ETL / CDC

학습목차

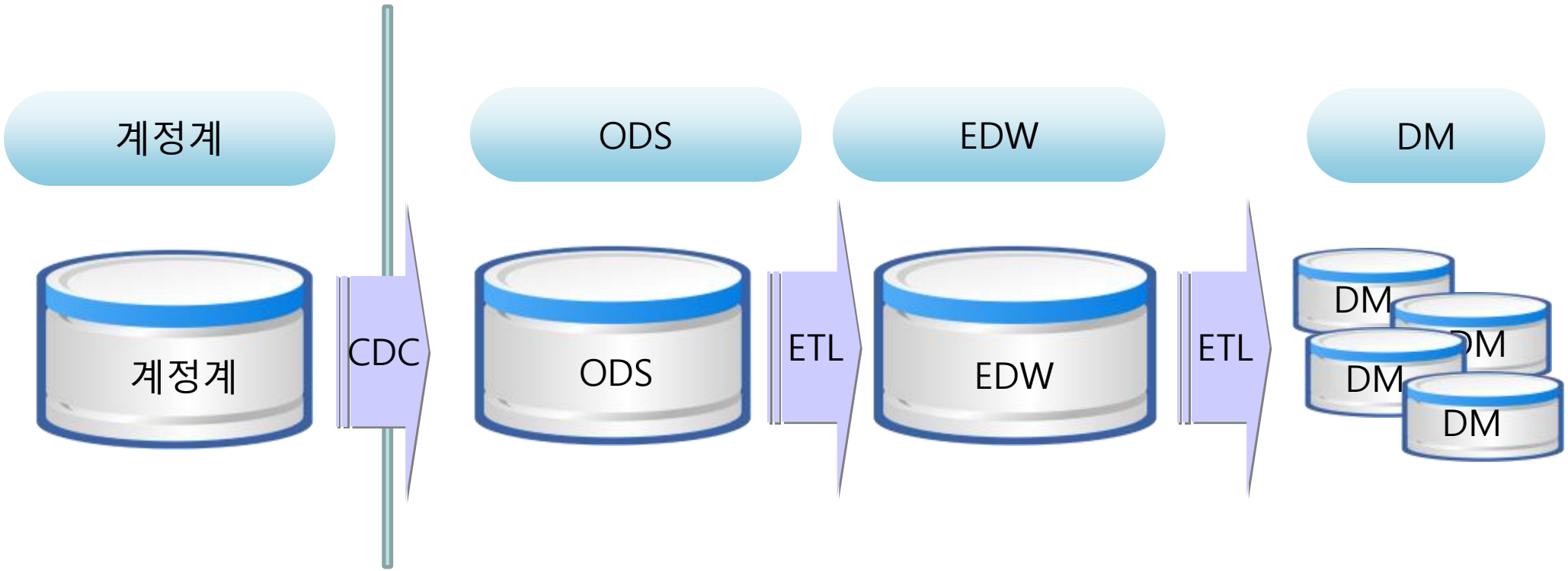
1 ETL 개념

2 CDC 개념

3 대용량 고객 데이터 통합 방안

생각해 보기

- 왜 시스템 및 데이터 처리 방법을 알아야 할까?



[출처] [자체 제작]

학습목차

1 ETL 개념

학습목표

1. ETL가 무엇인지 설명 한다.
2. ELT가 무엇인지 설명한다.
3. ETL 과 ELT 변화 배경을 설명 한다.

ETL 의미

1) ETL 뜻

- 추출, 변환, 적재(Extract, transform, load)
  - **추출, 변환, 적재**(extract, transform, load, ETL)는 컴퓨팅에서 데이터베이스 이용의 한 과정으로 특히 데이터 웨어하우스에 사용 된다.
  - 동일 기종 또는 타기종의 데이터 소스로부터 데이터를 추출한다.
  - 조회 또는 분석을 목적으로 적절한 포맷이나 구조로 데이터를 저장하기 위해 데이터를 변환한다.
  - 최종 대상(데이터베이스, 특히 운영 데이터 스토어, 데이터 마트, 데이터 웨어하우스)으로 변환 데이터를 적재한다.

[출처] https://ko.wikipedia.org/wiki/ETL

1_ETL 의미

2) ETL의 장단점 (1/2)

