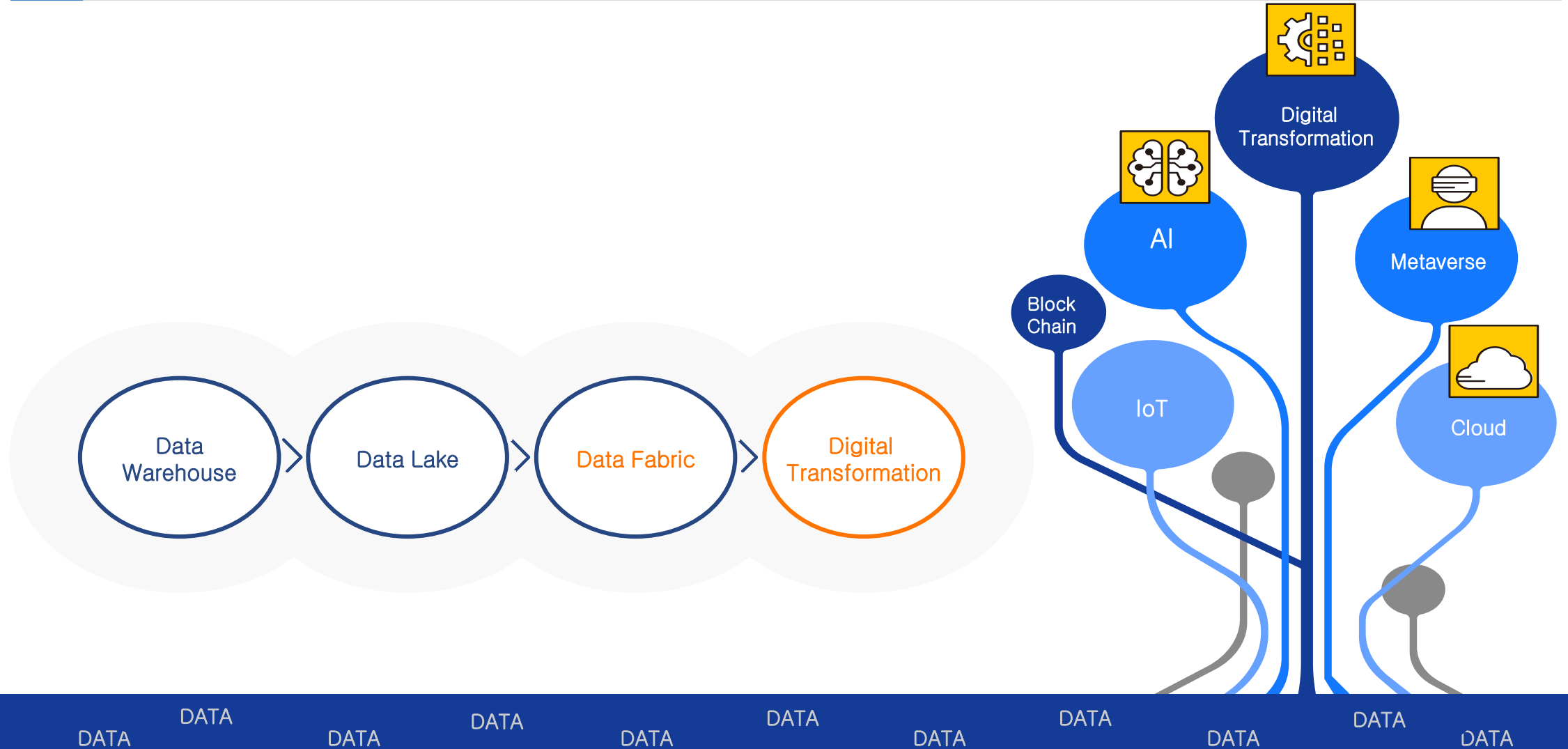


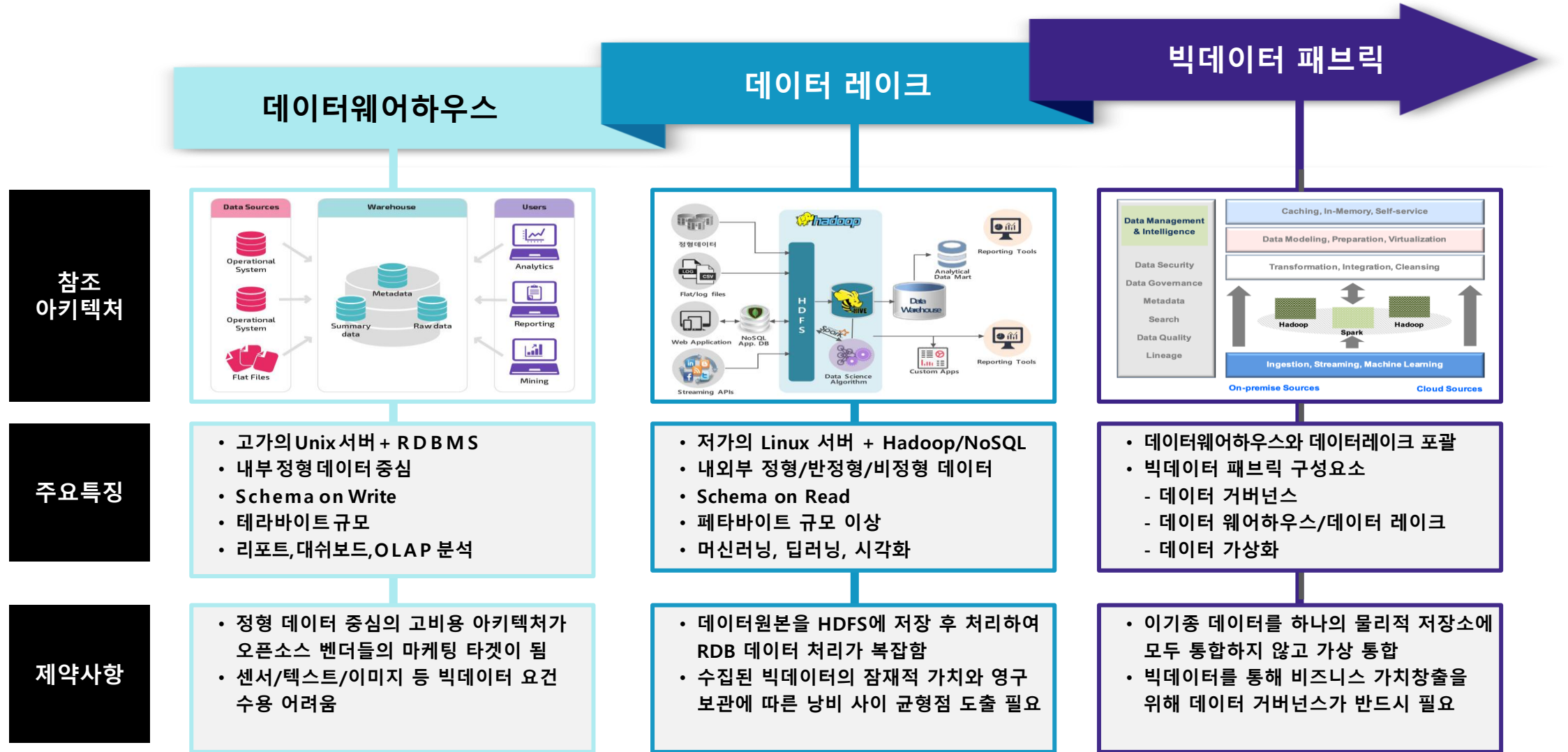


데이터의 의미

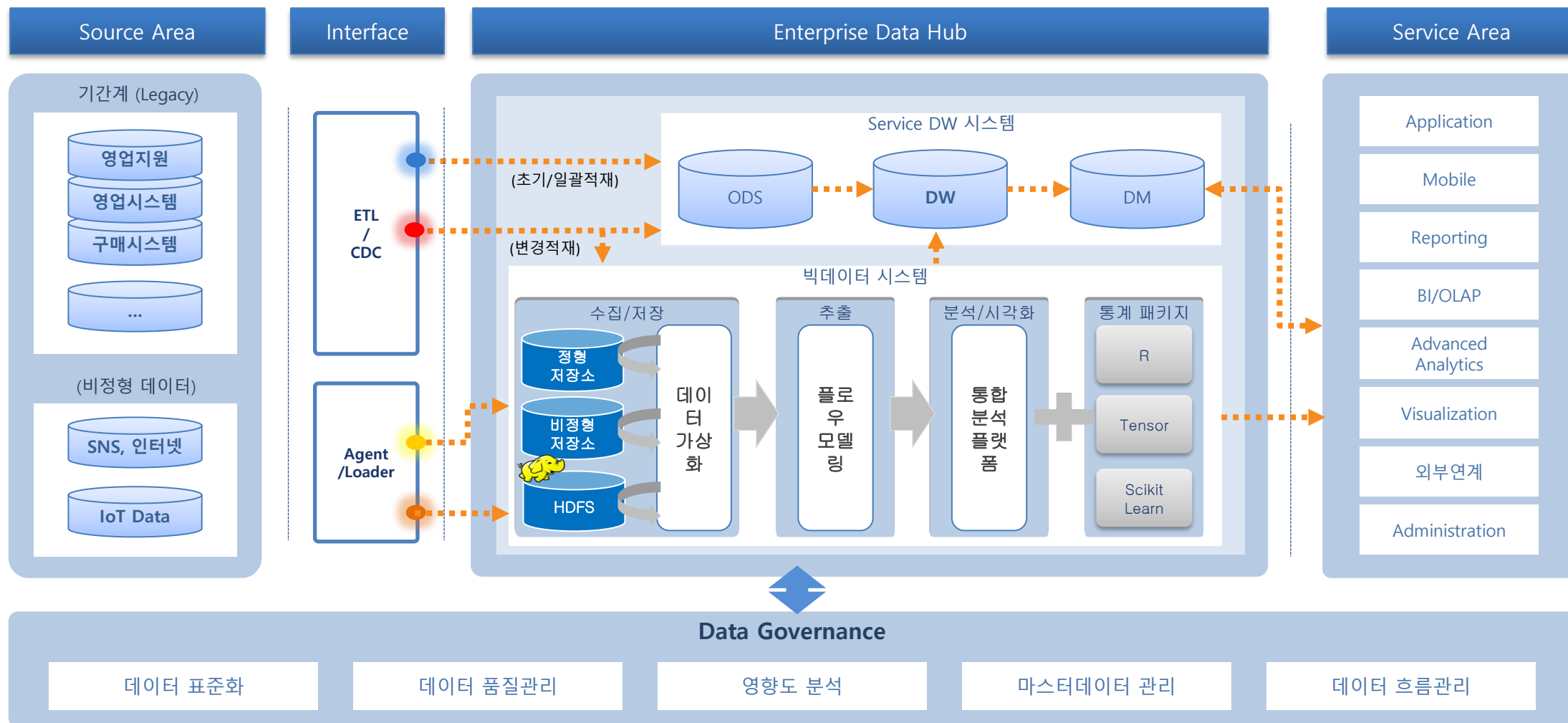
생각해 보기



생각해 보기

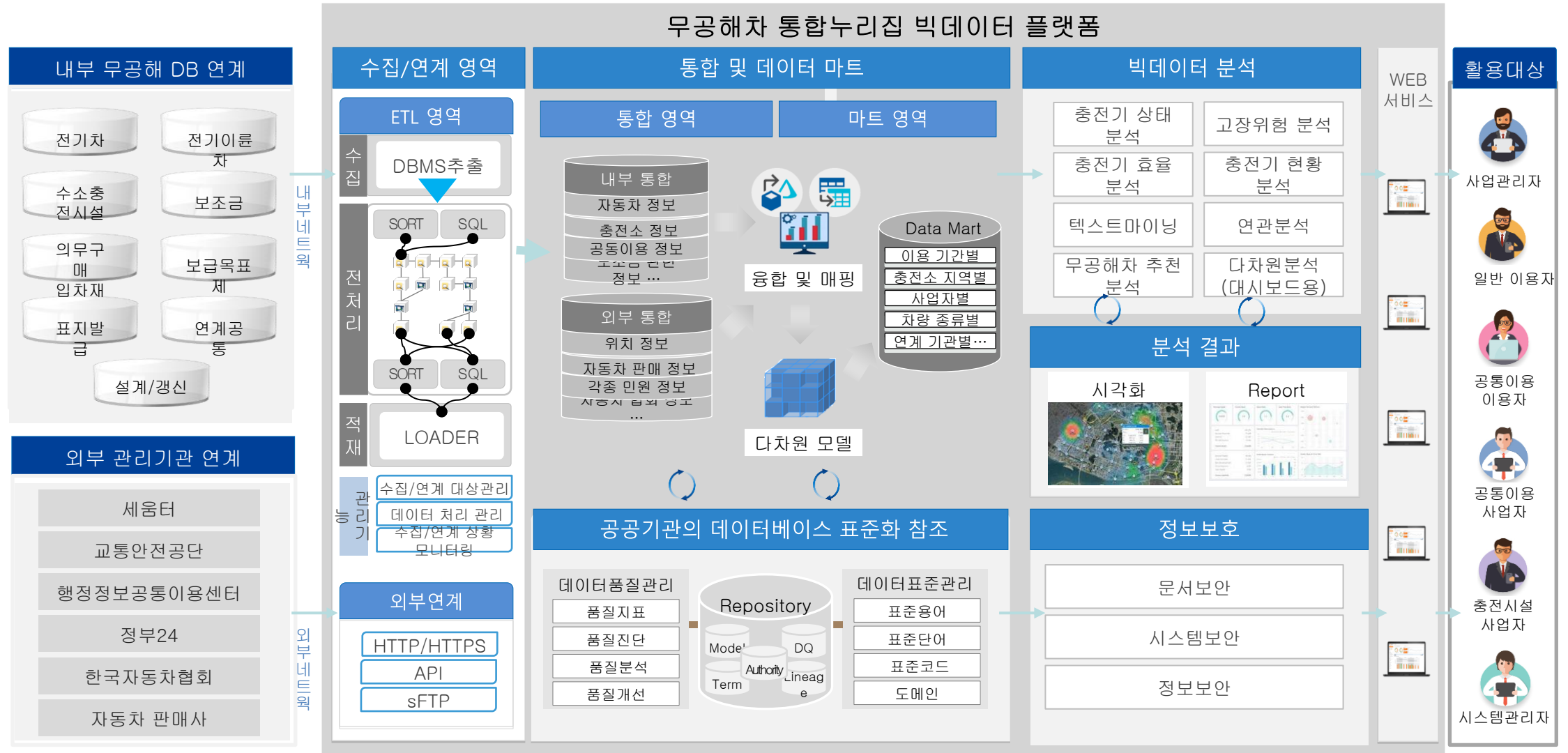


생각해 보기



생각해 보기

한국 환경 공단





DW 와 Lake 구축 시 고려 사항

DW 구축 시 고려 사항

전략적 고려 사항

- **경영 전략과의 정렬:** DW는 단순한 데이터 저장소가 아니라 전략적 의사결정을 지원하는 시스템이므로, 기업의 목표와 방향성과 일치해야 합니다.
- **경영진의 지원:** 경영진의 강력한 의지와 지속적인 관심이 프로젝트 성공의 핵심입니다.

DW 구축 시 고려 사항

기술적 고려 사항

- **데이터 품질 확보:** Legacy 시스템에서 추출한 데이터의 정제(Cleaning)와 검증이 필수입니다. 신뢰할 수 있는 데이터 없이는 분석도 무의미하죠.
- **ETL 성능:** Extract, Transform, Load 과정의 성능과 안정성은 DW의 기반입니다. 대용량 처리에 적합한 구조와 스케줄링이 필요합니다.
- **모델링 방식:** Star Schema, Snowflake Schema 등 분석 목적에 맞는 DB 모델링을 선택해야 합니다.
- **메타데이터 관리:** 데이터의 출처, 구조, 의미를 설명하는 메타데이터는 DW의 활용성과 유지보수에 큰 영향을 줍니다.

DW 구축 시 고려 사항

사용자 측면 고려 사항

- **사용자 인터페이스(UI):** 분석자와 의사결정자가 쉽게 접근하고 활용할 수 있도록 직관적인 UI와 시각화 도구(그래프, 차트 등)가 필요합니다.
- **교육 및 변화관리:** DW는 기존 운영 시스템과 다르기 때문에 사용자 교육과 변화관리 전략이 중요합니다.

DW 구축 시 고려 사항

운영 및 통합 고려 사항

- **기존 시스템과의 연계:** ERP, CRM 등 기존 시스템과의 데이터 흐름을 고려한 통합 설계가 필요합니다.
- **지속적 관리 및 튜닝:** DW는 구축 후에도 지속적인 성능 튜닝과 데이터 품질 관리가 필요합니다.
- **데이터 레이크 활용:** 정형/비정형 데이터가 많을 경우, DW와 함께 데이터 레이크를 고려하는 것도 좋은 전략입니다.

DW 구축 시 고려 사항

- 구축 방식 선택

방식	특징
Top-Down	전사적 관점에서 DW 전체를 먼저 구축. 시간·비용 많이 들지만 통합성 높음
Bottom-Up	부서별 Data Mart부터 시작해 통합. 빠른 구축 가능, 향후 통합 설계 필요
Hybrid	DW와 Data Mart 병행 구축. 리스크 분산, 유연한 자원 배분 가능

Data Lake 구축 시 고려 사항

데이터 레이크(Data Lake)를 구축할 때는 단순히 데이터를 저장하는 것 이상의 전략적, 기술적, 운영적 요소들을 고려해야 합니다. 잘못 설계하면 데이터 늪(Data Swamp)이 되어버릴 수도 있습니다.

전략적 고려 사항

- **비즈니스 목적 명확화** 단순 저장이 아닌 분석, 머신러닝, 실시간 처리 등 어떤 목적을 위한 데이터 레이크인지 정의해야 합니다.
- **데이터 거버넌스 체계 수립** 데이터의 소유권, 접근 권한, 품질 기준 등을 명확히 해야 데이터 혼란을 막을 수 있어요.

Data Lake 구축 시 고려 사항

기술적 고려 사항

- **데이터 유형 다양성** 정형(SQL), 반정형(JSON, XML), 비정형(이미지, 영상, 로그 등) 데이터를 모두 수용할 수 있는 구조가 필요합니다.
- **스키마 온 리드 방식** 저장 시 스키마를 강제하지 않고, 분석 시점에 스키마를 적용하는 유연한 구조가 핵심입니다.
