

빅데이터 저장

성공회 대학교 김종율 교수

1. 데이터 웨어하우스

- 1) **정의:** 데이터 웨어하우스란 사용자의 의사결정에 도움을 주기 위해 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스를 말한다. 줄여서 DW로도 불린다(위키백과).
- 2) **데이터웨어하우스와 BI:** 데이터 웨어하우스는 조직 내의 방대한 데이터를 효율적으로 관리하여 주요 경영의사결정의 기초를 제공하는 역할을 수행한다. 따라서 필수적으로 데이터 웨어하우스는 비즈니스 인텔리전스(Business Intelligence, BI)라 불리는 일종의 프로세스를 동반한다. 정리하면, 경영진이 주요 경영의사결정을 하기 위해서는 데이터를 통합 관리하는 시스템인 데이터 웨어하우스가 필요하며, 이 데이터 웨어하우스로부터 얻은 정보를 분석 및 가공해서 개발자가 아닌 경영진이 보기 쉽게 시각화해서 보여주는 프로세스가 비즈니스 인텔리전스다.

3) 데이터 웨어하우스의 특징:

데이터 웨어하우스는 경영의사결정을 위해 태어났다고 해도 과언이 아니다. 기존의 데이터베이스 관리 소프트웨어인 DBMS는 신속하고 정확하게 고객과 시장을 분석해서 의사결정을 위한 기초로 활용하기에는 한계가 있었다. 이는 DBMS의 OLTP(Online Transaction Processing)시스템이 가지는 한계이기도 했다. 그래서 데이터 웨어하우스를 구축하고 OLAP(Online Analytical Processing) 시스템 기반의 비즈니스 인텔리전스 프로세스를 구행하는 쪽으로 데이터 관리는 변화하게 된다. OLAP는 데이터 분석 및 관리의 측면이지만, 비즈니스 인텔리전스는 경영진의 주요 경영 의사결정 측면에서 바라보는 좀 더 넓은 개념으로, BSC, OLAP, 데이터마이닝, ETL, DW 등을 모두 포괄하는 개념이다.

- 4) **데이터 웨어하우스의 특징:** 데이터웨어하우스는 사용자의 의사결정을 지원하기 위해 기업이 축적한 많은 데이터를 사용자 관점에서 주제별로 통합하여 운영시스템과 사용자(경영진) 사이의 별도의 장소에 저장해 놓은 데이터베이스 개념이기 때문에 업무나 부서 중심이 아니라 주제 중심이다. 또 혼재한 데이터를 통합하는 특징을 갖는다. 시간의 흐름에 따른 정보의 변화를 알 수 있으며, 원본 데이터를 훼손 및 변경하지 않고 오로지 경영의사결정을 위한 리포팅에 포커스를 둔다.

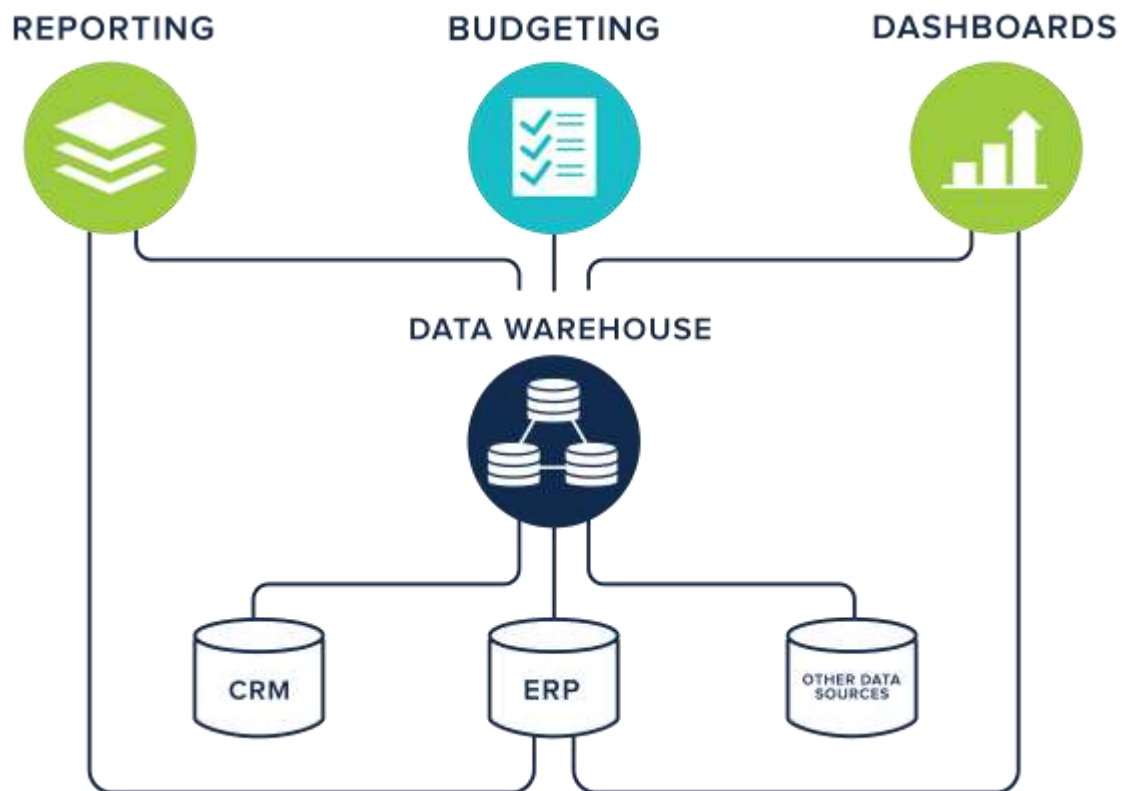
주제지향성=주제중심성 (Subject Oriented)	업무 중심이 아닌 주제 중심 최종 사용자가 이해하기 쉬운 형태를 가짐
통합성(Integrated)	혼재한 DB로부터의 데이터 통합
시계열성(Time Variant)	각각의 원천 DB는 최신 데이터를 보유하고 있지만, DW는 시간에