- 장점
  - 자원(데이터 보관 인프라)의 효율적 사용
  - Compliance 이슈 해결
  - 오랜 기간 발전된 강력한 도구들

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

1_ETL 의미

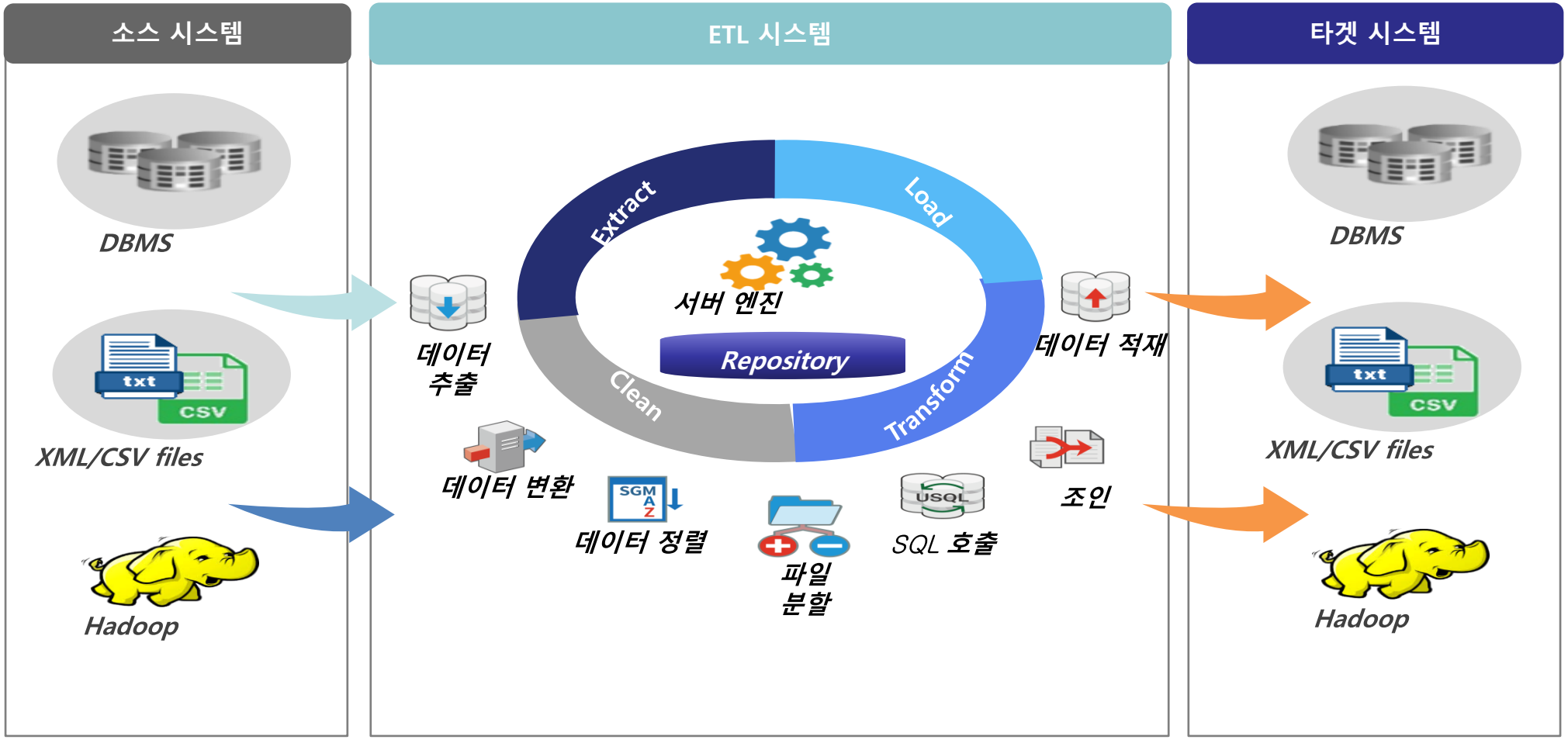
2) ETL의 장단점 (2/2)

- 단점
  - Disk 기반의 성능 문제 (속도 느림)
  - 잦은 관리 필요
  - 높은 수정 비용

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

1_ETL 의미

3) ETL Flow



[출처] 자체제작

1_ETL 의미

4) ETL 솔루션 (1/7)

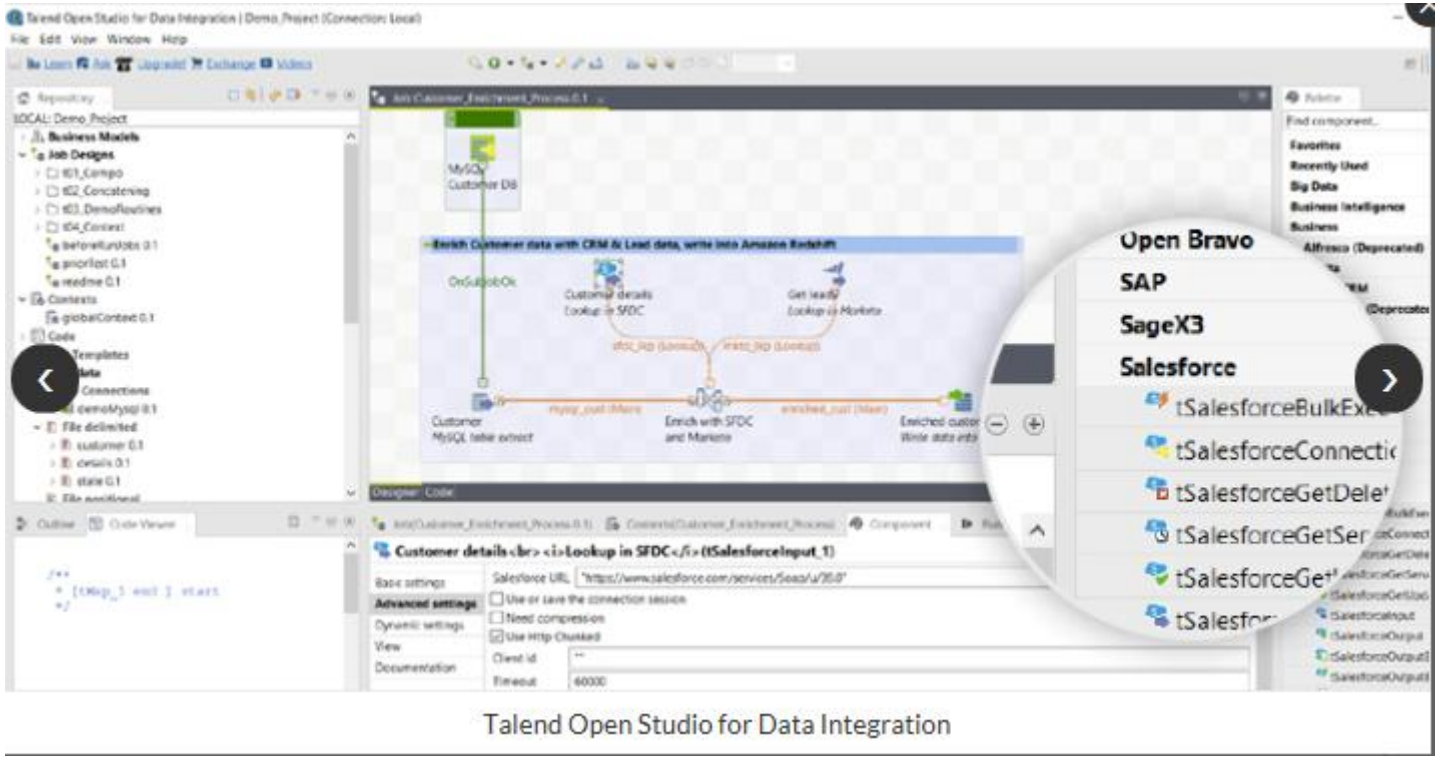
- 오픈 소스 도구
  - Talend Open Studio for Data Integration.
  - Pentaho Data Integration
  - Apache NIFI
  - Jaspersoft ETL
  - KNIME
  - Rhino ETL
  - StreamSets
  - InnoQuartz ETL

