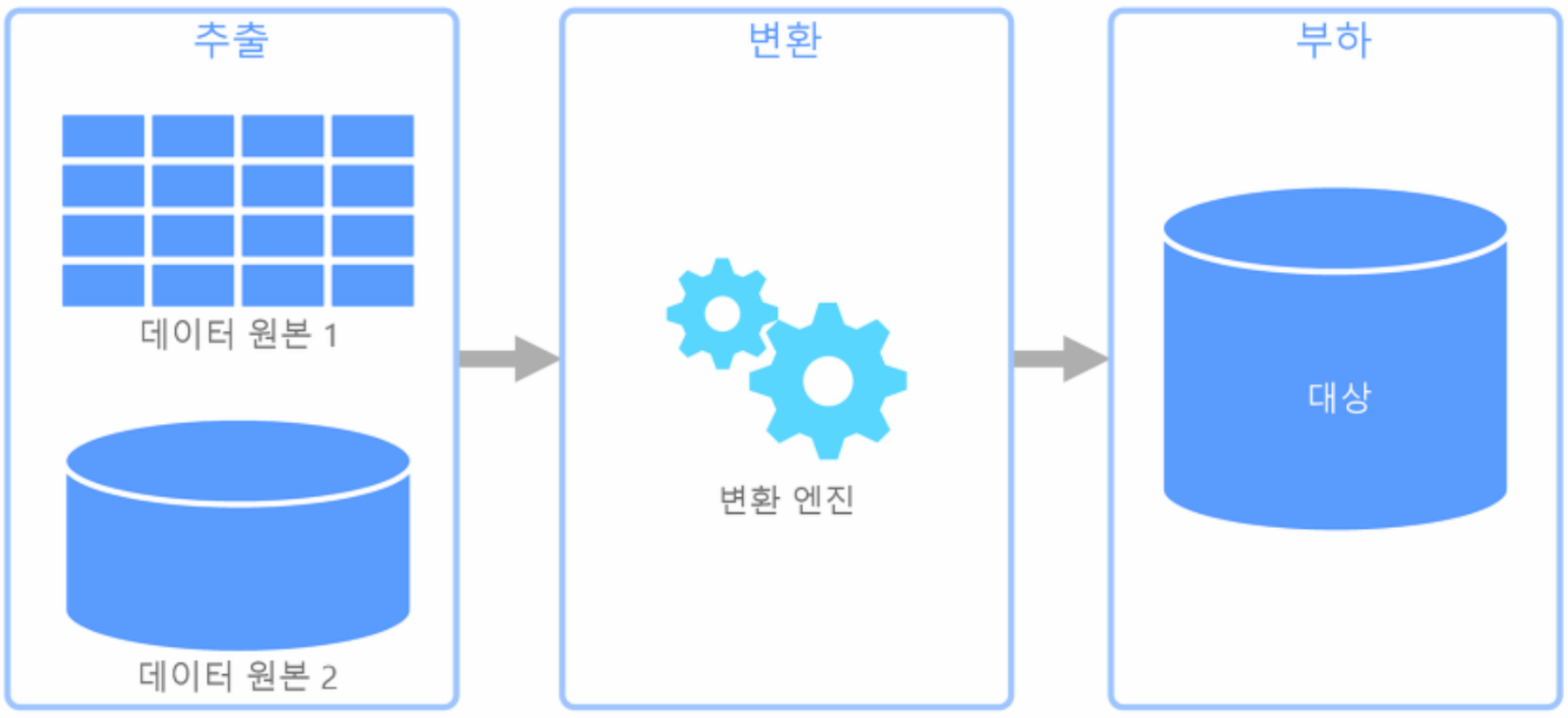
ETL

- **E**xtraction, **T**ransformation, **L**oading 의 약자 (**추출, 변환, 적재**)
다양한 데이터 원천으로부터 `데이터를 추출 및 변환` 하여 **ODS**(운영 데이터 스토어: **O**perational **D**ata **S**tore), **DW**(데이터 웨어하우스 : **D**ata **W**arehouse), **DM**(데이터 마트 : **D**ata **M**art) 등에 데이터를 적재하는 작업의 핵심 구성요소.
- 일반적으로 발생하는 데이터 변환에는 **필터링, 정렬, 집계, 데이터 조인, 데이터 정리, 중복 제거 및 데이터 유효성 검사**등의 다양한 작업이 포함된다.



참고사항

ETL은 대용량 데이터에 대한 일괄작업(Batch, 실시간의 반대 개념)을 통해 정형 데이터를 통합한다.
And 정형 데이터의 실시간 (혹은 근접 실시간 처리)와 통합에 관한 기술은 **CDC**, **EAI**

