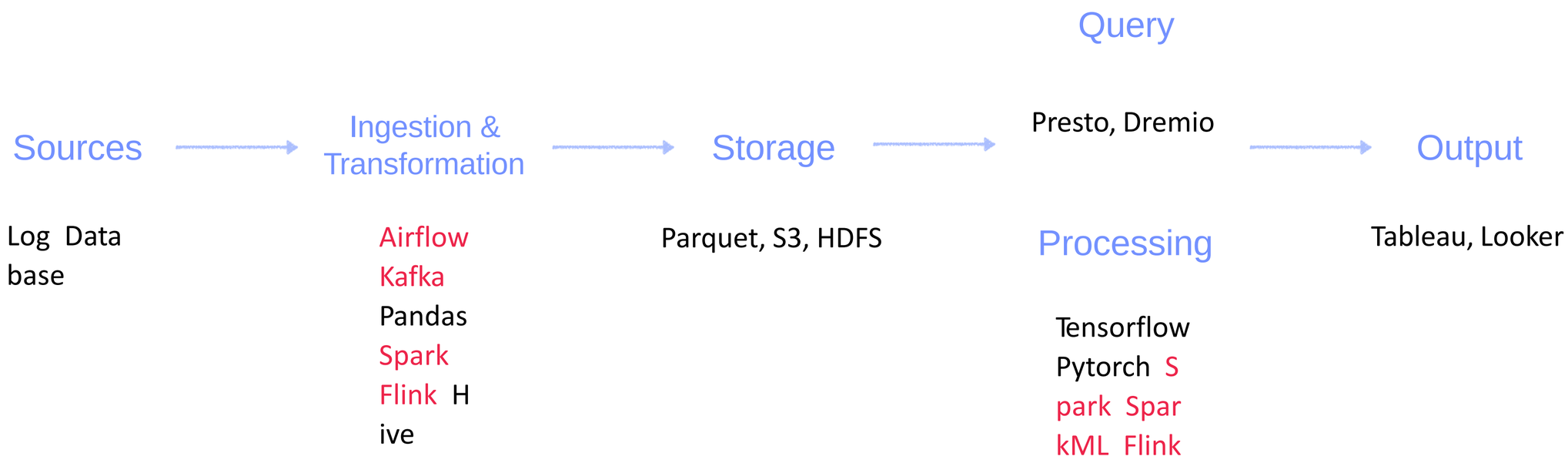
데이터는 어떻게 흘러갈까 - 도구들

일반적인 엔지니어링은 “수집 및 변환” 그리고 “데이터 처리” 에 집중



데이터는 어떻게 흘러갈까 - 도구들

앞으로 배울 내용들



앞으로 배울 내용

1. **Spark**와 데이터 병렬-분산 처리
2. **Airflow**와 데이터 오케스트레이션
3. **Kafka**와 이벤트 스트리밍
4. **Flink**와 분산 스트림 프로세싱