	다른 변경 정보(이력 데이터)를 보유
비휘발성=영속성 (Non-volatile)	DW의 데이터는 최초 저장 후에는 삭제나 변경되지 않는 읽기 전용 (read only)속성을 가짐 분석 및 리포팅을 위한 기능에 충실

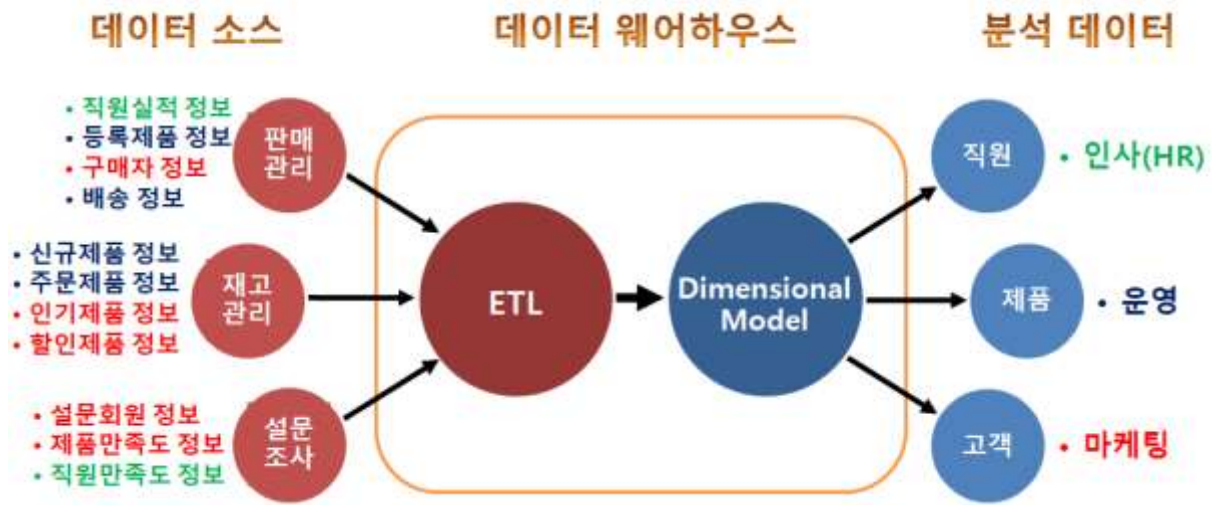
5) 데이터 웨어하우스 구성

데이터 웨어하우스는 전사적자원관리(ERP), 고객관계관리(CRM), 공급망관리(SCM) 등 기업에서 활용되는 다양한 시스템에서 생성되는 데이터를 한 곳에 담아두고, 분석이 필요할 때 이 창고의 데이터를 대상으로 분석하는 방식이다.

CRM, ERP, Data-Source 등의 원천 DB로부터 데이터 웨어하우스에 적재하는 대표적인 방식으로 ETL(Extraction, Transformation and load)과 CDC(Change Data Capture)가 있다. 이 둘은 CRM, ERP, Data-source 등의 원천 DB로부터 데이터 정보를 추출하여 데이터웨어하우스에 적재하는 개념은 동일하나, 그 방법과 목적, 적재 수준에서 서로 차이점이 있다.