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

1_ETL 의미

4) ETL 솔루션 (2/7)

- Talend Open Studio for Data Integration



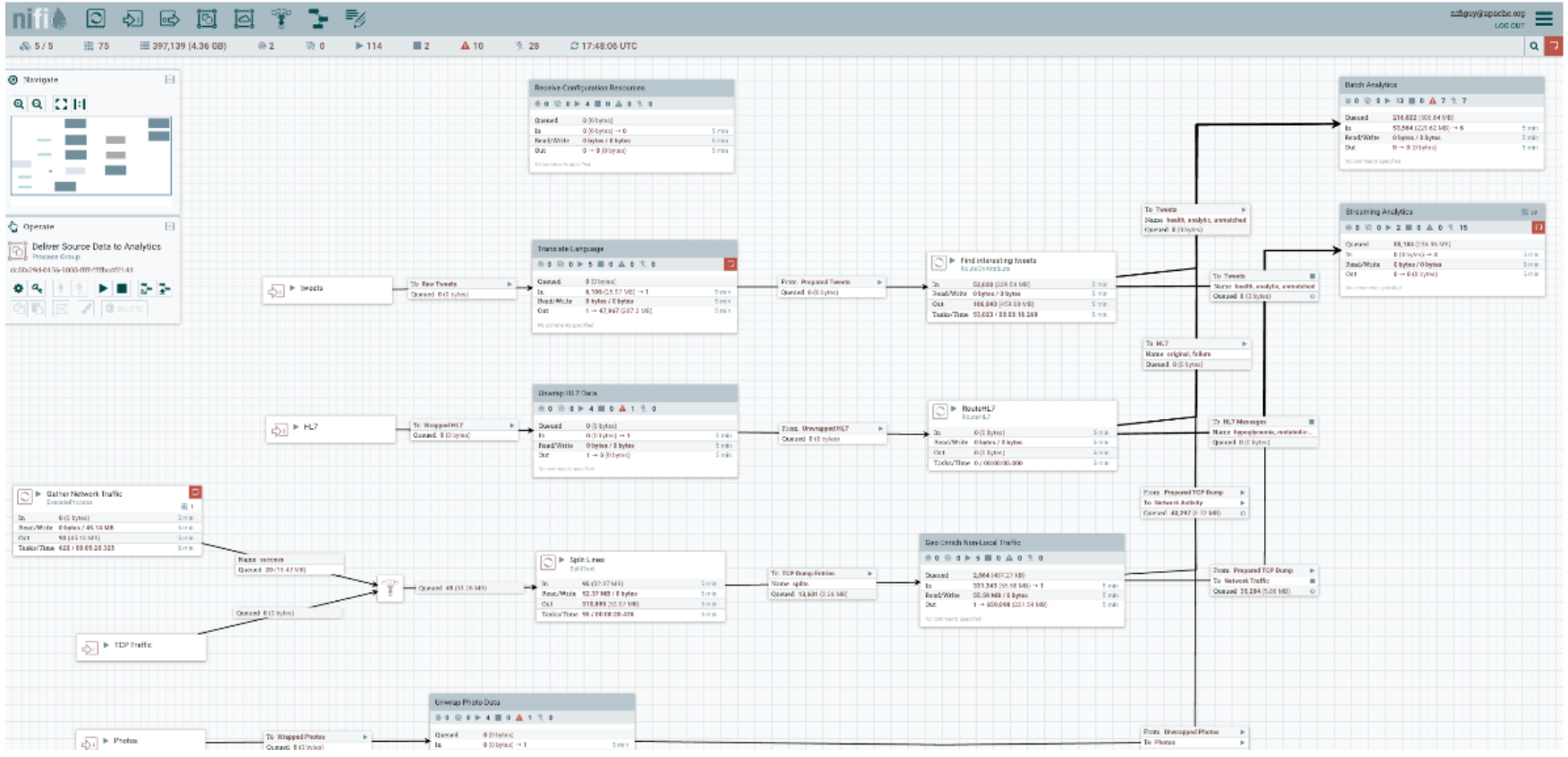
Talend Open Studio for Data Integration

[출처] https://sourceforge.net/projects/talend-studio/

1_ETL 의미

4) ETL 솔루션 (3/7)

- Apache NIFI



[출처] https://nifi.apache.org/assets/images/flow.png

1_ETL 의미

4) ETL 솔루션 (4/7)

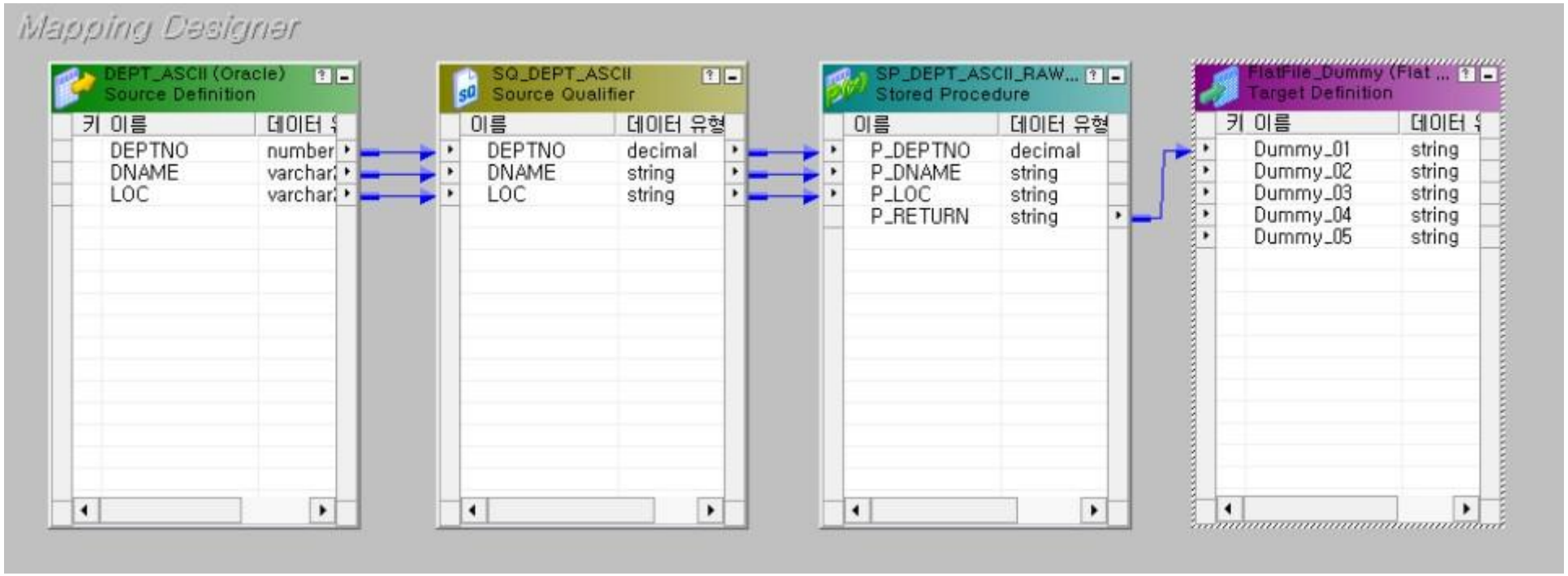
- 상용 소스 도구
  - Informatica PowerCenter
  - Talend
  - InnoQuartz
  - IBM 인포스피어 데이터스테이지
  - 오라클 데이터 인티그레이터 (ODI)
  - SAP 비즈니스 오브젝트 데이터 서비스
  - SAS Data Integration Studio
  - TeraStream (국산)