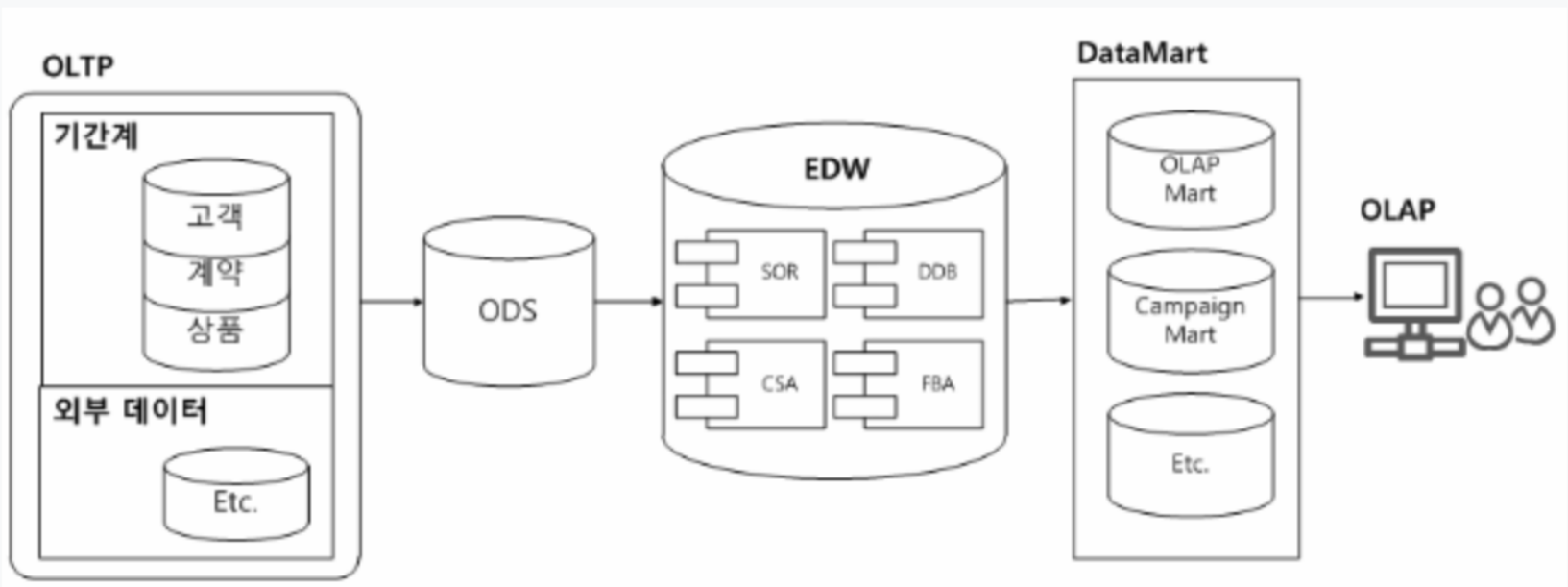
CDC

- **C**hanged **D**ata **C**apture의 약자. 데이터 캡처 기술

> 실시간으로 원천 데이터의 변경사항을 감지하여 이관하는 작업.
> 다양한 방법으로 캡처를 진행하나, 로그를 읽어 변경을 반영하는 방법과 통신을 통한 변경이 주로 쓰인다.

CDC 솔루션의 원칙

복제 대상인 소스DB의 성능에 거의 영향을 주지 않고 타깃 DB로 실시간 복제를 실행.
이때 소스DB에 부하를 주지 않도록 ① 별도의 타깃 DB를 생성하고 ② 최초 1회만 마이그레이션을 수행한 후 ③ 소스DB의 데이터 변경내역 만을 읽어서 타깃DB에서 같은 작업을 수행함으로써 데이터의 정합성을 유지.



CDC 와 ETL

CDC와 ETL 모두 원천(Source) 데이터 정보를 추출하여 목표(Target) 시스템에 적재하는 개념은 동일하지만, 방법 / 사용목적 / 적재수준 에서 차이가 발생한다.

ETL

1. 주목적 : ~~기간계~~ 시스템의 하루일과를 다 끝내고 난 뒤 저녁에 ~~기간계~~ Batch 프로그램 까지 다 실행되어 필요한 집계처리 까지 마무리한 상태에서 이들 데이터(원장 / 거래 / 집계)를 정보계 시스템에 넘겨 어떤 처리를 하고자 할 때 사용하는 방식

2. 사용 기술 : CDC와 달리 시스템 테이블에 접근하는게 아니라 ~~기간계~~의 원장/거래/집계 테이블을 대상으로 그날의 변경분을 찾아 그걸 정보계 시스템에 전달하는 방식을 주로 사용.

   만약 변경분을 인식할 수 있는 컬럼이 존재하지 않는다면, 매일매일 해당 테이블의 모든 데이터를 ALL-COPY 방식으로 적재해야 한다.

3. 적재 수준 : ETL은 원장/거래/집계 테이블을 원천데이터로 하기 때문에 적재주기(시간 / 일 / 월) 에 따라 적재 수준이 정해지게 된다.

   즉, 적재 주기가 2시간 간격으로 정해져 있다면, 적재 수준은 2시간 마다의 최종 데이터로 한정되게 된다.

CDC

1. 주목적 : 실시간(Real-Time) or 준실시간(Near Real Time) 으로 원천시스템(**기간계, 업무계**) 데이터를 읽어들여 **정보계 시스템 or 후선업무 시스템** 등에 정보를 넘겨 어떤 처리를 하고자 할 때 사용하는 방식.

2. 사용 기술 : 모든 거래 발생 이벤트를 추적해야 하기 때문에 DBMS 종류에 관계없이 변경 이력을 관리하는 `DB Archive Log` 를 주로 사용. CDC 솔루션의 데몬이 떠서 일정 시간별로 Archive Log를 읽어들여 타겟시스템에 쌓는 방식으로 대부분 작동.

3. 적재 수준 : 원천 테이블이 아닌 시스템 테이블인 Archive Log 를 읽어 처리하므로, **적재 주기(시간 / 일 / 월) 에 관계없이 모든 원천 데이터의 변경사항이 적재 대상**이 되게 된다.