- **스토리지 선택** 오브젝트 스토리지(S3 등) vs. 분산 파일 시스템(HDFS 등) — 비용, 성능, 확장성에 따라 선택해야 합니다.
- **아키텍처 설계** 단일 중앙화 vs. 분산형 레이크 — 조직 규모, 규제, 네트워크 환경에 따라 결정해야 합니다.

Data Lake 구축 시 고려 사항

보안 및 관리 고려 사항

- **데이터 보안 및 개인정보 보호** 민감 정보 암호화, 접근 제어, 감사 로그 등 보안 정책을 철저히 수립해야 합니다.
- **데이터 품질 관리** 중복, 오류, 누락 데이터를 방지하기 위한 정제 및 검증 프로세스가 필요합니다.
- **모니터링 및 메타데이터 관리** 저장된 데이터의 출처, 구조, 사용 이력 등을 추적할 수 있는 메타데이터 시스템이 중요합니다.

Data Lake 구축 시 고려 사항

운영 및 확장 고려 사항

- **ETL/ELT 전략** 실시간 스트리밍 수집(Kafka 등) vs. 배치 수집 — 워크로드에 따라 적절한 방식 선택.
- **확장성과 고가용성** 클라우드 기반으로 구축 시 자동 확장(Auto Scaling), 장애 복구(Disaster Recovery) 설계가 필수입니다.
- **도구 및 분석 환경 연계** Spark, Presto, Hive, ML 플랫폼 등과의 연동을 고려해 분석 환경을 구축해야 합니다.

Data Lake 구축 시 고려 사항

구축 시 체크리스트

항목	질문 예시
아키텍처 설계	중앙 집중형으로 갈 것인가, 분산형으로 갈 것인가?
데이터 수집 방식	실시간 스트리밍이 필요한가, 배치 처리로 충분한가?
스토리지 선택	오브젝트 스토리지 vs. 분산 파일 시스템 중 어떤 것이 적합한가?
보안 및 규제 대응	GDPR, 개인정보보호법 등 규제에 어떻게 대응할 것인가?
분석 도구 연계	어떤 BI/ML 도구와 연동할 것인가?



ODS , DW 개념

학습목표

1 ODS를 설명한다.

2 DW(EDW)를 설명한다.

2 고객 데이터 통합 방법을 학습한다.

1_ ODS의 의미

1) ODS 데이터

■ ODS 데이터

- DW 구축에서 Source 데이터에 일정한 가공 과정을 거쳐 작성
- 조직의 단기적 의사결정을 지원할 수 있음
- DW를 구축하기 위한 중간적 역할을 담당

1_ ODS의 의미

2) ODS 정의 및 특징

■ ODS 정의

- 주제중심적(Subject-oriented)이고,
- 통합적(Integrated)이며,
- 최근의 ([Current](#) valued), 휘발성(Volatile), 상세(Detailed) 데이터의 [집합](#)

■ ODS 특징

- 일상적(day-to-day, up-to-the-second)인 [의사결정](#)을 지원
- [데이터웨어하우스](#)로의 이동 통로(Migration Path) 제공

2_ DW(EDW)의 의미

1) DW의 의미

■ DW 의미

- 기업의 대단위 데이터를 주제별로 통합 축적하여 별도의 장소에 저장해 놓은 것
- 단순한 데이터의 저장고가 아니라, 관계형 DB를 근간으로 많은 데이터를 다차원 분석하여 의사결정에 도움을 줌
- 조직내에 집적된 각종 데이터를 다차원적으로 분석함으로써
- 서로 다른 정보(생산,구매,주문,영업 등)들의 연관성을 신속하게 찾아내어,
- 의사결정을 지원하는 도구

2_ DW(EDW)의 의미

2) DW 용어유래

■ DW 용어 유래

- 80년대 중반 IBM이 자사 하드웨어에 Information Warehouse라는 용어 사용
- 기존의 OLTP 시스템은 다년간 추세분석과 같은 방대한 과거 데이터 유지 및 비정형 질의에 대한 처리에 한계 봉착
- 1992년 Inmon이 처음으로 개념 및 정의 제시
 - 기업의 의사결정 과정을 지원하기 위해 . 주제 중심적(Subject-oriented)이고, . 통합적(Integrated)이며, 시간성(Time-variant)을 갖고, 비휘발성(Non-volatile)인 자료의 저장소