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

1_ETL 의미

4) ETL 솔루션 (5/7)

- Informatica PowerCenter

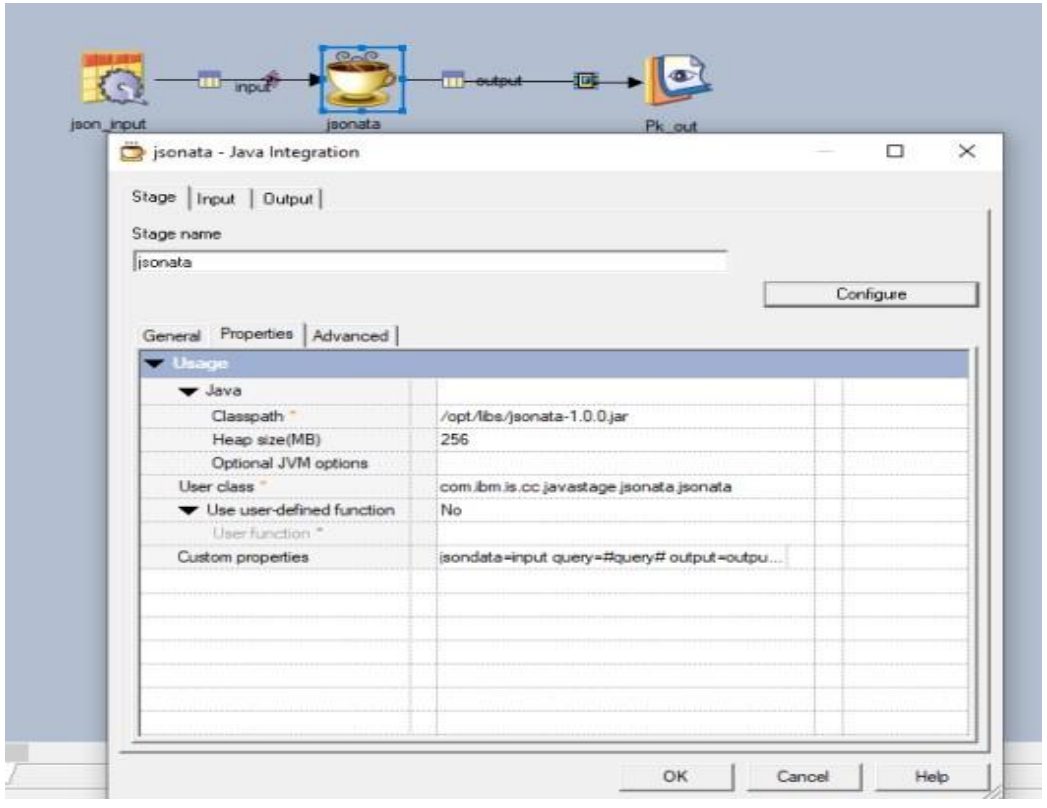


[출처] https://search.naver.com/search.naver?sm=tab_hty.top&where=image&query=powercenter&oquery=Change+Data+Capture&tqi=hR7oOlp0JXVssf9cfhlsssssstGK-212369#imgId=image_sas%3Ablog885065%7C146%7C30160414649_806781298

1_ETL 의미

4) ETL 솔루션 (6/7)