[데이터 웨어하우스 구성]



출처: <https://chankim.tistory.com/10>

데이터 소스에서 데이터를 수집하기 위해, 여러 데이터베이스에서 필요한 데이터 소스를 추출(Extract)하고 통합 저장할 수 있도록 동일한 구조로 변환(Transform)한 다음 테이블(Relation)에 적재(Load)하는 통합(재구조화)

이 과정이 데이터 웨어하우스 안에서 이뤄지지만, 먼저 ETL이 이루어지고 데이터 웨어하우스로 들어오는 경우도 있다.

그 다음 과정은 모델링으로 ETL 이후 테이블(Relation)을 차원모델(Dimensional Model)로 저장한다. 차원모델링의 목적은 사용자가 데이터를 분석할 때 빠르고 쉽게 데이터 검색을 할 수 있도록 데이터베이스를 최적화하는 것으로 데이터 소스에서 ETL을 통해 가공된 데이터는 정규화된 데이터베이스에 저장된다.

사용자가 분석을 하려면 숫자정보(수량, 가격 등)의 집계가 필요한데 정규화된 데이터는 집계 시 JOIN이 많이 발생해 속도가 느려지는 문제가 있기 때문에 차원 모델링은 분석에 최적화되도록 숫자정보 중심으로 테이블을 비정규화 시킨다. (숫자 중심의 팩트 테이블과 그 주위에 차원 테이블로 구성)

6) 데이터 웨어하우스의 테이블 모델링 기법

- 스타 스키마(Star Schema)

조인 스키마(Schema)라고도 하며, 단일 테이블(Fact Table)을 중심으로 다수의 차원 테이블(Dimensional Table)이 연결되어 있다.

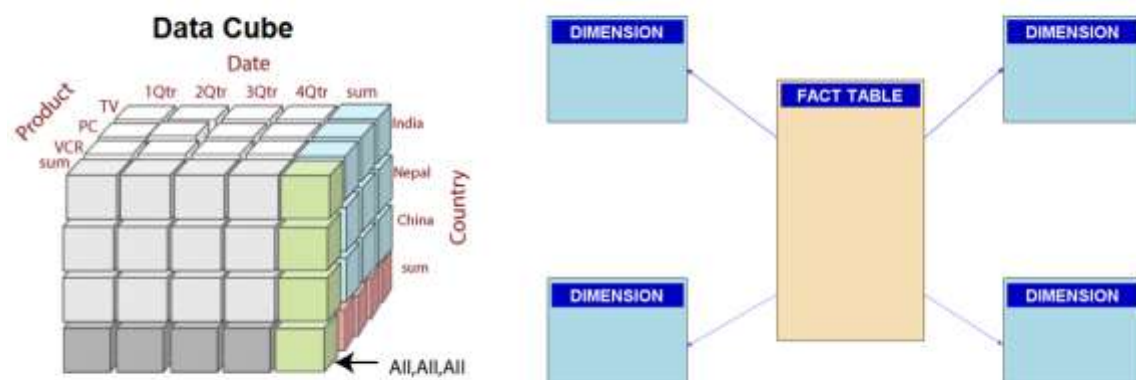
사실 테이블 내에 있는 데이터에 대한 검색 기준으로 차원 테이블의 열을 사용함으로써 응답에 필요한 조인의 횟수를 최소화하지만, 확장성과 유연성이 제한된다.

스타 스키마의 사실 테이블은 보통 제3 정규형으로 모델링하며, 차원 테이블은 제2 정규형으로

모델링하는 것이 일반적이다.

스타 스키마 장단점:

장점	복잡도가 낮아서 이해하기 쉽다 쿼리 작성이 용이하고 조인의 테이블 수가 적다
단점	차원 테이블들의 비정규화로 데이터 중복이 발생하여 상대적으로 데이터 적재에 시간이 많이 소요된다



다차원 큐브 모델을 사용하면 데이터를 모델에 적재하는 시점에 미리 계산을 수행하고 필요한 인덱싱 작업을 진행하기 때문에 높은 성능을 발휘할 수 있다. 그래서 대용량의 데이터를 처리할 때에는 다차원 큐브 모델을 많이 선택했으나 데이터를 적재하는 데에 오랜 시간이 걸린다는 단점이 존재한다.

관계형 모델을 사용하는 경우에는 RDMBS의 특징에 따라 백업이나 복원하는 것이 용이하고 익숙한 데이터베이스 형식으로 모델링을 수행할 수 있다는 장점이 있다. 전통적인 큐브 모델보다 성능이 떨어진다는 단점이 존재했지만 최근에는 클라우드 환경의 사용과 메모리의 증가로 다차원 큐브 모델과 성능적인 부분에서도 큰 차이가 없기 때문에 관계형 모델을 사용한 모델링이 많이 진행되고 있다.