OLTP : on-line transaction processing

2_ DW(EDW)의 의미

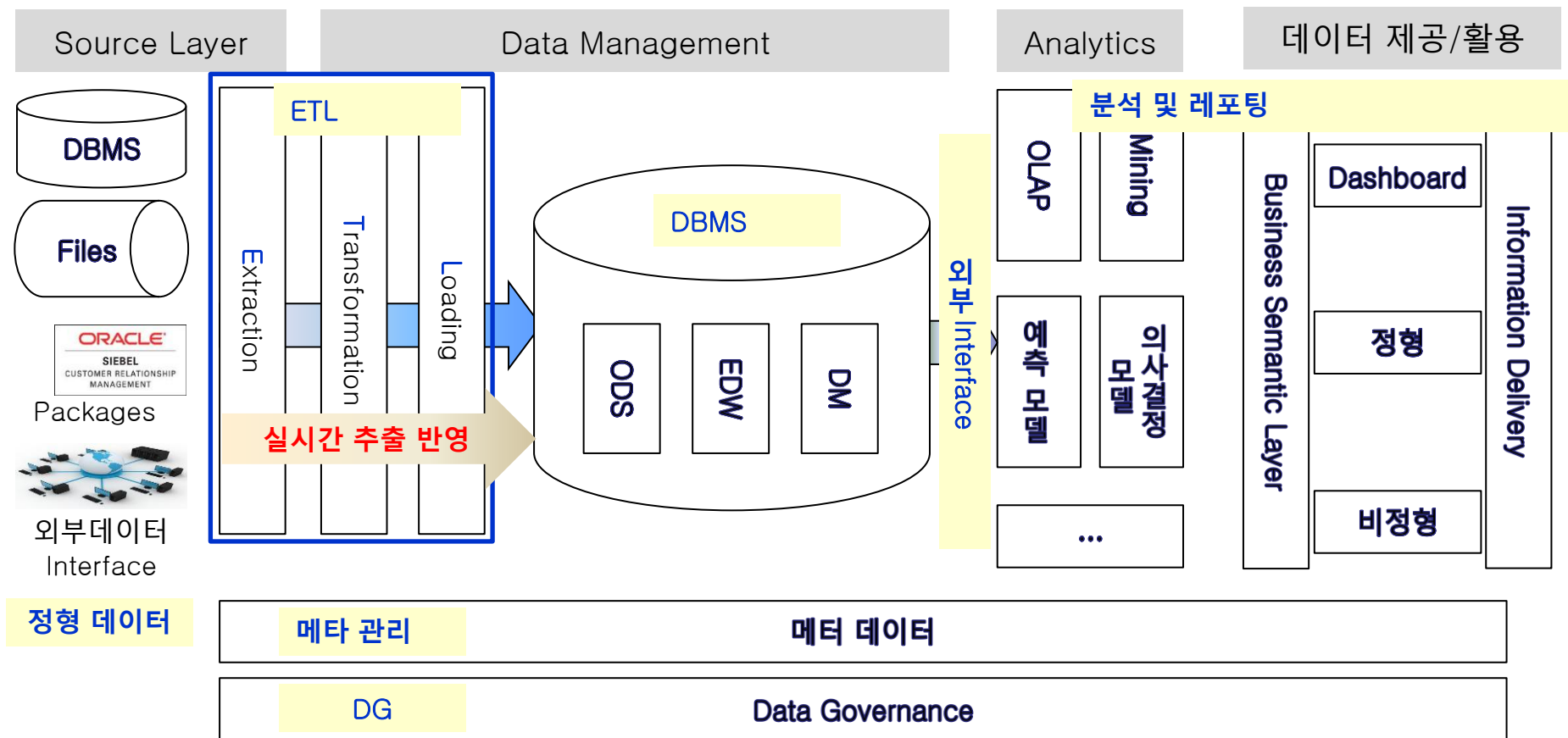
3) DW 특징

■ DW 특징

- 주제지향성 (Subject-oriented) : 고객, 제품 등과 같이 주제 중심으로 구성 (주제별 분류)
- 통합성 (Integrated) : 일관성 있는 데이터의 정의, 레이아웃, 관계성, 키 구조 등 (하나의 의미로 통합)
- 시계열성 (Time-varient) : DW 내의 데이터는 스냅샷 형태의 데이터로 묵시적, 명시적으로 시간 항목을 가지며, 장기간에 걸쳐 존재하게 함 (연속성 있는 이력 데이터 관리)
- 비휘발성 (Non-volatile) : 일단 읽기전용 형태(스냅샷 등)로 만들어지면 갱신이 이루어지지 않음

2_ DW(EDW)의 의미

4) DW Flow



2_ DW(EDW)의 의미

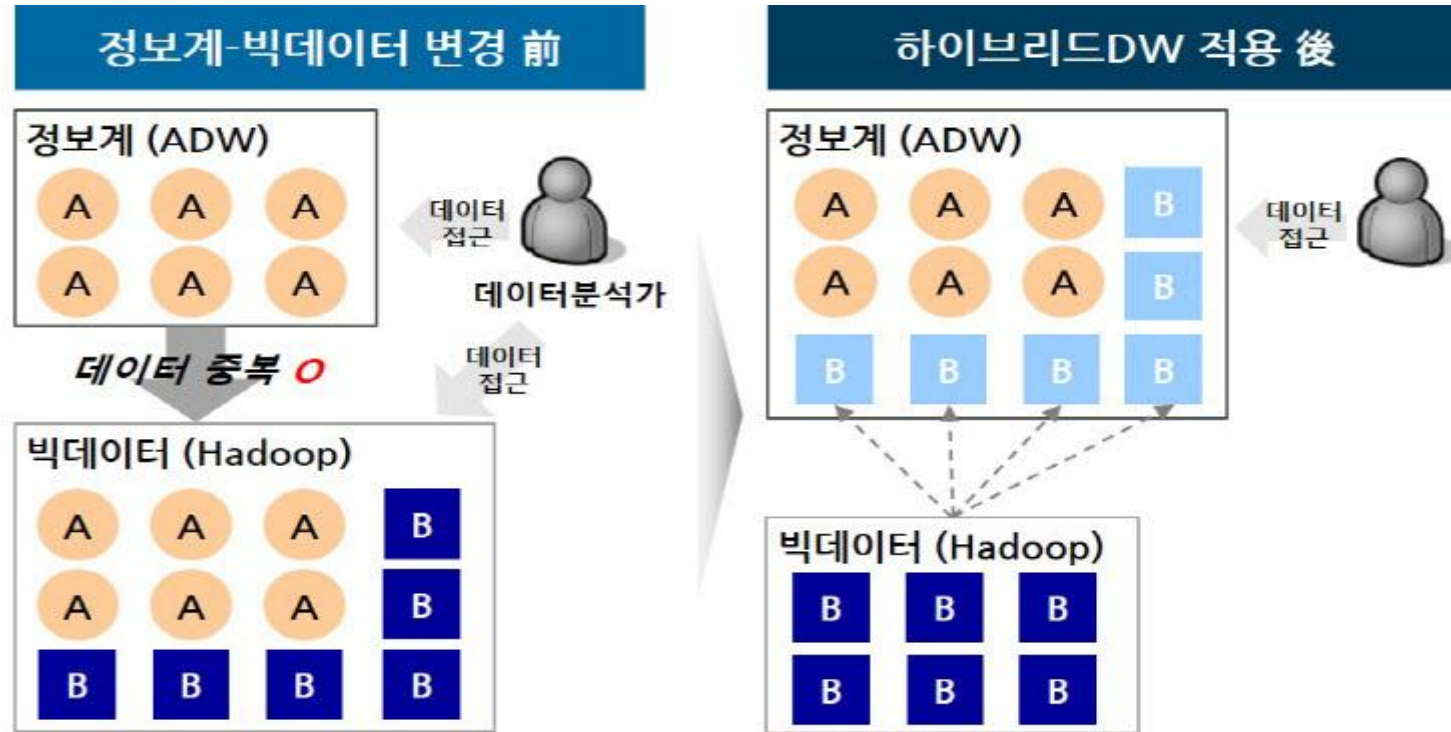
5) 발전적 DW

- 우리은행, 미래지향적 빅데이터 시스템 구축 (1/2)
 - 국내은행 최초로 빅데이터 플랫폼과 EDW(Enterprise Data Warehouse)를 결합하고 하이브리드 DW(Data Warehouse) 아키텍처 구축 프로젝트를 성공적으로 완료
 - 하둡(Hadoop) 분산정보저장시스템과 기존 EDW를 통합해 다양하고 많은 정보를 저장할 수 있는 데이터레이크(Data Lake)를 새롭게 구축했으며,
 - 이를 통해 데이터 중복 적재에 따른 자원 낭비를 해소하고 데이터 분석, 설계, 서비스 구현 등에서 50% 이상 속도를 높임

2_ DW(EDW)의 의미

5) 발전적 DW

- 우리는, 미래지향적 빅데이터 시스템 구축 (2/2)



[그림] 우리은행, 하이브리드 DW(Data Warehouse) 구조도



메타 데이터와 데이터 표준화

1 메타데이터 의미

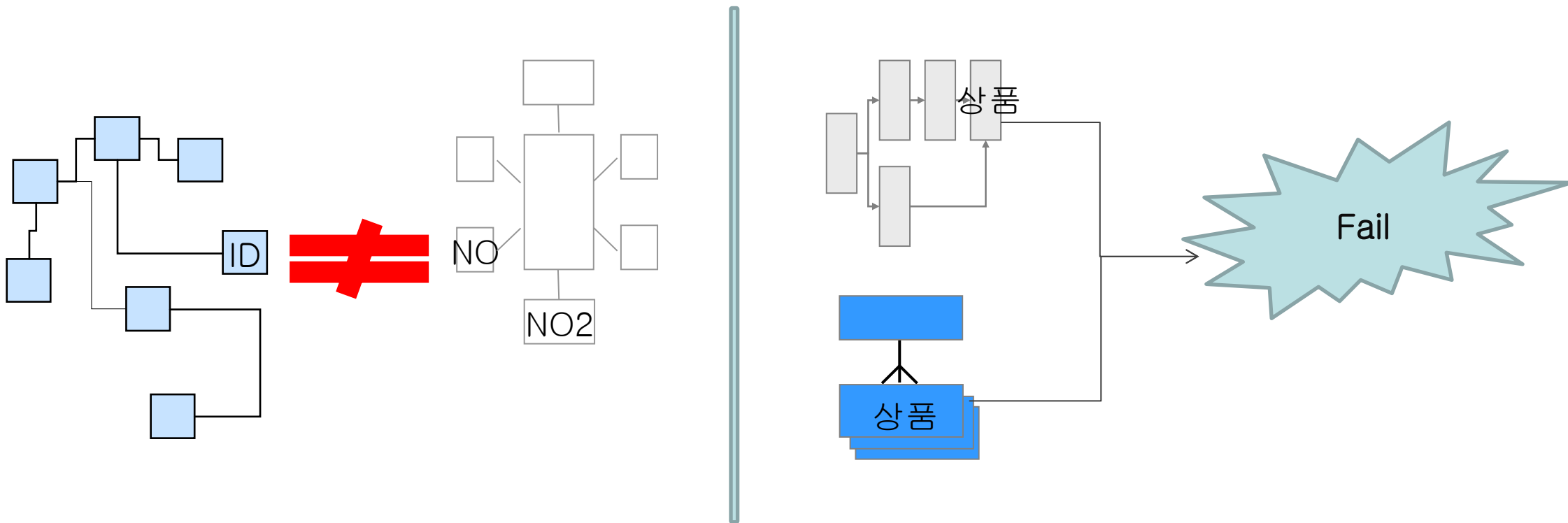
2 데이터 표준화 정의

3 데이터 표준화 절차

생각해 보기

- 왜 데이터 처리 시 결과값이 맞지 않을까?

Select * From cust Where cust.ID = tb_cust.no;



학습목차

1

메타데이터 의미

학습목표

- 1 데이터 관리 체계를 설명 한다.
- 2 메타데이터 관리 정의를 설명 한다.
- 3 메타데이터 관리 시스템 구축을 설명 한다.

1_데이터 관리체계 정의

1) 메타데이터 의미 (1/2)

■ 데이터에 대한 데이터

- 동영상, 소리, 문서 등과 같이 실제로 존재하거나 사용할 수 있는 데이터는 아니지만, 이러한 실제 데이터와 직접적 또는 간접적으로 연관된 정보를 제공해주는 데이터를 말한다.
- 일반적으로 가장 많이 사용되고 있는 메타데이터의 정의는 “데이터에 대한 데이터”이며 기능적인 면을 강조하였을 때는 “데이터에 대한 구조화된 데이터”로 정의된다.
 - 쉬운 예로, 영화를 보려고 DVD 타이틀을 구매하였을 때 포장에 붙어있는 제목, 상영시간, 등급, 제작사, 감독, 줄거리 등에 관한 정보를 메타데이터라고 할 수 있다.