- IBM 인포스피어 데이터스테이지

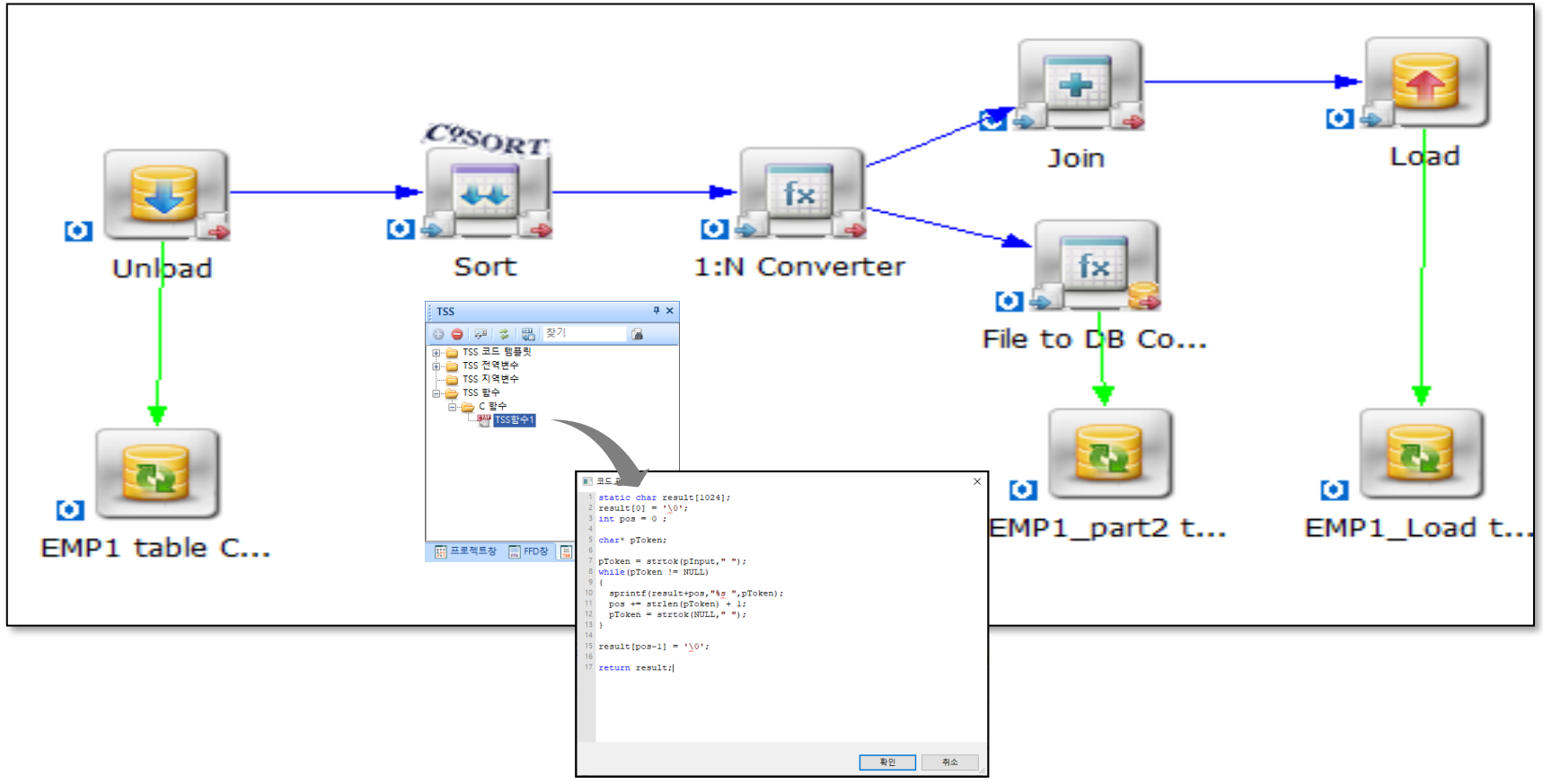


[출처] https://search.naver.com/search.naver?sm=tab_hty.top&where=image&query=datastage&oquery=powercenter&tqi=hR7o1wp0JXoss4MI41RssssstX8-260519#imgId=image_sas%3Awebhttps%3A%2F%2Fgithub.com%2FNSIT-BIM%2FJsonataStage_396994885

1_ETL 의미

4) ETL 솔루션 (7/7)

- TeraStream (국산)



[출처] 자체 제작

2_ELT 의미

1) ELT 뜻

- ETL과 달리 데이터를 추출(Extract)한 이후에 변환없이 그대로 저장(Load)한 후 원하는 방식으로 변환(Transform)/분석하는 방식이다.
- 즉, 모든 데이터를 가능한 저장한 후에 다양한 목적에 따라 원하는 변환을 쉽게 수행 가능하다. (기존 ETL은 새로운 분석요건이 생기면 ETL의 Transformation기능을 반드시 수정해야 가능한 구조였음.)

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

2_ELT 의미

2) Why ELT?

- 최근 AI의 도입이 증가하면서 많은 기업이 데이터 분석가를 양성하며, 분석에 필요한 다양한 데이터를 제공한다.
- 하지만, 기존 ETL을 도입한 기업들은 분석가들이 원하는 데이터가 변경될 때 마다 새로운 Transformation을 개발하여 적용하는데 많은 시간이 소요된다. (물론 데이터 품질도 문제..) 이를 해결하게 해주는 새로운 방식이 ELT 방식이며, 이를 public cloud service에서는 Cloud DW라는 서비스로 접근하고 있다.

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

2_ELT 의미

3) ELT를 가능하게 하는 기술들

- **제약 없는 스토리지 및 컴퓨팅 리소스 활용 (Cloud 기반)**
  - 데이터 저장 공간에 제약이 없게 되고, 어떤 데이터라도 충분한 컴퓨팅 자원을 통해서 쉽게 조회가 가능한 환경을 제공
  - 기존에는 처리할 수 없었던 비정형, 대용량 데이터 처리가 가능한 기술(hadoop, spark 등)의 등장으로 저장후에 모든 작업이 가능함

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

2_ELT 의미

4) ELT의 장단점 (1/2)

- 장점
  - 빠른 데이터 전송 (Extraction & Loading)
  - 낮은 데이터 처리 비용
  - 낮은 운영비용
  - 높은 유연성

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

2_ELT 의미

4) ELT의 장단점 (2/2)

- 단점
  - Overgeneralization(과도한 일반화)
  - 보안 이슈
  - Compliance 이슈
  - 조회 속도(latency) 늦음

[출처] https://blog.naver.com/freepsw/222276087707
ETL vs ELT : 최근 트렌드 변화|작성자 freepsw

3_ETL 과 ELT 변화 배경

1) ETL 과 ELT의 차이점

	ETL	ELT
프로세스	추출, 변환, 적재	추출, 적재, 변환
자료구조	전처리 된 데이터 / 데이터 웨어 하우스 지원	원천 데이터 (Raw Data) / 데이터 레이크 지원
데이터 활용목적	현재 사용 중	미결정 상태
접근성	변경하기 쉽지 않고 비용도 많이 소요됨.	접근성 높고 신속한 업데이트
사용자	비즈니스 현업 전문가	데이터 과학자
시스템의 데이터 가용성	데이터 웨어하우스 및 ETL 프로세스를 생성할 때 필요하다고 결정한 데이터만 변환하고 로드합니다 .	모든 데이터를 즉시 로드할 수 있으며 사용자는 나중에 변환 및 분석할 데이터를 결정할 수 있습니다.
데이터 지원	관계형 SQL 기반 구조	정형, 비정형 등 모든 데이터 유형을 수집
데이터 크기	소량의 데이터로 정교한 데이터 변환에 사용	대용량 데이터에 사용
정보 로드 대기 시간	적재 후 데이터 변환에 다소 시간이 걸리며, ELT보다 느립니다. 그러나 데이터가 로드되면 정보 분석이 ELT보다 빠릅니다.	변환을 기다릴 필요가 없고 데이터가 대상 데이터 시스템에 한 번만 로드되기 때문에 데이터를 빠르게 처리 할 수 있습니다. 그러나 정보 분석은 ETL보다 느립니다.
유지보수	프로세스의 지속적인 유지 관리가 필요하다.	클라우드 기반이며 자동화된 솔루션을 통합하므로 유지 관리가 거의 필요가 없다.