- 스노우플레이크 스키마

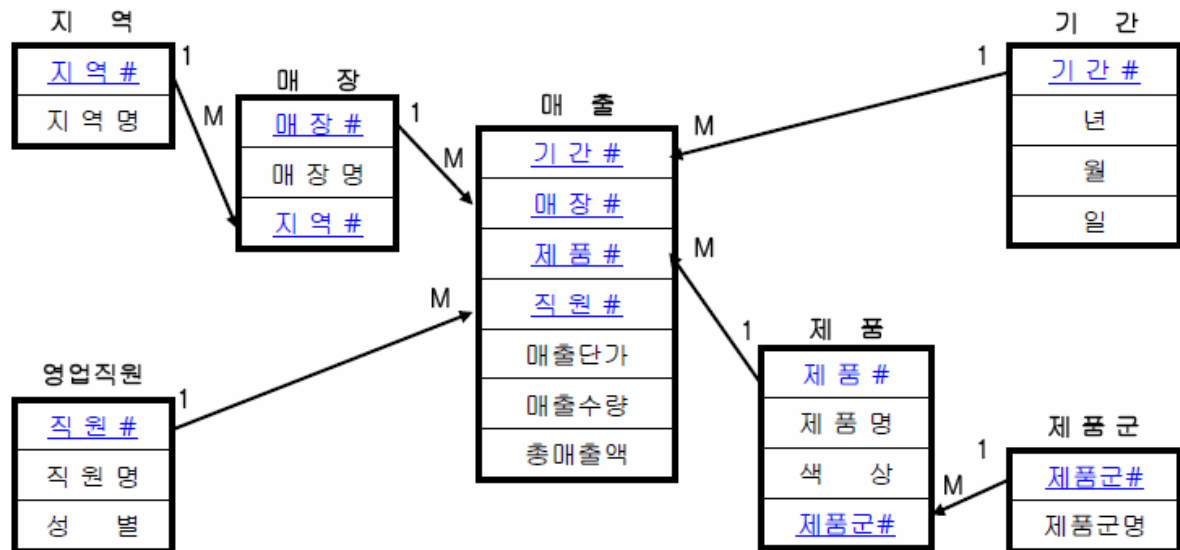
스타 스키마의 차원 테이블이 제3 정규형으로 정규화된 형태.

사실 테이블은 그대로 유지되는데, 꼬리에 물고 차원 테이블이 계속 등장한다.

그래서 저장공간이 최소화되고 유연성이 증가하지만, 복잡하고 결과 검증이 어렵다.

장점	데이터 중복 제거, 디스크 절감 적재시간 단축
----	------------------------------

단점	스키마 구조의 복잡성 증가 조인 테이블 개수 증가(쿼리 작성의 난이도 증가)
----	---



- 정규화: 어떤 대상을 일정한 규칙이나 기준에 따르는 정규적인 상태로 바꾸거나 비 정상적인 대상을 정상적으로 되돌리는 과정

- 1차 정규화: 반복되는 속성 제거

- 좌측 테이블은 비정규 테이블임('수강과목' 필드가 원자값을 갖지 않으므로)
- 속성 '수강과목'을 분해하고 '학번'과 '수강과목'의 복합키를 기본키로 지정함으로써, 우측 테이블과 같은 1차 정규형 테이블을 도출할 수 있음

학번	이름	과목명
1111	홍길동	데이터베이스, 운영체제
2222	강감찬	운영체제, 네트워크, 자료구조

➔

학번	이름	과목명
1111	홍길동	데이터베이스
1111	홍길동	운영체제
2222	강감찬	운영체제
2222	강감찬	네트워크
2222	강감찬	자료구조

- 2차 정규화 : 부분 함수 종속성 제거

- 좌측 테이블은 2차 정규형이 아님
- (학번, 과목코드)→(과목명)의 함수 종속에서 (과목코드)→(과목명)의 부분 함수 종속이 존재하기 때문
- 이를 우측과 같이 두 개의 테이블로 분할함으로써 2차 정규형 테이블을 도출함

학번	이름	과목코드	과목명	학점
1111	홍길동	D11	데이터베이스	A
1111	홍길동	O22	운영체제	B
2222	강감찬	O22	운영체제	A
2222	강감찬	N33	네트워크	A
2222	강감찬	D44	자료구조	B