1_데이터 관리체계 정의

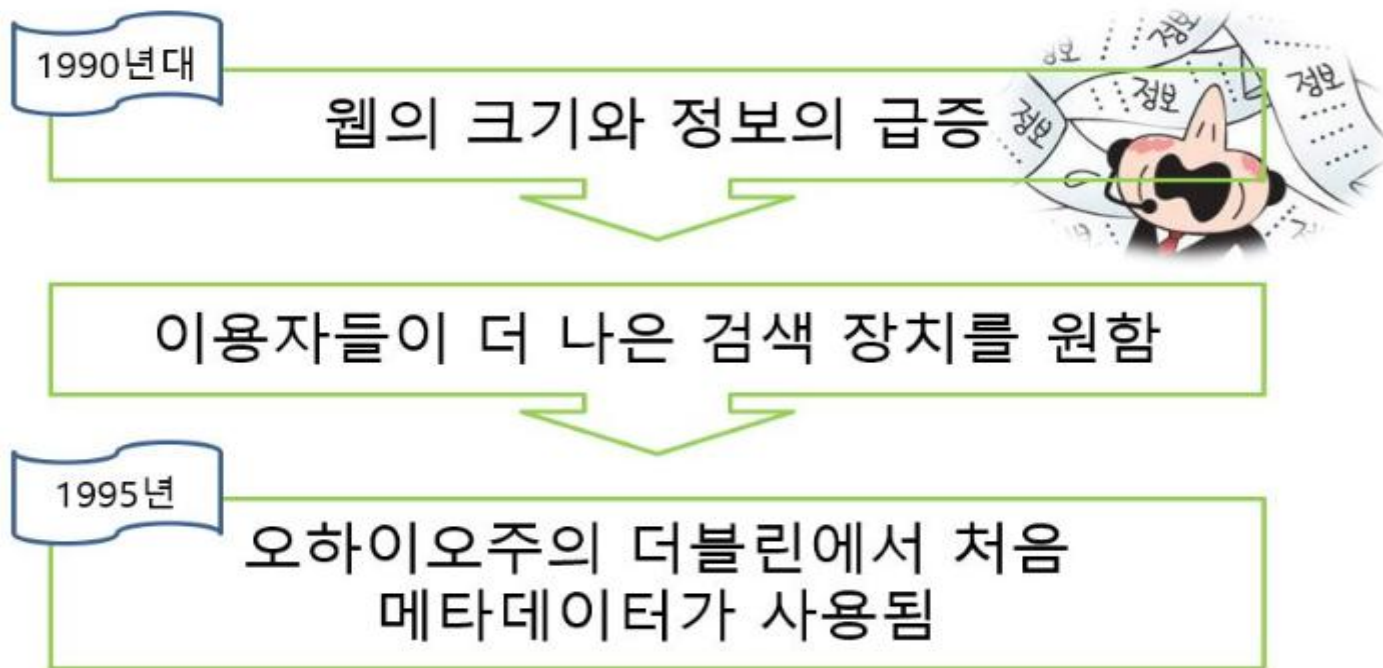
1) 메타데이터 의미 (2/2)

- 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터
 - 속성정보라고도 한다.
 - 대량의 정보 가운데에서 찾고 있는 정보를 효율적으로 찾아내서 이용하기 위해 일정한 규칙에 따라 콘텐츠에 대하여 부여되는 데이터이다.
 - 여기에는 콘텐츠의 위치와 내용, 작성자에 관한 정보, 권리 조건, 이용 조건, 이용 내력 등이 기록되어 있다.
 - 컴퓨터에서는 보통 메타데이터를 데이터를 표현하기 위한 목적과 데이터를 빨리 찾기 위한 목적으로 사용하고 있다.

1_데이터 관리체계 정의

2) 메타데이터 등장

메타데이터의 등장



1_데이터 관리체계 정의

3) 데이터, 메타데이터



메타데이터	데이터
고객번호	20190800001
고객명	홍길동
성별	남자
가입일자	2019.8.30
고객등급	최우수등급
직업	회사원
주민등록번호	72010-11177***
휴대전화번호	010-4545-2828

관리할 데이터를 정의

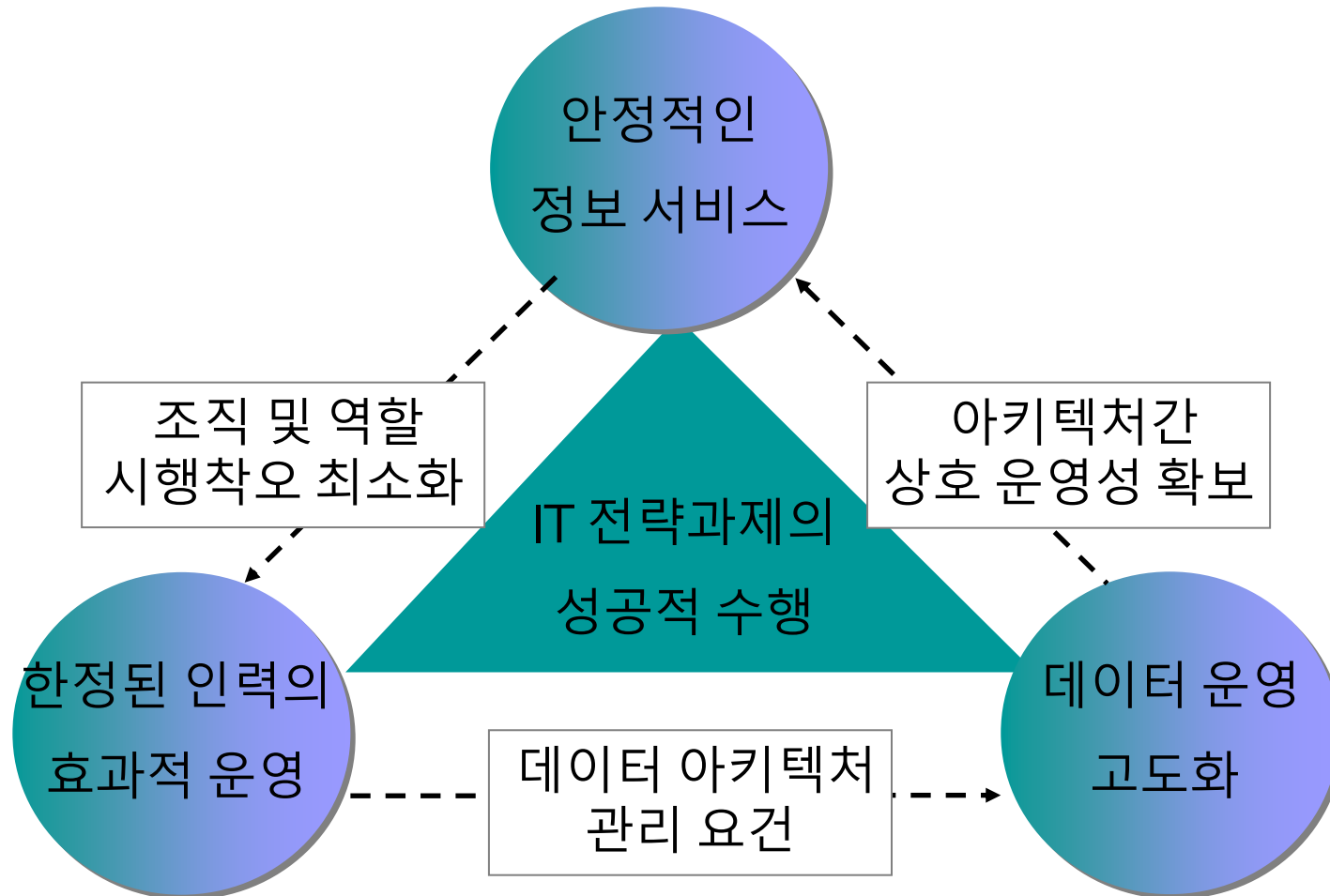
Karen Coyle

어떤 목적을 가지고 만들어진 데이터
(Constructed data with a purpose)

1_데이터 관리체계 정의

4) 데이터 관리 체계

■ 데이터 관리 체계 개요



1_데이터 관리 체계 정의

5) 데이터아키텍처 관리체계

관리 항목

데이터요구사항, 데이터 표준, 데이터 모델, 데이터베이스 운영, 데이터 발생규칙, 데이터 품질 등의 데이터 아키텍처의 목적 달성 여부를 평가하기 위해 필요한 관리 항목들을 정의 합니다.

관리 프로세스

데이터 아키텍처의 운영관리 목적을 달성하기 위한 통제절차에 대한 프로세스를 정의 합니다.

관리 조직

관리 조직은 전사의 데이터 아키텍처를 주관하는 상시 관리 조직과, 프로젝트의 데이터 아키텍처를 관리하는 비상시 조직으로 구분 합니다.

1_ 데이터 관리 체계 정의

6) 데이터 관리 체계의 수립 목표

■ 수립 목표

- 데이터 요구사항관리, 데이터 표준 및 품질관리, 데이터 모델 및 발생규칙, DB 운영 및 성능관리 영역에 대해 아키텍처 관리 항목과 우선순위를 정의
- 데이터 아키텍처 관리를 통한 전사 데이터의 활용을 강화

1_ 데이터 관리 체계 정의

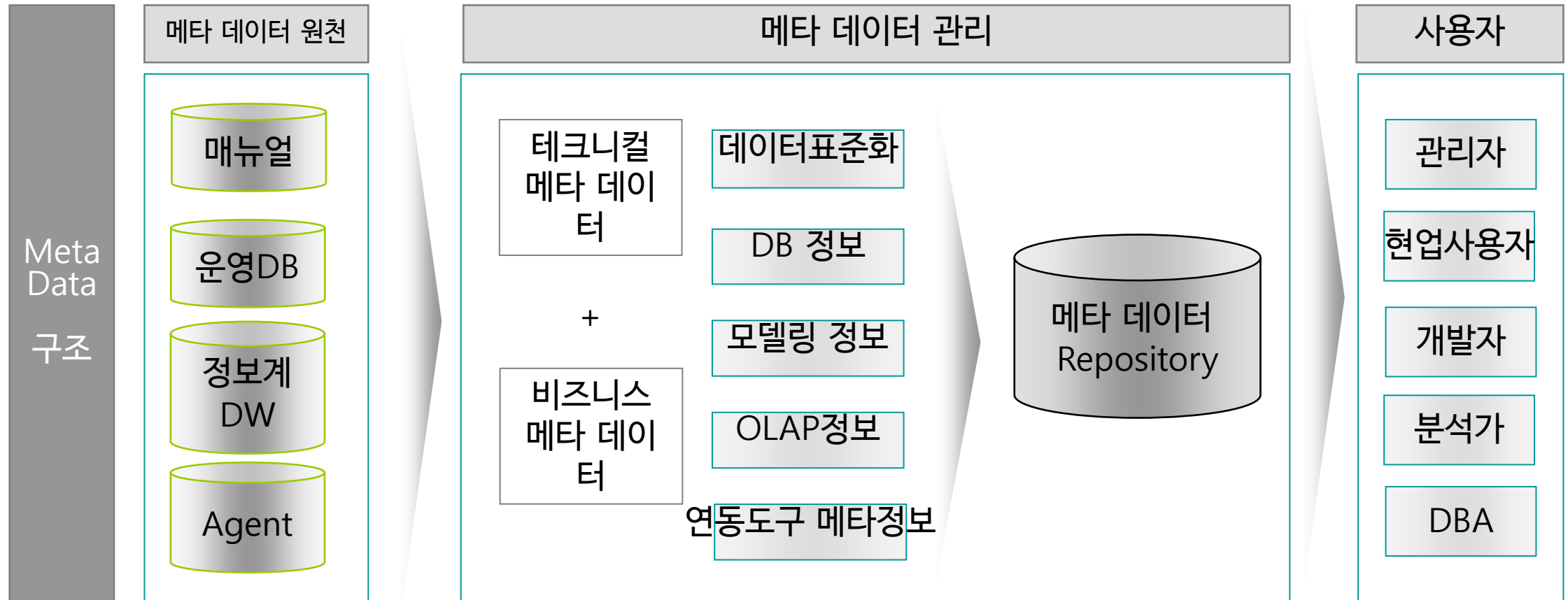
7) 데이터관리 체계 수행방법

- 데이터 관리 체계 수행 방법
 - 현행 데이터 관리 자료 취합
 - Focus Group 인터뷰
 - To-Be 데이터 관리 방향 수립
 - 데이터 관리 체계 수립
 - 데이터 저장소 배치도