[출처] https://blog.bizspring.co.kr/34792

3_ETL 과 ELT 변화 배경

2) ETL에서 ELT로 흐름이 변화된 배경 (1/2)

- 대량의 데이터 발생
  - 기업형 ETL(DW)는 복잡한 분석, 보고 및 운영을 수행하는데 있어 대부분의 조직에 기본적으로 사용되는 솔루션이었으나,
  - 대용량의 광범위하고 다양한 데이터가 표준이 되는 빅데이터 시대에 데이터를 처리하는데 시간이 오래 걸리는 ETL 프로세는 부적합해졌습니다.

[출처] https://blog.bizspring.co.kr/34792

3_ETL 과 ELT 변화 배경

2) ETL에서 ELT로 흐름이 변화된 배경 (2/2)

- **리소스 /유지보수 비용 부담 해소**
  - ETL(DW)는 ELT는 ETL 데이터하우스(DW)에 비해 클라우드 기반의 합리적인 가격으로 데이터 지속성을 생성하는데 효과적이며,
  - 유지보수 비용이 적어 기업은 모든 비정형 데이터를 클라우드에 저장하고 데이터를 유연하게 처리할 수 있게 되었습니다.
  - 기업에서는 데이터 처리 비용 부담 해소로 더 이상 변환 단계에서 데이터를 줄이거나 필터링할 필요가 없어졌으며, 이러한 배경에서 ETL에서 ELT로 흐름이 변화되고 있습니다.

[출처] https://blog.bizspring.co.kr/34792

3_ETL 과 ELT 변화 배경

3) ETL 과 ELT 사용 사례 (1/2)

- ETL 사용 사례
  - ETL 프로세스는 대상의 기존 데이터에 맞게 (데이터는 모양, 데이터 형식, 언어, 표준시간대등) 변환하고, 정교한 데이터 변환을 수행합니다. 정밀도를 높일 수 있습니다.
  - ETL 프로세스는 새 데이터를 기존 데이터와 결합하여 보고를 최신 상태로 유지하거나, 기존 데이터에 대한 추가 인사이트를 제공할 수 있습니다.
  - 보고 도구나 서비스 같은 애플리케이션이 데이터를 원하는 형식으로 사용할 수 있습니다.

[출처] https://blog.bizspring.co.kr/34792

3_ETL 과 ELT 변화 배경

3) ETL 과 ELT 사용 사례 (1/2)

- ELT 사용 사례
  - ELT는 구조화된 데이터와 구조화되지 않은 데이터의 방대한 양에서 가장 잘 작동합니다. 대상 시스템이 클라우드 기반인 한 ELT 솔루션을 사용하여 이러한 엄청난 양의 데이터를 더 빠르게 처리할 수 있습니다.
  - 필요한 처리 능력을 처리할 수 있는 리소스가 있는 조직입니다.
    ETL을 사용하면 대부분의 처리가 데이터가 웨어하우스에 도착하기 전에 파이프라인에 있는 동안 발생합니다.
    ELT는 데이터가 이미 데이터 레이크에 도착하면 작업을 수행합니다.
  - 모든 데이터가 한 곳에서 가능한 한 빨리 필요한 회사입니다.
    좋은 데이터든 나쁜 데이터든 모든 데이터가 나중에 변환을 위해 데이터 레이크에 저장됩니다.

[출처] https://blog.bizspring.co.kr/34792

학습목차

2 CDC 개념

학습목표

1 CDC가 무엇인지 설명 한다.

2 EAI가 무엇인지 설명한다.

1. CDC의 의미

1) CDC 뜻

- Change Data Capture (CDC)
  - identifies and captures only the source data that has changed and moves that data to the target system. CDC can be used to reduce the resources required during the ETL “extract” step; it can also be used independently to move data that has been transformed into a data lake or other repository in real time

    (변경된 소스 데이터만 식별 및 캡처하고 해당 데이터를 대상 시스템으로 이동합니다. CDC는 ETL "추출" 단계에서 필요한 리소스를 줄이는 데 사용할 수 있습니다. 데이터 레이크 또는 다른 리포지토리로 변환된 데이터를 실시간으로 이동하는 데 독립적으로 사용할 수도 있습니다.)