→

학번	이름	과목코드	학점
1111	홍길동	D11	A
1111	홍길동	O22	B
2222	강감찬	O22	A
2222	강감찬	N33	A
2222	강감찬	D44	B

+

과목코드	과목명
D11	데이터베이스
O22	운영체제
O22	운영체제
N33	네트워크
D44	자료구조

- 3차 정규화 : 이행 함수 종속성 제거

- 좌측 테이블은 3차 정규형이 아님
- (학번)→(학과), (학과)→(학과사무실)의 이행 함수 종속이 존재하기 때문
- 이를 우측과 같이 두 개의 테이블로 분할함으로써 3차 정규형 테이블을 도출함

학번	이름	과목코드	학점
1111	홍길동	컴퓨터공학과	공학관
2222	강감찬	컴퓨터공학과	공학관
3333	유관순	물리학과	자연관

→

학번	이름	학점
1111	홍길동	컴퓨터공학과
2222	강감찬	컴퓨터공학과
3333	유관순	물리학과

+

과목코드	과목명
컴퓨터공학과	공학관
컴퓨터공학과	공학관
물리학과	자연관

- 보이스 코드 정규화 : 결정자 함수 종속성 제거

- 좌측 테이블에서 (교수, 과목명)→(교재명), (교재명)→(과목명)의 함수 종속 관계가 존재한다고 가정
- 이때 좌측 테이블은 보이스-코드 정규형이 아님
- 결정자가 후보키가 아닌 함수 종속(교재명)→(과목명)이 존재하기 때문
- 이를 우측과 같이 두 개의 테이블로 분할함으로써 보이스-코드 정규형 테이블을 도출

교수	과목명	교재명
P1	자료구조	Book1
P1	네트워크	Book2
P2	네트워크	Book3
P2	프로그래밍	Book4
P3	프로그래밍	Book4

→

교수	교재명
P1	Book1
P1	Book2
P2	Book3
P2	Book4
P3	Book4

+

교재명	과목명
Book1	자료구조
Book2	네트워크
Book3	네트워크
Book4	프로그래밍

- 4차 정규화 : 다중값 종속성 제거

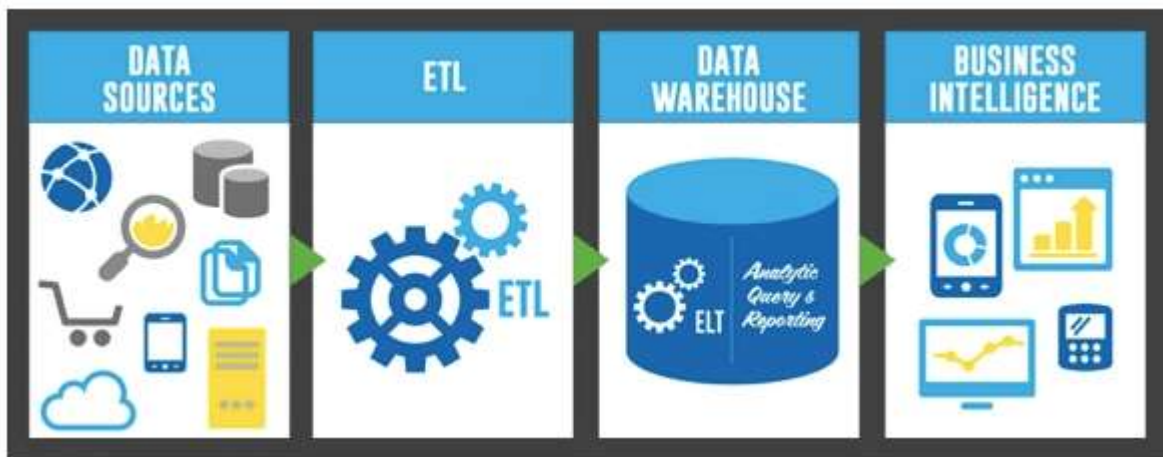
- 5차 정규화 : 결합 종속성 제거

7) 데이터 웨어하우스에서의 데이터 처리 프로세스

- ETL(Extraction, Transformation and Load)