2_ 메타데이터 관리 정의

1) 메타데이터 관리 정의

■ 메타데이터 정의



2_ 메타데이터 관리 정의

2) 메타데이터 관리 지침수립

■ 메타데이터 관리 지침수립

데이터 관리현황 파악

- 데이터 관리 현황파악 및 개선사항 도출

메타데이터 정책/지침 수립

- 데이터 관리 방향 및 원칙 정의

메타데이터 관리 기능

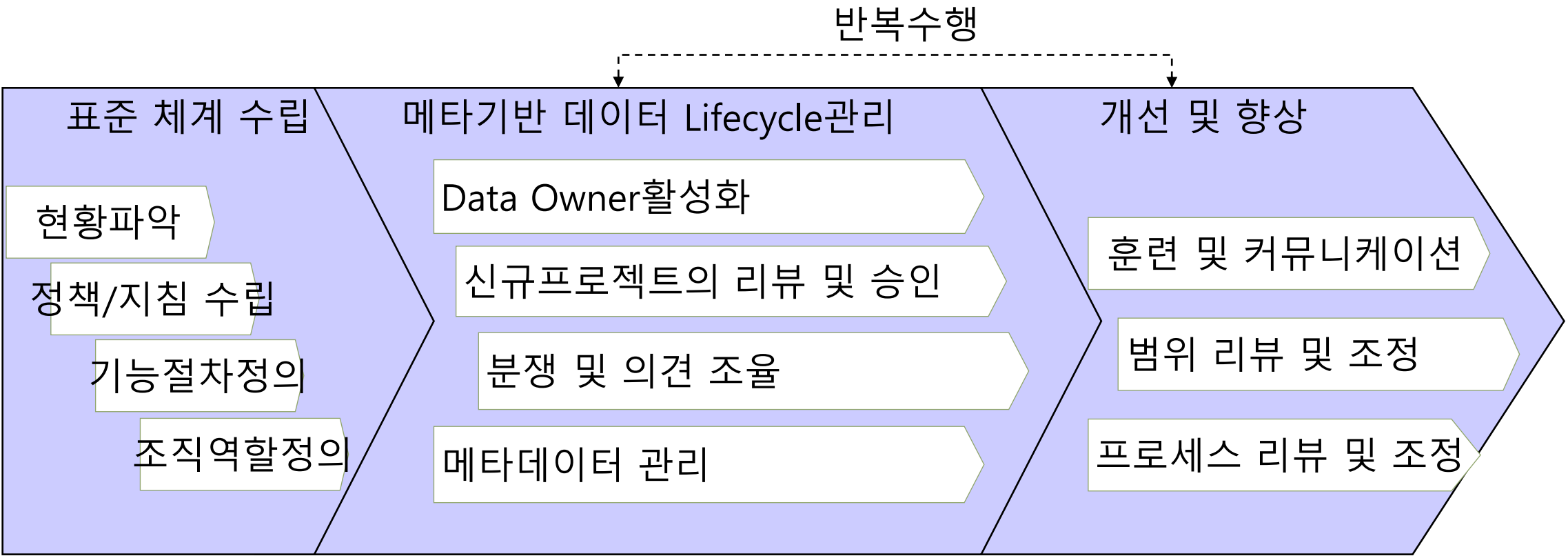
- 데이터 표준관리
- 데이터 모델관리
- 데이터 보안관리
- 메타데이터 관리
- 데이터 품질관리

메타데이터 관리 조직

- ◆ 데이터 관리조직 및 역할정의

2_ 메타데이터 관리 정의

3) 메타데이터 관리 프로세스



2_ 메타데이터 관리 정의

4) 메타데이터 관리 시스템 구축효과

■ 구축 효과

- 정량적 효과
 - 정보시스템 개발 생산성 증대
 - 비생산적인 작업 감소
 - 데이터 중복 감소
 - 중복 프로세스 감소
- 정성적 효과
 - 정보시스템 전반적 관리기능의 개선
 - 표준 준수도의 향상
 - 작업결과에 대한 타 팀에의 전달효과 증대
 - 프로젝트 실패 가능성 줄임

3_메타데이터 관리 시스템 구축

1) 구축 방법

- 1단계

- 데이터 표준화/품질 정의

- 2단계

- 데이터품질 강화

- 3단계

- 통합 데이터관리

3_메타데이터 관리 시스템 구축

2) 프로세스 정의

- 사용자 편리성
 - 포탈을 통한 통합관리
- 표준항목 관리
 - 표준 단어 , 표준 용어 ,도메인(코드)
- 모델 관리
 - 모델링 도구 연동 , 모델 분류구조관리 , 테이블 별 DDL 및 이력관리
- DB 관리
 - 데이터베이스 오브젝트관리 , 물리모델과 DB간의 정합성 관리 , DB의 변경이력관리

3_메타데이터 관리 시스템 구축

3) 메타데이터 추출

- 메타데이터 추출



META DATA POPULATOR & Application Parser

통합 메타 레파지토리



학습목차

2 데이터 표준화 정의

학습목표

1 데이터 표준화가 무엇인지를 설명 한다.

2 데이터 용어 표준화를 설명 한다.

3 데이터 사전 구성을 설명 한다.

1. 데이터 체계 수립

1) 데이터 표준화 개요

■ 데이터 표준 정의

- 해당 데이터를 가장 잘 이해할 수 있도록 업무적인 관점에서 데이터에 대한 상세한 설명

■ 데이터 표준 명칭

- 데이터를 유일하게 식별하는 이름으로 명칭을 부여. 명칭만으로 의미를 파악할 수 있도록 구체적 이며 명확하게 정의

■ 데이터 표준 형식 및 규칙

- 데이터 표현 형태의 정의를 통해 데이터 입력 오류와 통제 위험을 최소화하는 역할.
- 발생 가능한 데이터 값을 사전에 정의함으로써 데이터 입력 오류와 통제 위험을 최소화하는 역할

1. 데이터 체계 수립

2) 데이터 표준화의 특징

■ 데이터 관리의 시작점

- 누구나 같은 의미로 이해할 수 있도록 데이터 정의 방법의 표준을 정의하는 것.

■ 데이터 표준 근거 마련

- 전사 합의에 의해 정의된 표준은 가급적 예외를 두지 말고 실행 될 수 있도록 법령, 제도, 지침을 근거로 마련.

■ 메타데이터 관리시스템과 연계

- 표준화 후 결과물(표준사전, 관리프로세스, 역할정의, 대상시스템 정보)은 메타관리시스템과 밀접하므로 컨설팅 중에 도입되는 메타관리시스템의 기능 및 특성을 잘 이해하여 연계

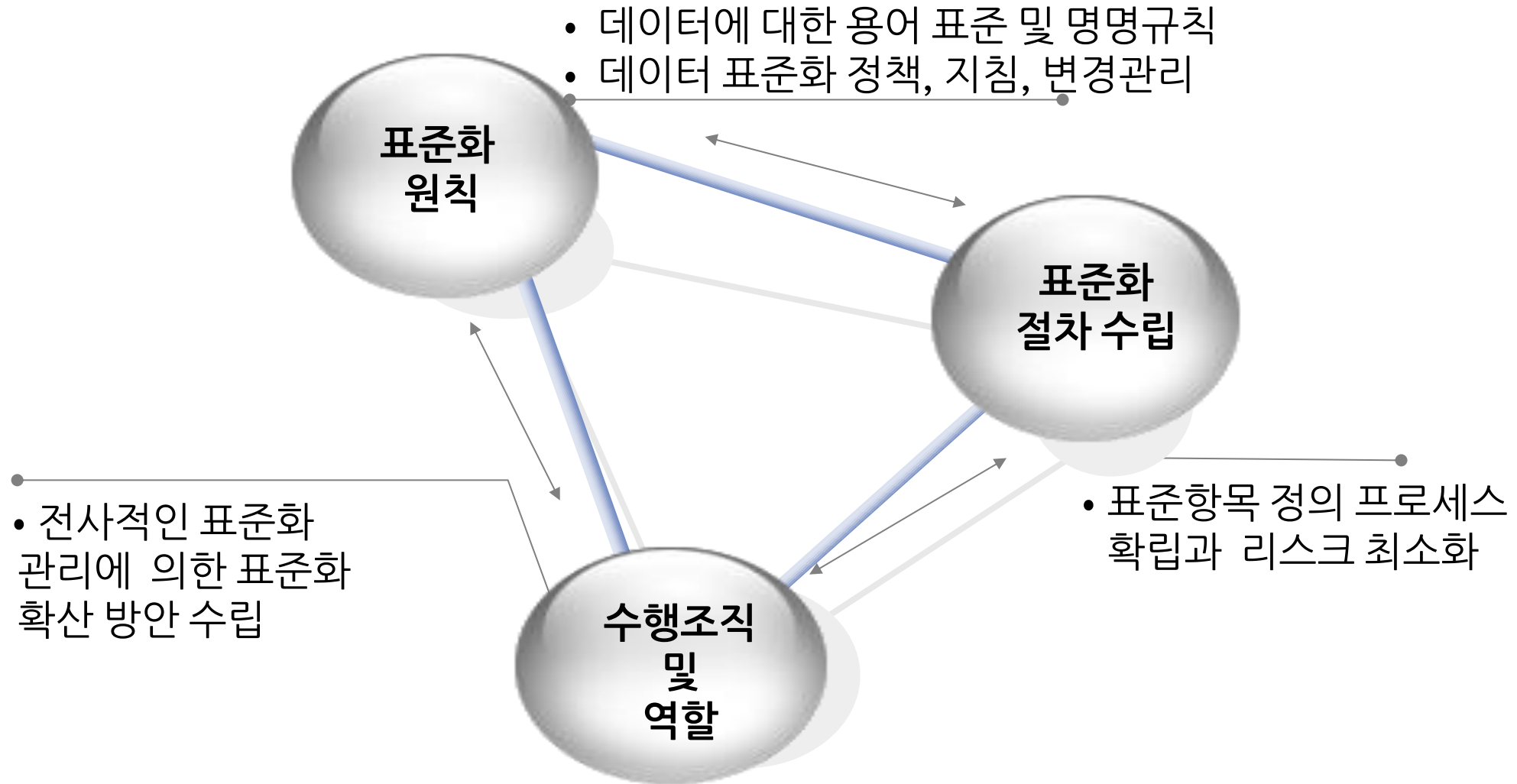
1. 데이터 체계 수립

3) 데이터 표준화의 목표

- 전사 차원의 “체계적인 데이터 표준 관리체계 ” 수립
 - 전사 데이터 아키텍처 기반에 따라 데이터 표준화 역할 및 프로세스 체계 정의함으로써
 - 일관성 있는 데이터 표준
 - 데이터 통합 모델 관리
 - 모델-DB간 정합성 확보
 - 구조적 데이터 품질 관리를 확보를 목표

1. 데이터 체계 수립

4) 데이터 표준화 지침 작성 기준 (1/2)



1. 데이터 체계 수립

4) 데이터 표준화 지침 작성 기준 (2/2)

- 표준화 지침 및 사전 작성 원칙
 - 표준화 원칙
 - 표준화 체계, 단어, 도메인, 용어, 코드
 - 표준화 절차 수립
 - 프로세스
 - 수행 조직 및 역할
 - R&R

2. 데이터 용어 표준화

1) 표준화 개념

■ 개념

- 데이터 플랫폼에서 사용하고 관리하는 데이터에 대해서
- 누구나 같은 의미로 이해하고 같은 방법으로 사용할 수 있는 원칙(기준)을 정하는 것을 의미.
- 데이터 플랫폼에서 보유하고 있는 자료사전을 데이터표준 중심으로 통합하여 사용자에게 일관되고 정확한 데이터의 의미를 제공하는 일련의 과정

2. 데이터 용어 표준화

2) 표준화 대상

- 표준단어

- 의미를 가지는 최소 단위이다.

- 표준용어

- 데이터 플랫폼에서 사용하는 각종 용어로서 독립적이고 구체적인 의미를 가지고 있으며 표준단어들의 조합 을 통해 만들어진다 .

- 표준도메인

- 표준용어가 가지고 있는 데이터의 성질을 그룹핑하여 데이터의 일관성을 유지하는 역할을 한다.

- 통합도메인

- 전체 시스템에 통합하여 관리되는 코드를 말한다.