[출처] https://www.ibm.com/cloud/learn/etl

1. CDC의 의미

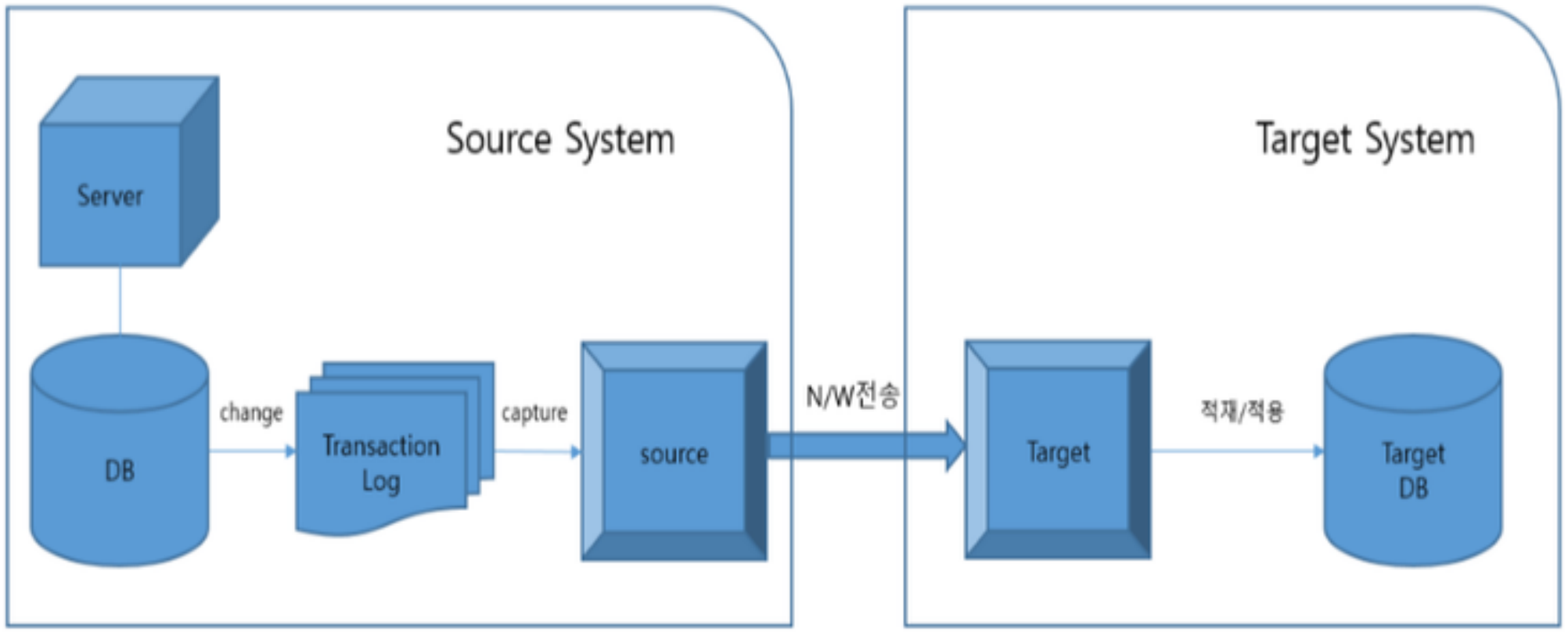
2) CDC 특징

- Change Data Capture (CDC) **특징**
  - 운영중인 시스템에 부하를 주지 않고 실시간 전달
  - 다이나믹 웨어하우싱, BI, 레포팅을 위한 분석 지원
  - 타임스탬프 기록이나 테이블 구조 변경이 필요 없음
  - 데이터 변경 사항이 기록된 로그 만을 전송

[출처] https://blog.naver.com/stjdsmtjs/221644127400
CDC(=변경 데이터 캡쳐=Change data capture)|**작성자** 공부의 신

1. CDC의 의미

3) CDC Flow



[출처] https://blog.naver.com/stjdsmtjs/221644127400

CDC(=변경 데이터 캡쳐=Change data capture)|**작성자** 공부의 신

1. CDC의 의미

4) CDC 솔루션 (1/4)

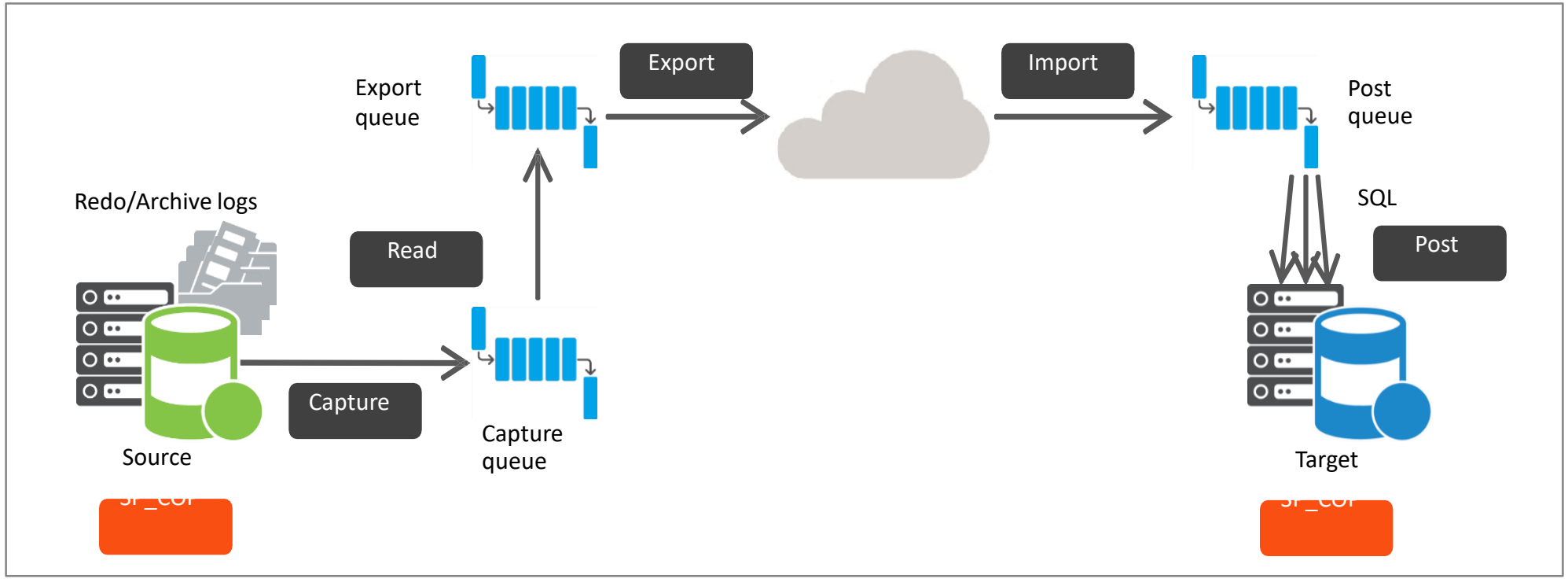
- Change Data Capture (CDC) 솔루션
  - SharePlex
  - Ark-CDC
  - DeltaStream

[출처] https://blog.naver.com/stjdsmtjs/221644127400
CDC(=변경 데이터 캡쳐=Change data capture)|작성자 공부의 신

1. CDC의 의미

4) CDC 솔루션 (2/4)

- SharePlex 아키텍쳐



[출처] SharePlex 제품소개서

1. CDC의 의미

4) CDC 솔루션 (3/4)