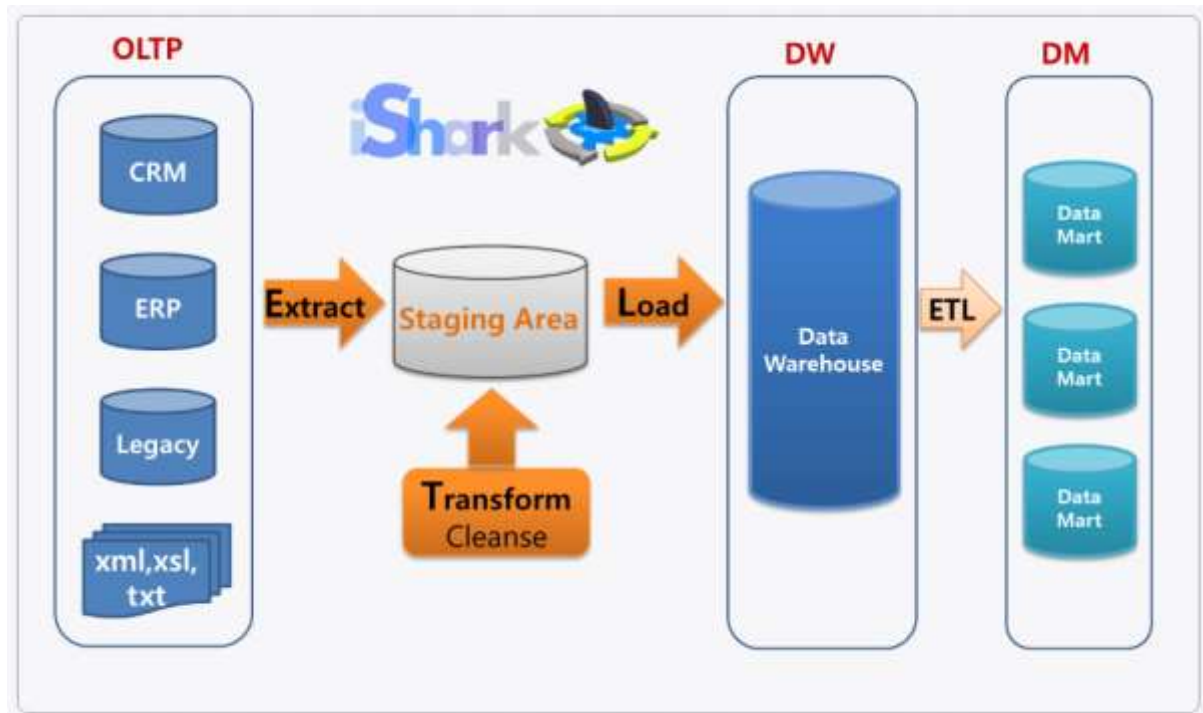
ETL 은 데이터 이동과 변환 절차와 관련된 업계 표준 용어로, 데이터 원천으로부터 데이터를 추출 및 변환하여 운영 데이터 스토어(ODS:Operational Data Store), 데이터 웨어하우스(Data Warehouse), 데이터 마크()Data Mart)에 데이터를 적재하는 작업을 말한다.

ETL 은 크게 일괄 ETL(Batch ETL)과 실시간 ETL(Real Time ETL)로 구분된다. 이때 대용량 데이터를 처리하기 위해 MPP(Massive parallel Processing, ETL 작업 단계에서 프로그램을 여러 파트로 나누어 다수의 프로세서가 동시에 처리될 수 있게 하는 일종의 병렬처리 프로세스)를 지원할 수 있다.



- ETL 기능

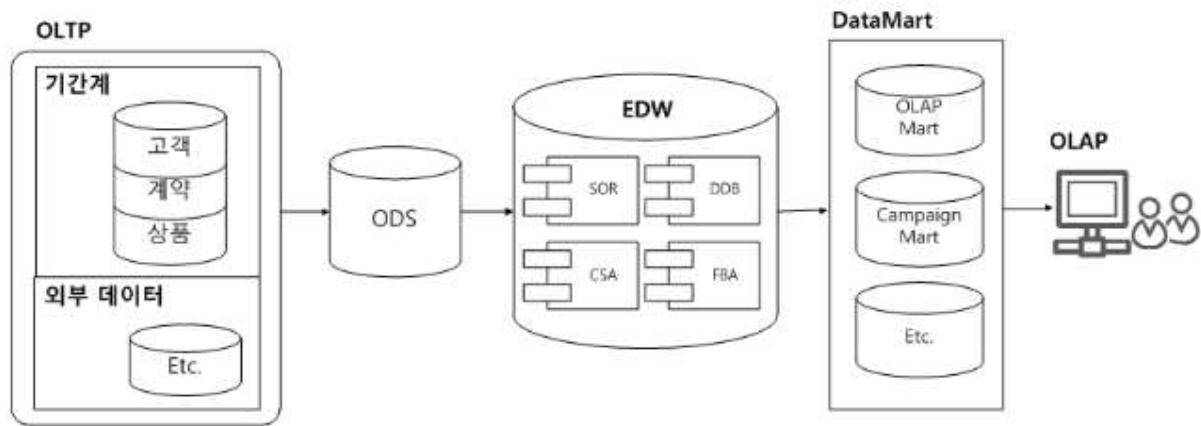
추출(Extraction)	하나 또는 그 이상의 데이터 원천으로부터 데이터를 획득
변형(Transformation)	데이터 클렌징, 형식 변환, 표준화, 통합 또는 다수 애플리케이션에 내장된 비즈니스를 적용 등
적재>Loading)	위 변형 단계 처리가 완료된 데이터를 특정 목표 시스템에 적재