2. 데이터 용어 표준화

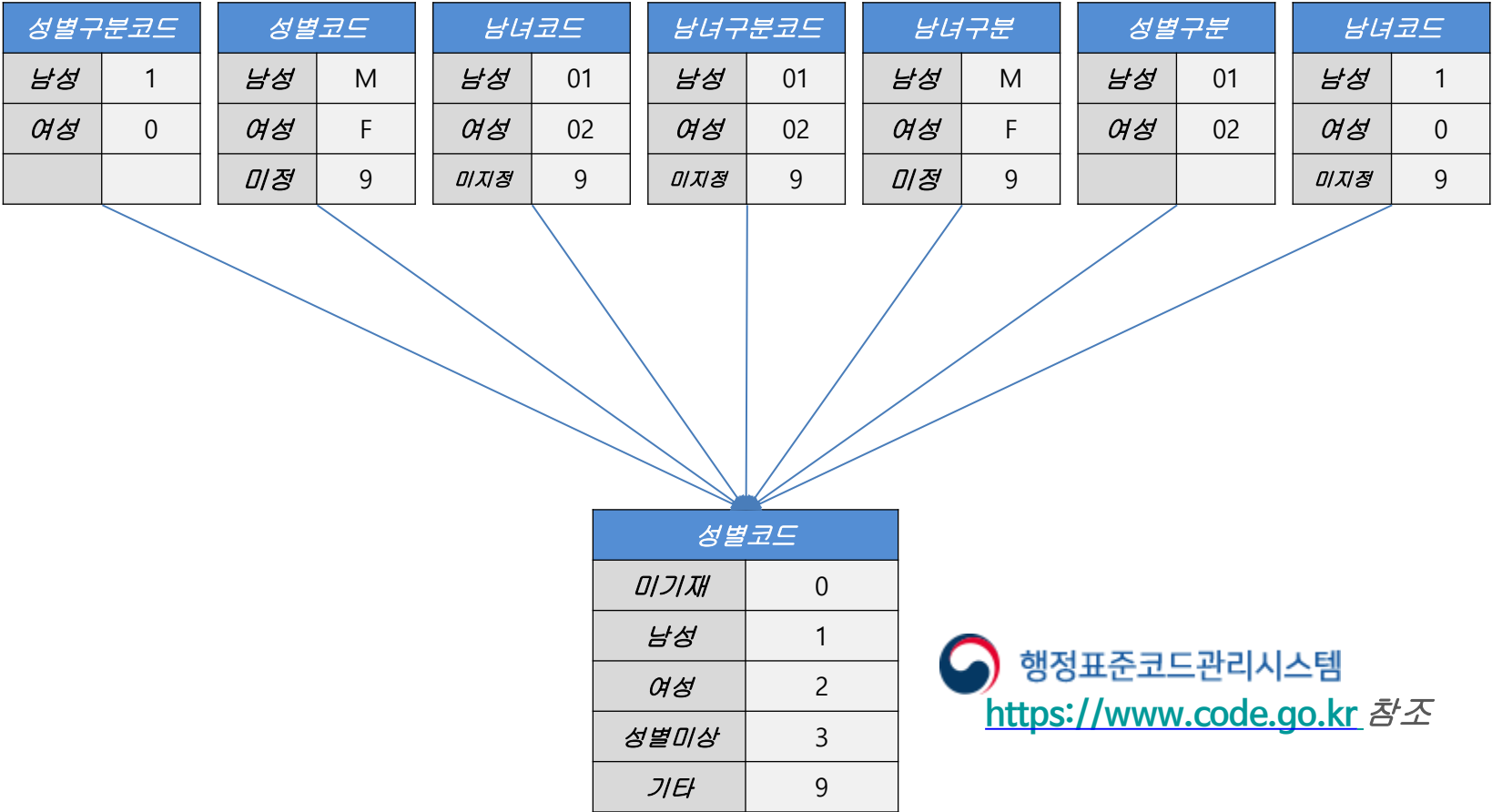
3) 메타데이터에 대한 표준 정의

	System A	System B	System C	System D	System E
고객명	고객명	고객성명	고객이름	고객명	고객성명
주민등록번호	주민등록번호	주민번호	주민번호	주민등록번호	고객식별번호
고객등급코드	고객등급	고객등급구분코드	고객등급코드	고객분류코드	고객등급
고객가입일자	가입일자	고객가입일자	등록일자	입력일자	고객가입일자
휴대전화	휴대전화번호	휴대폰번호	핸드폰번호	휴대전화	휴대전화번호
사원번호	사원번호	직원번호	조직원번호	직원번호	사원번호

표준용어
고객명
주민등록번호
고객등급코드
고객가입일자
휴대전화
직업
직원번호

2. 데이터 용어 표준화

4) 코드 표준 정의



 행정표준코드관리시스템
<https://www.code.go.kr> 참조

2. 데이터 용어 표준화

5) 표준화의 범위 (1/2)

■ 표준화 범위

- 데이터표준화 관리 대상은 전사에 걸쳐 수집된 모든 데이터에 대한 정확한 정의와 데이터사전이다.
- 최근에는 빅데이터 플랫폼 구축에 따라 오브젝트는 각 센터 및 플랫폼 업체에서 제공하는 데이터가 작성되는 것에서부터 사용자의 데이터 활용 때까지 일련의 과정이 포함되며,
- 논리 설계로써 모델링의 명명규칙(Naming Rule)과, 물리 설계에 해당하는 물리적인 오브젝트 의 명명규칙(Naming Rule)을 포함된다.

2. 데이터 용어 표준화

5) 표준화의 범위 (2/2)

- 논리 데이터 모델

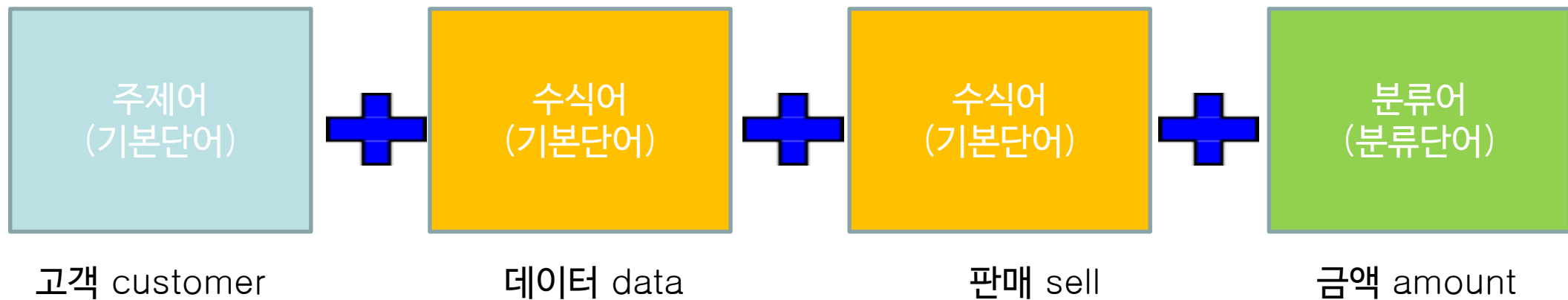
- 주제영역 (Subject Area)
- 엔티티 (Entity)
- 속성 (Attribute)

- 물리 데이터 모델

- 데이터베이스 (Database)
- 테이블스페이스 (Tablespace)
- 테이블 (Table) , 컬럼 (Column)
- 인덱스 (Index) , 뷰 (View)

3. 데이터 사전 구성

1) 표준 용어 구성



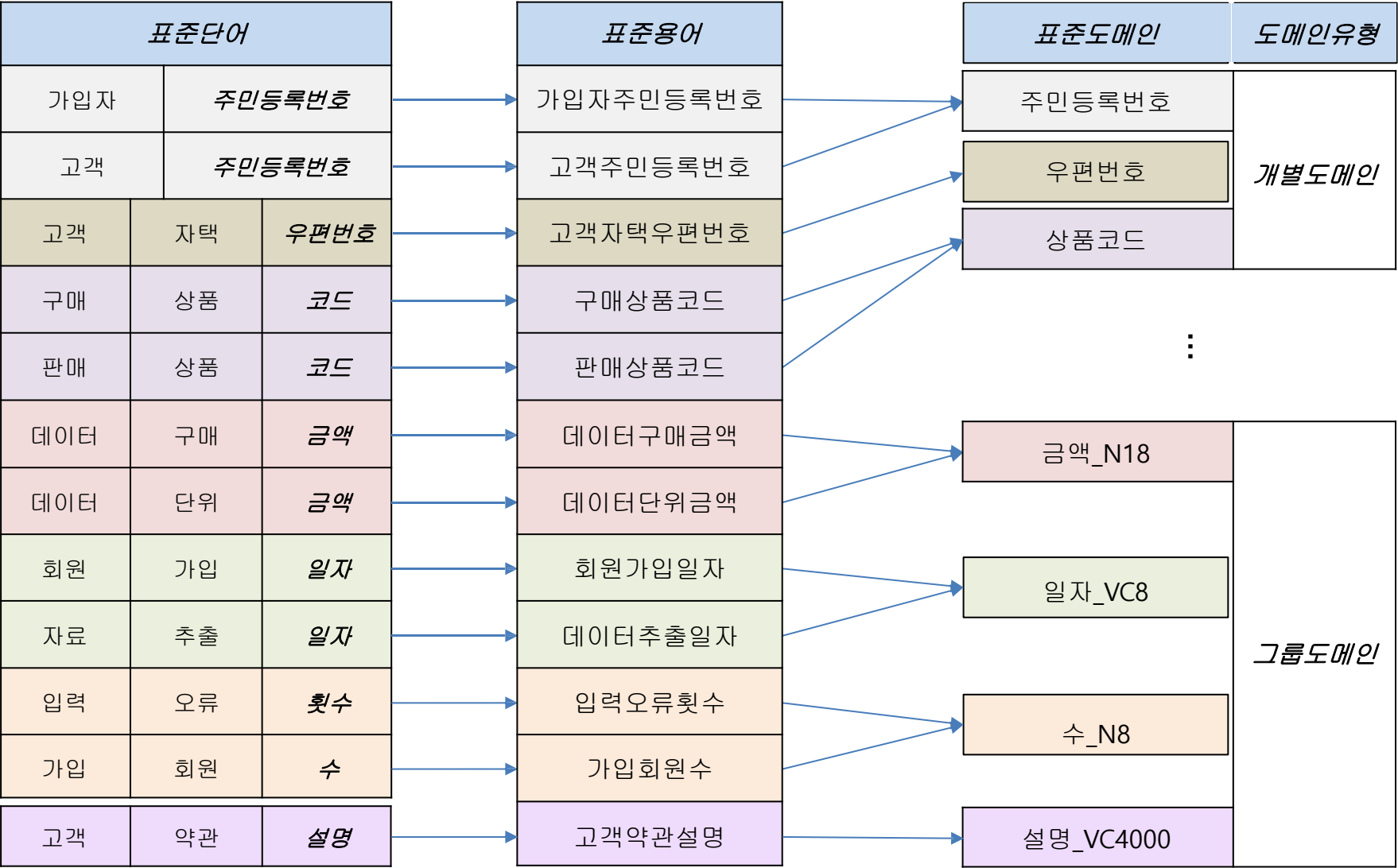
3. 데이터 사전 구성

2) 표준 단어 구성

단어명	영문약어명	단어영문명	단어종류	정의
고객	CUST	customer	기본단어	고객 데이터
가입	JNG	joining	기본단어	가입절차를 통해 정식으로 사용자로 등록
금액	AMT	amount	분류단어	돈의 액수

3. 데이터 사전 구성

3) 표준 도메인 구성



3. 데이터 사전 구성

4) 표준 용어 구성 (1/3)

- 용어의 길이는 데이터베이스 시스템의 제약으로 인해, 용어영문약어명을 기준으로 이 최대 28까지만 허용한다.
- 용어영문명은 첫글자는 대문자로 시작하며 나머지 글자는 소문자를 표준으로 사용한다.
- 용어영문명에 약어명 또는 관용적으로 대문자를 사용하는 경우에는 대문자를 사용할 수 있다.
- 용어영문명을 구성하는 마지막 단어는 반드시 분류단어를 사용하여 구성한다.
- “-“, “_” 를 제외한 특수기호는 사용하지 않는 것을 표준으로 한다.