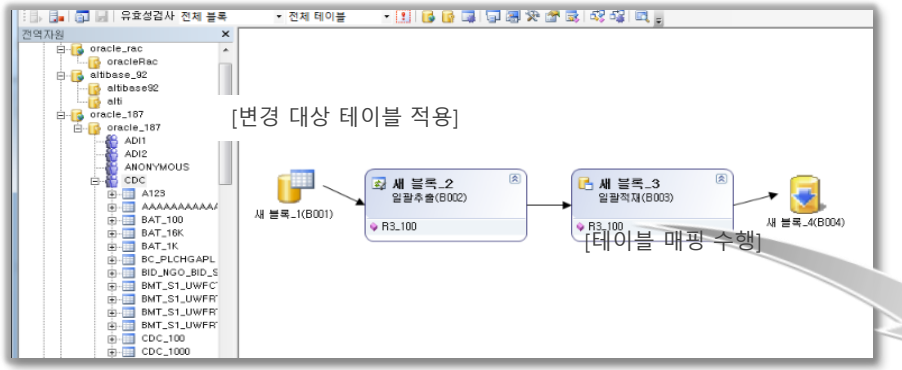
- Ark-CDC 아키텍쳐

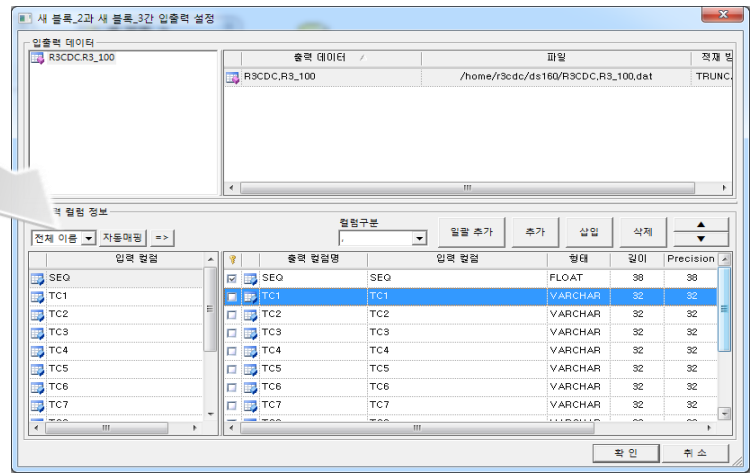
[출처] www.idatabank.com 제품소개서

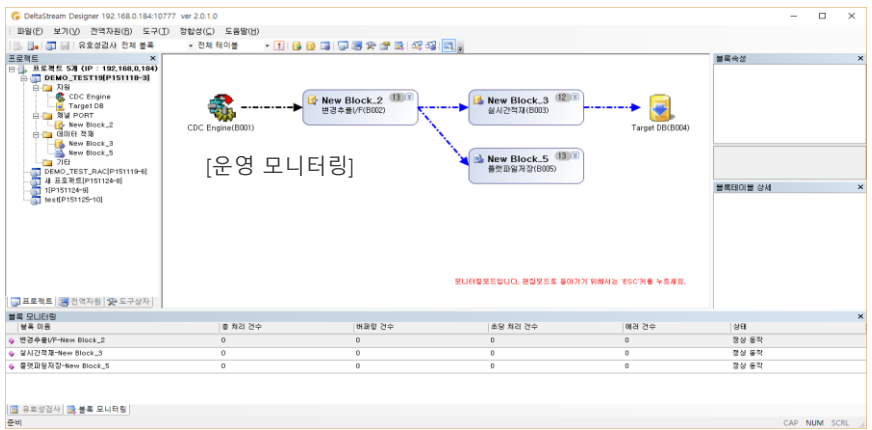
1. CDC의 의미

4) CDC 솔루션 (4/4)

- DeltaStream **아키텍쳐**







[출처] https://blog.naver.com/stjdsmtjs/221644127400

CDC(=변경 데이터 캡쳐=Change data capture)|**작성자** 공부의 신

1. CDC의 의미

5) CDC 도입시 고려사항

- 고려 사항

유의사항	설명
No Logging 트랜잭션 대응	- 성능향상을 위한 No-Logging처리 시, 전체 데이터 재 동기화가 효율적임
대량 데이터 처리 성능	- 동기화 지연 발생으로 갱신주기별 분리 수용
테이블 구조의 변경 연계	- 스키마 변경 오류 대응
암호화 데이터 처리	- 복호화 필요성 고려
Supplemental Log설정	- 설정 변경 시 REDO Log의 길이 증가(영향도 확인)
양방향 동기화	- 동일 테이블에 대한 주의

[출처] https://blog.naver.com/stjdsmtjs/221644127400
CDC(=변경 데이터 캡쳐=Change data capture)|작성자 공부의 신

2. EAI 의미

1) EAI 개요 (1/2)

- 개요
  - 정의 : 기업정보시스템들의 데이터를 연계, 통합하는 소프트웨어/정보시스템 아키텍처 프레임워크
  - 방식 : Hub and spoke 방식

    – 미들웨어(Hub)를 이용하는 일반적인 방식의 EAI.

    – 허브를 채용하여 유지보수에 뛰어나지만 허브에 이상이 생기면 전체 기능에 장애가 생김
  - 구성 요소 : 정보시스템, Adapter, BUS, Broker, Transformer

** EAI(Enterprise Application Integration)

[출처] https://blog.naver.com/kmrie/221543069778

2. EAI 의미

1) EAI 개요 (2/2)

- Hub and spoke 방식
  - 허브 앤 스포크 시스템이란?
    자전거 바퀴살(Spoke)이 중심축(Hub)으로 모이는 것처럼 물류가 거점으로 집중된 후 다시 개별 지점으로 이동하는 운송!

    국가물류통합정보센터 물류용어사전

[출처] http://cjkx.tistory.com/CJ대한통운운수대통
[허브 앤 스포크(Hub And Spoke)]|작성자 aika2085

2. EAI 의미

2) 활용 시 기대 효과

- Hub 활용 시 기대 효과
  - 향후 정보시스템 개발/운영 비용 절감
  - 정보시스템의 지속적 발전 기반 확보
  - 인터넷 비즈니스를 위한 기본 토대

[출처] https://blog.naver.com/kmrie/221543069778