8) ODS

- ODS 개념 및 특징:

- ODS 는 데이터에 대한 추가작업을 위해 다양한 데이터 원천들로부터 데이터를 추출 및 통합한 데이터베이스
- ODS 내의 데이터는 향후 비즈니스 지원을 위해 정보 시스템으로 이관되거나, 다양한 보고서 생성을 위해 데이터 웨어하우스로 이관된다.
- 다양한 원천들로부터 데이터가 구성되기 때문에, ODS 를 위한 데이터 통합은 일반적으로 데이터 클린징, 중복제거, 비즈니스 룰 대비 데이터 무결성 점검 등의 작업들을 포함한다.
- ODS 는 일반적으로 실시간(Real Time) 또는 실시간 근접(Near Real Time) 트랜잭션 데이터 혹은 가격 등의 원자성(개별성)을 지닌 하위 수준 데이터를 저장하기 위해 설계된다.



9) CDC (change Data Capture)

CDC 는 데이터베이스 내 데이터에 대한 변경을 식별해 필요한 후속처리(데이터 전송/공유 등)를 자동화하는 기술 또는 설계 기법이자 구조다.

원천데이터를 DW, DM 등에 적재한다는 의미에서는 ETL 과 비슷하지만, 실시간(Real Time) 혹은 준실시간(Near Real Time)으로 적재하 점에서 큰 차이가 있다.

CDC 는 실시간 또는 준실시간 데이터 통합을 기반으로 하는 데이터 웨어하우스 및 기타 데이터 저장소 구축에 폭넓게 활용된다.

- ETL 과 CDC 의 차이

	ETL	CDC
공통 목적	원천 데이터를 SW, DM 등에 적재	원천 데이터를 SW, DM 등에 적재
특징	실시간이 아닌 정해진 시점의 완료된 데이터를 적재	실시간(Real Time)혹은 준실시간(Near Real Time)으로 적재
용도 예	거래 집계, 일일 회계 집계, 원장 등 이벤트 단위가 아닌 소스 데이터 취합 용도	이상 감지 경고, 금융 거래 이상 경고 변경된 이벤트 감지 용도
기술	변경 데이터만 적재 혹은 All Copy	변경 이력을 관리하는 DB Archive log
적재 수준	적재 시점의 적재 수준(시, 일, 월 등)	시점에 관계 없이 모든 원천 데이터의 변경 로그 기록을 적재

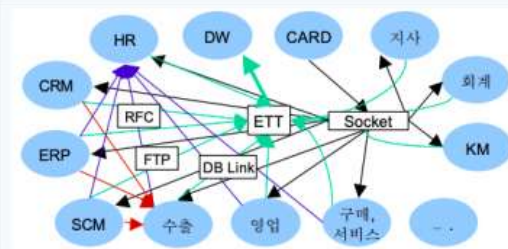
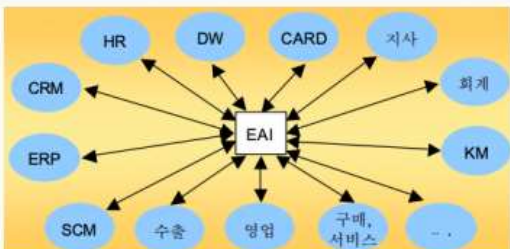
10) EAI (Enterprise Application Integration)

엔터프라이즈의 미들웨어를 인프라로 하여 다양한 이질적 기업환경(애플리케이션, 데이터, 플랫폼 및 네트워크 등)을 통합하여 하나의 시스템으로 관리 운영할 수 있는 유기적인 시스템(가트너)

한 기업 내의 ERP, CRM, SCP, 인트라넷 등의 시스템 간에는 서로 데이터를 주고 받아야할 필요성이 존재합니다. 이와 같이 기업 내 필요한 여러 어플리케이션 간에 상호 연동이 가능하도록 통합하는 솔루션