3. 데이터 사전 구성

4) 표준 용어 구성 (2/3)

- 영문약어명의 첫글자는 숫자로 정의하지 않는다.
- 반복적으로 발생하는 용어명은 업무단어나 도메인에 일련번호를 부여하여 정의한다.

(예 : 입금계좌번호1, 입금계좌번호2)

- 데이터의 성격은 동일하지만 업무적 의미가 다른 경우에는 일련번호로 구분하지 않고 구체적인 용어영문명을 부여한다.

(예 : 고객주소1, 고객주소2 => 고객우편번호주소, 고객상세주소)

3. 데이터 사전 구성

4) 표준 용어 구성 (3/3)

- FROM과 TO를 의미하는 용어는 가능한 시작/종료로 정의한다.

(예 : 이자계산시작일자, 이자계산종료일자)

- 개수를 의미하는 용어를 사용할 경우에는 분류단어로 수(count)를 표준으로 용어를 구성한다.

(예 : 부활고객수 (Reinstatement customer count), 계약자약정건수 (Contractor stipulation count))

학습목차

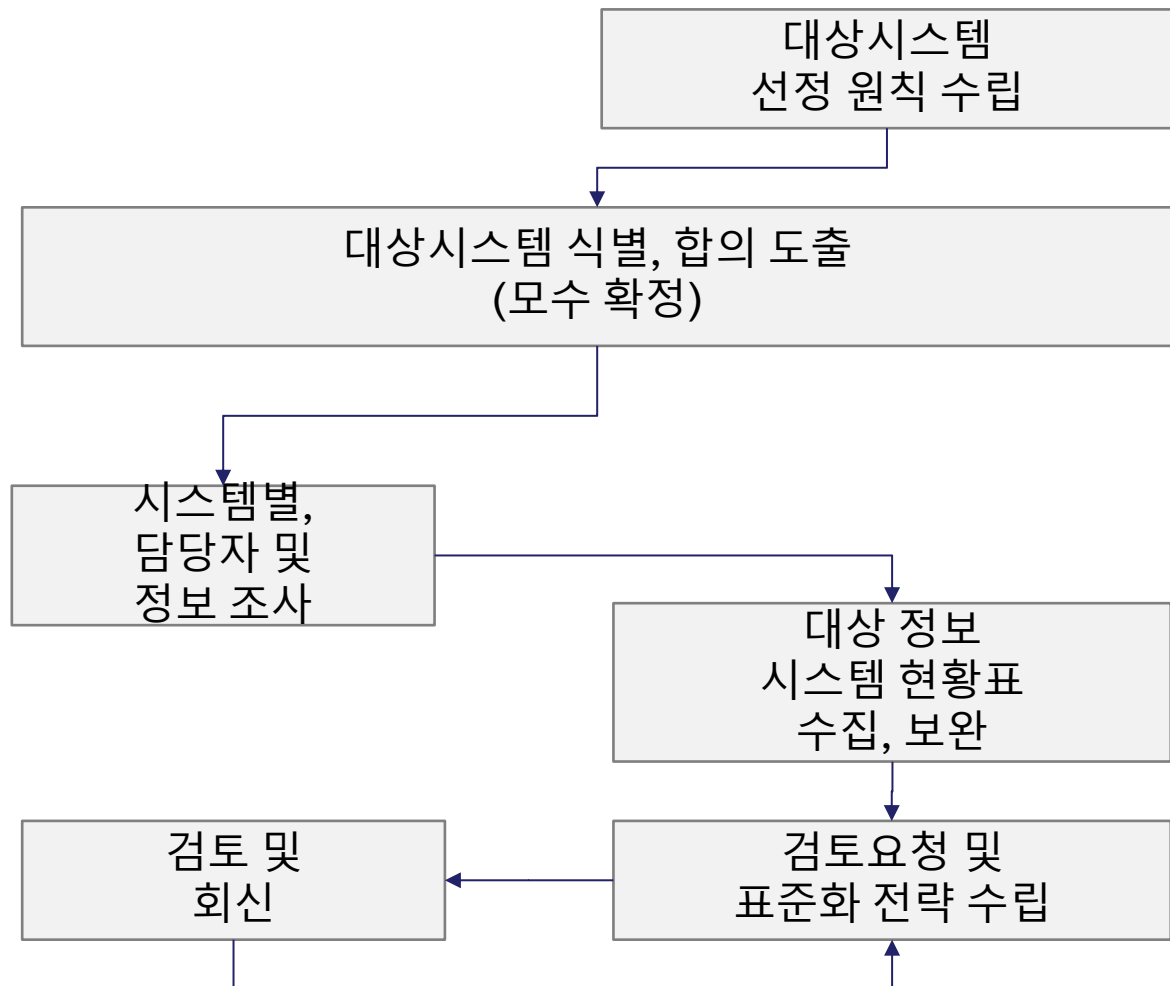
3 데이터 표준화 절차

학습목표

1 데이터 표준화 컨설팅 프로세스를 서술 한다.

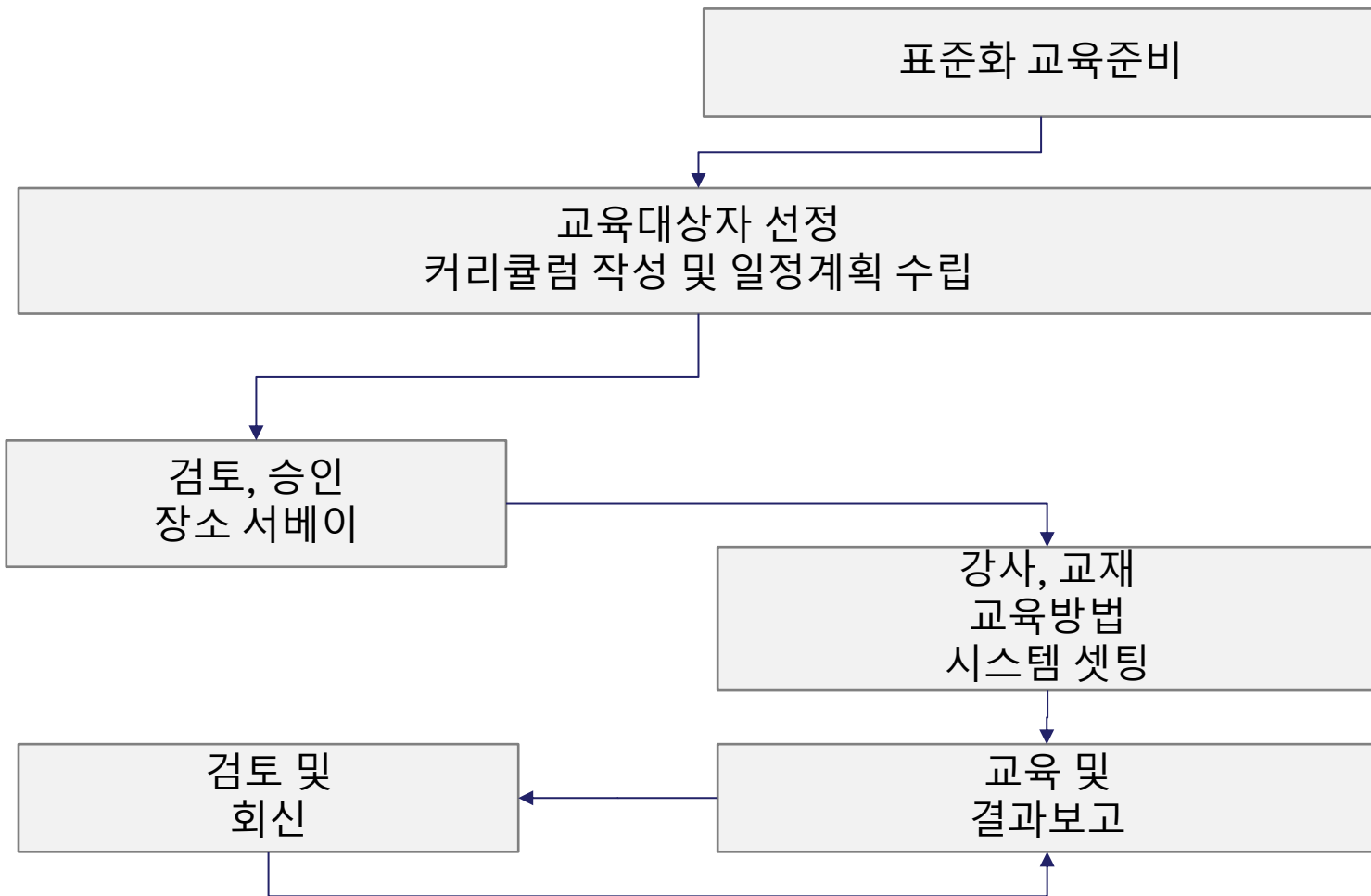
1_ 데이터 표준화 프로세스

1) 현황 조사 단계 (데이터 표준화 준비)



1_ 데이터 표준화 프로세스

1) 현황 조사 단계 (데이터 표준화 방법론 교육)



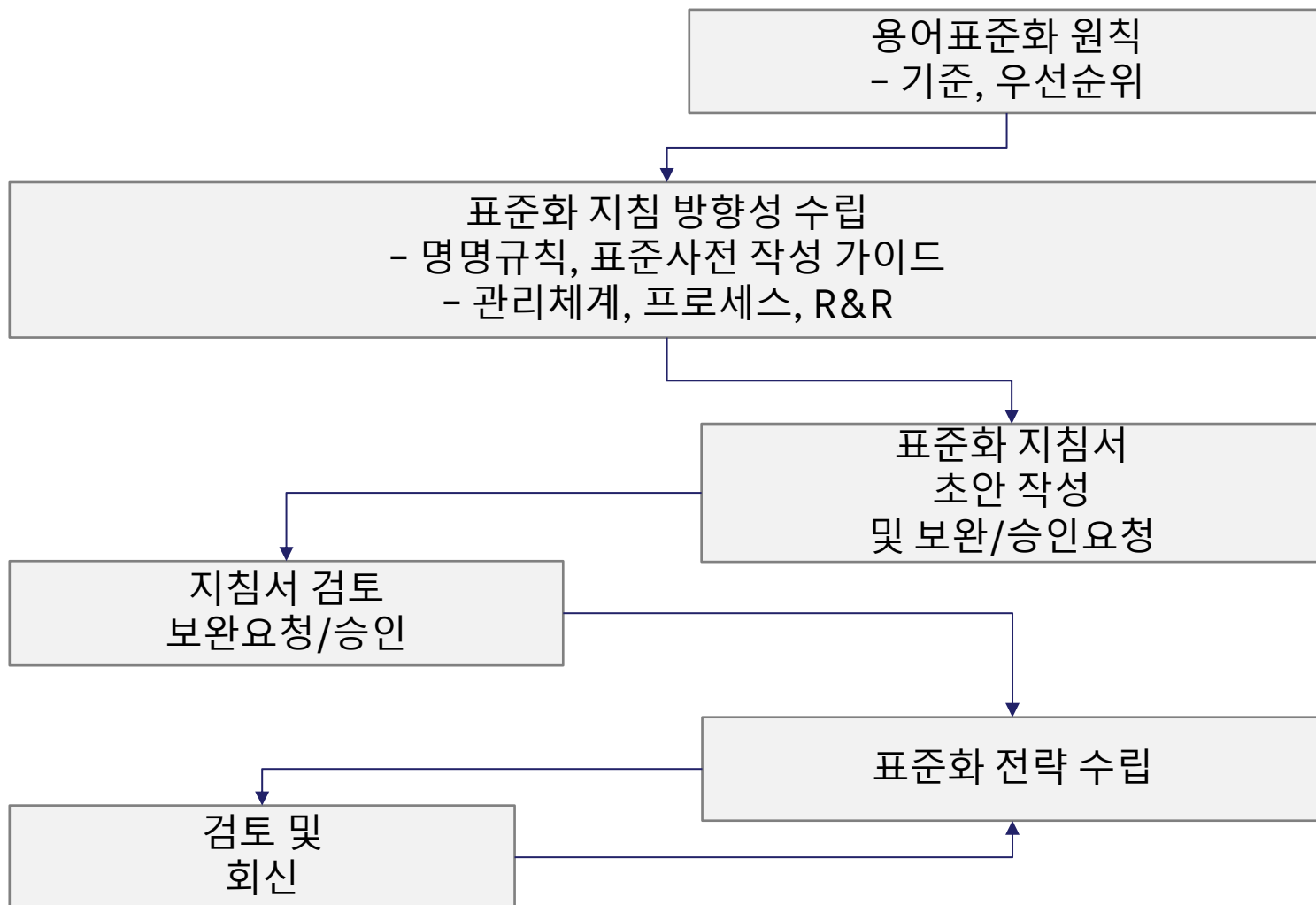
1_ 데이터 표준화 프로세스

1) 현황 조사 단계 (데이터 수집(1))