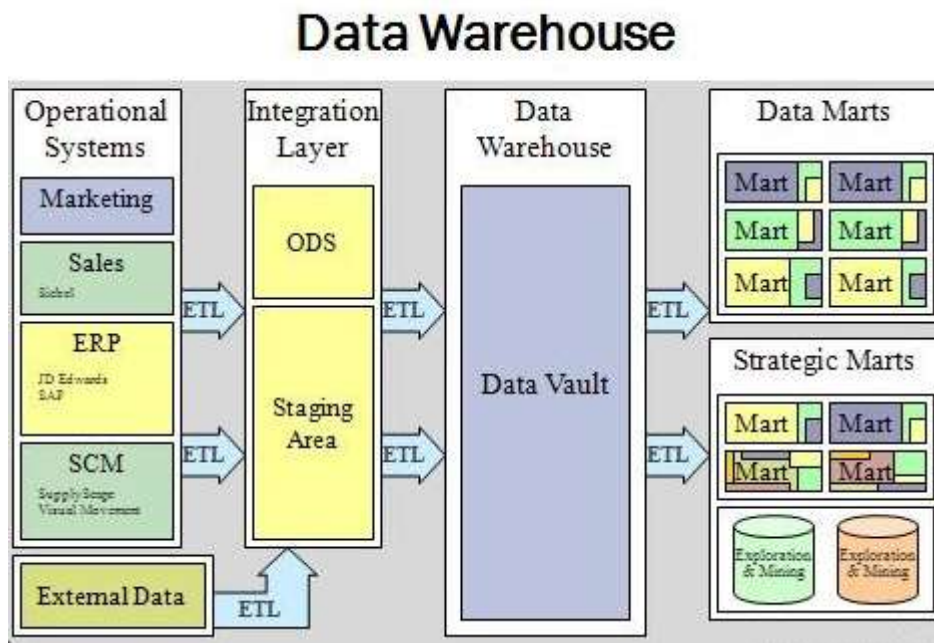
예를 들어, 데이터를 주고받기 위해 각 시스템 간에 개별적으로 서로 통신을 한다면 시스템 간에 개별적인 연결이 상당히 많이 생성되게 됩니다. 이런 경우 유지보수의 어려움과 시스템 간 통신 등의 다양한 문제가 발생할 수 있으며, 이런 문제점을 해결하기 위해 EAI 라는 솔루션을 적용한다.

〈 EAI 도입 전 후 비교 〉

구분	도입 전	도입 후
개념도		
구성	<ul style="list-style-type: none"> 지역/업무별 시스템 단위 운영관리 Interface 분산화/비표준화/복잡화 Data 정합성 관리 및 Maintenance 어려움 개발 및 Maintenance 비용 증가 	<ul style="list-style-type: none"> 전사 통합 운영관리 Interface 통합화/표준화/단순화 Data 정합성 보증 및 Maintenance 용이 개발 및 Maintenance 비용 절감
기반 기술	<ul style="list-style-type: none"> FTP, DB Link, Socket 통신등] 구성 방식의 혼재 및 업무별 API 개발 업무별 단위 시스템에 종속된 Maintenance 	<ul style="list-style-type: none"> 통일된 EAI 도구에 의한 구성 및 Interface 개발 Interface의 통합 관리를 고려한 Maintenance

- EAI 특징

기존 Point to Point 데이터 연계	EAI 연계
단위 업무 위주의 개발로 인해 업무 간 통합성 및 이해 부족	실시간 정보 및 프로세스 동기화
데이터 정합성 관리의 어려움	통합 모니터링 관리
개발 및 유지보수 비용 증가	추가 개발 및 유지보수 비용 절감
중복 작업의 산재 가능성	통합된 커넥션 관리
정보 공유 인식 부재	데이터 정합성 및 무결성 유지
	장애 대응 용이



2. 데이터레이크(Data Lake)

빅데이터 시대에 접어들면서 나날이 증가하는 방대한 데이터와 새로운 포맷의 데이터를 수집하고 축적·활용해야 한다는 환경의 니즈가 증가하면서 기업들은 종전의 ETL 과 DW 구축 및 관리만으로는 한계에 다다랐다. 이러한 이유로 많은 기업이 정형 데이터로 구성된 전통적인 데이터 외에 수많은 비정형 데이터(소셜 텍스트, 센서 데이터, 이미지, 동영상 등)를 실시간으로 수집, 정제, 통합, 분석해 활용하기 위해 데이터 레이크라는 새로운 개념의 데이터 관리 플랫폼을 도입하고 있다.

[데이터 창고\(DW\)는 잊어라...데이터 호수를 맞이하라 - Byline Network](#)

[데이터 레이크, 새로운 데이터웨어하우스가 된다 - ITWorld Korea](#)

[\[주말판\] 데이터 레이크, 앞으로 어떤 방향으로 흘러갈까? \(boannews.com\)](#)