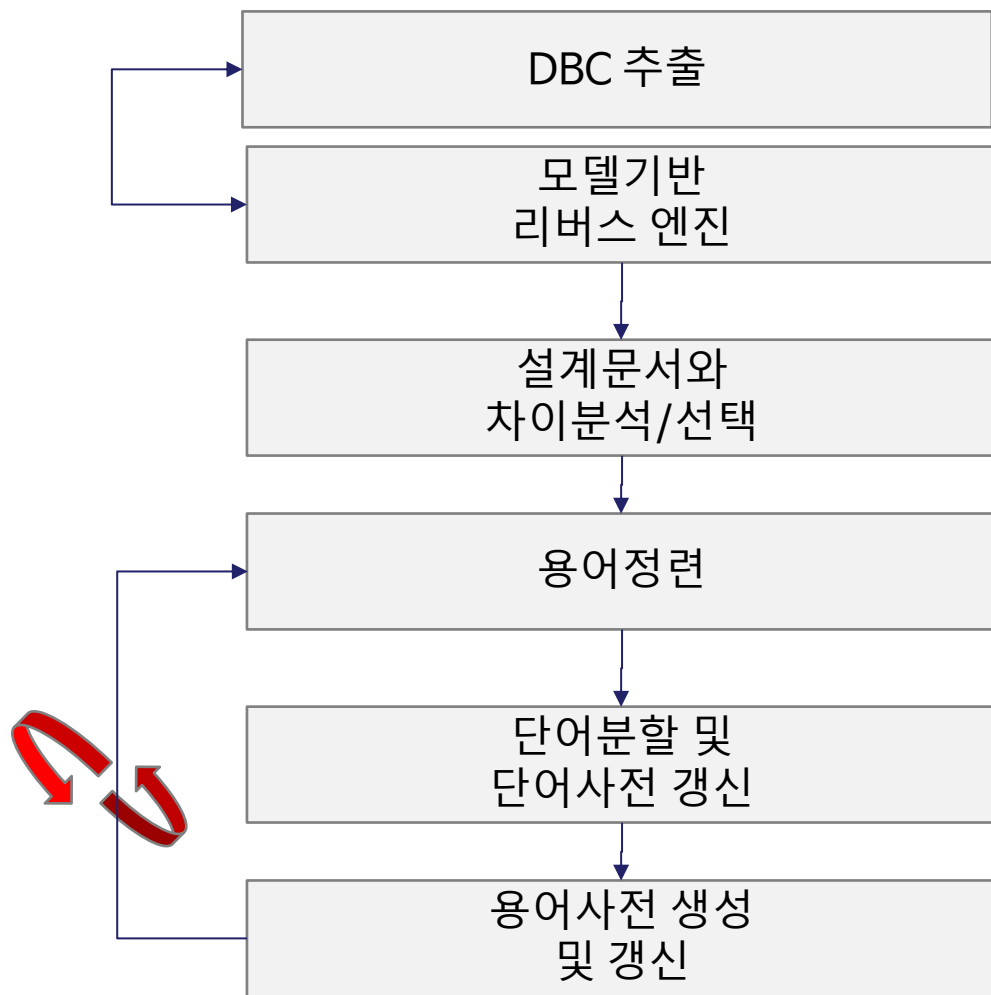
1_ 데이터 표준화 프로세스

2) 데이터사전 수집, 정련, 생성 단계 (표준화 지침서 작성)



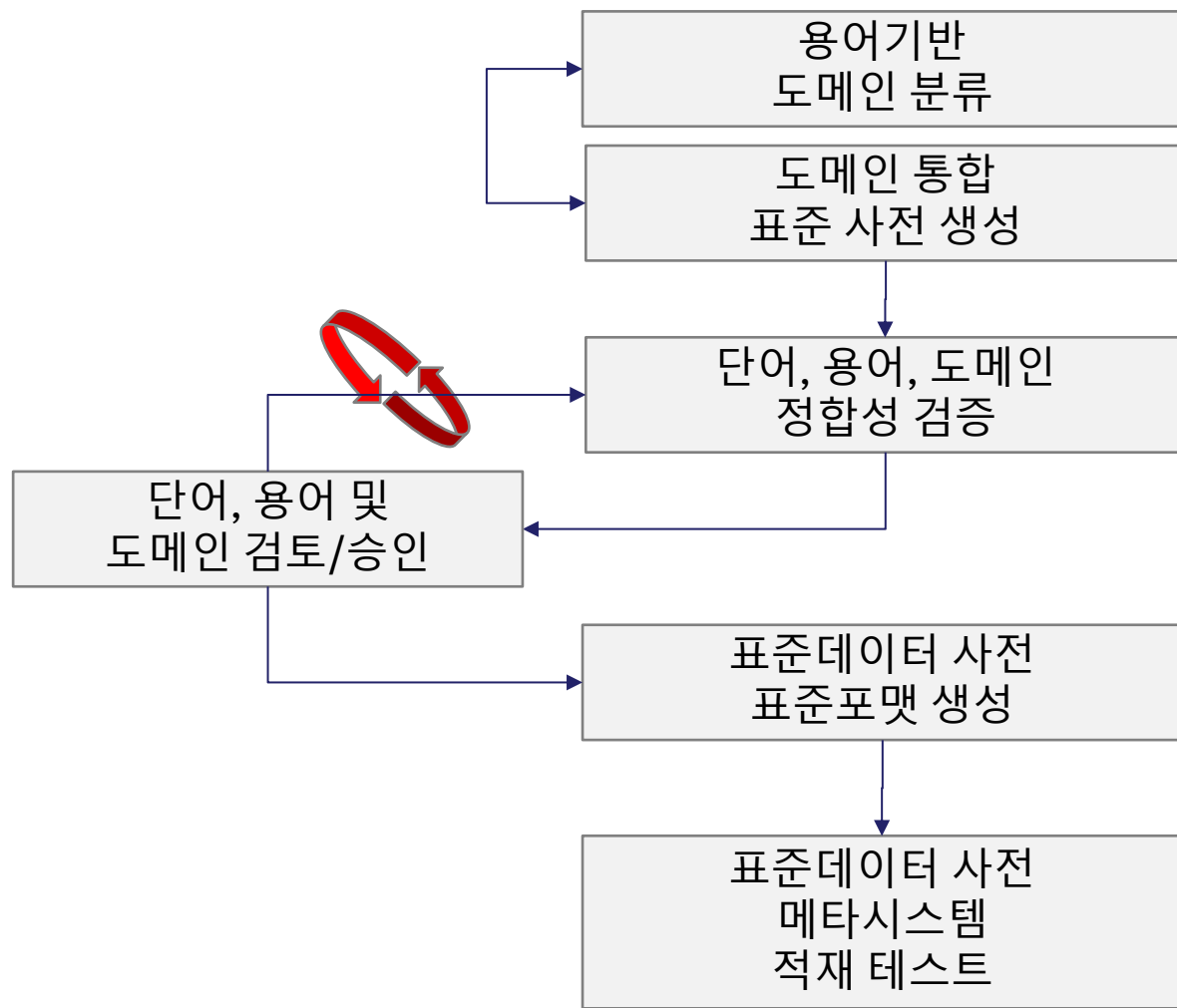
1_ 데이터 표준화 프로세스

2) 데이터사전 수집, 정련, 생성 단계 (데이터 수집 및 표준사전 생성)



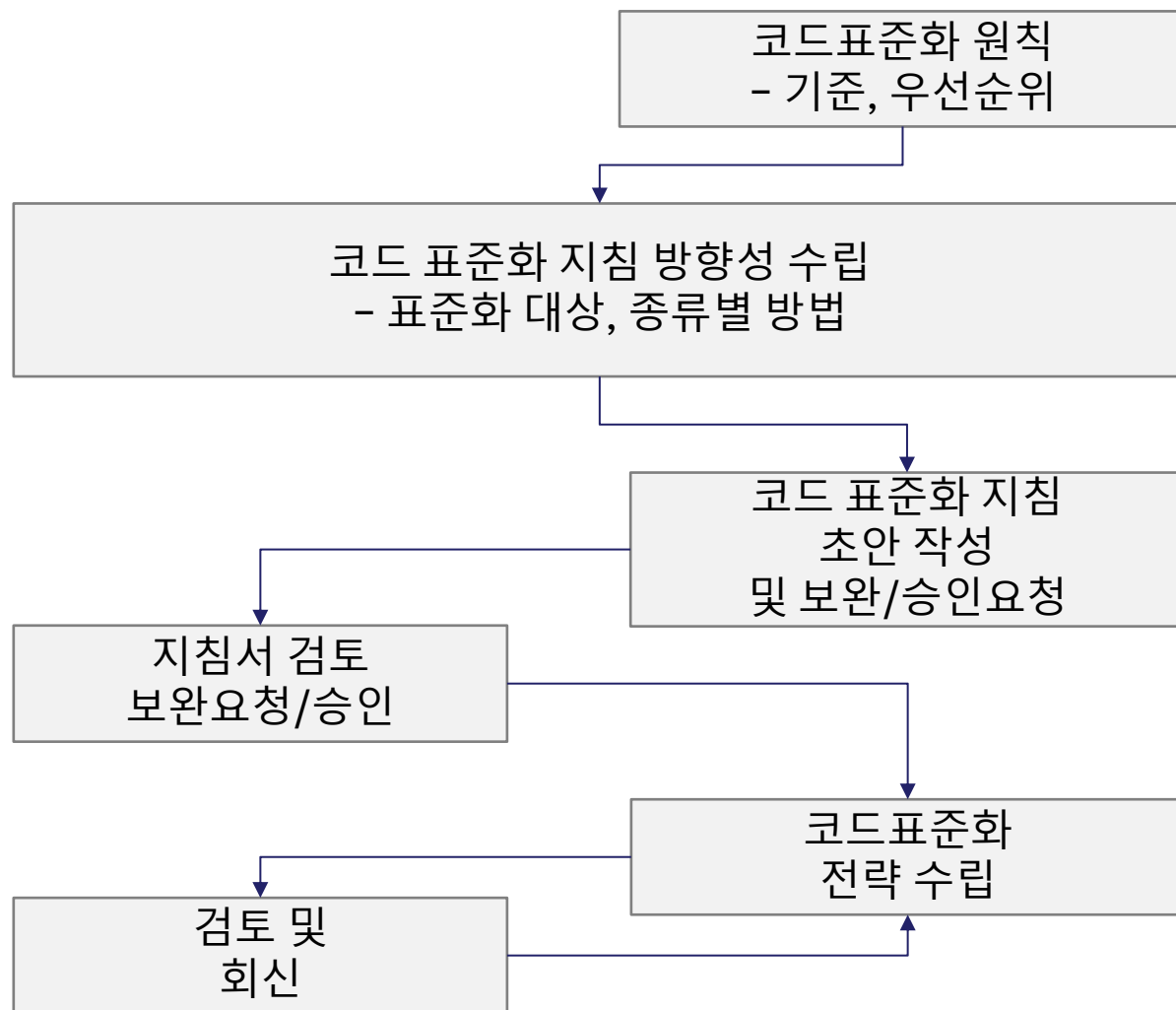
1_ 데이터 표준화 프로세스

2) 데이터사전 수집, 정련, 생성 단계 (표준사전 생성 및 승인)



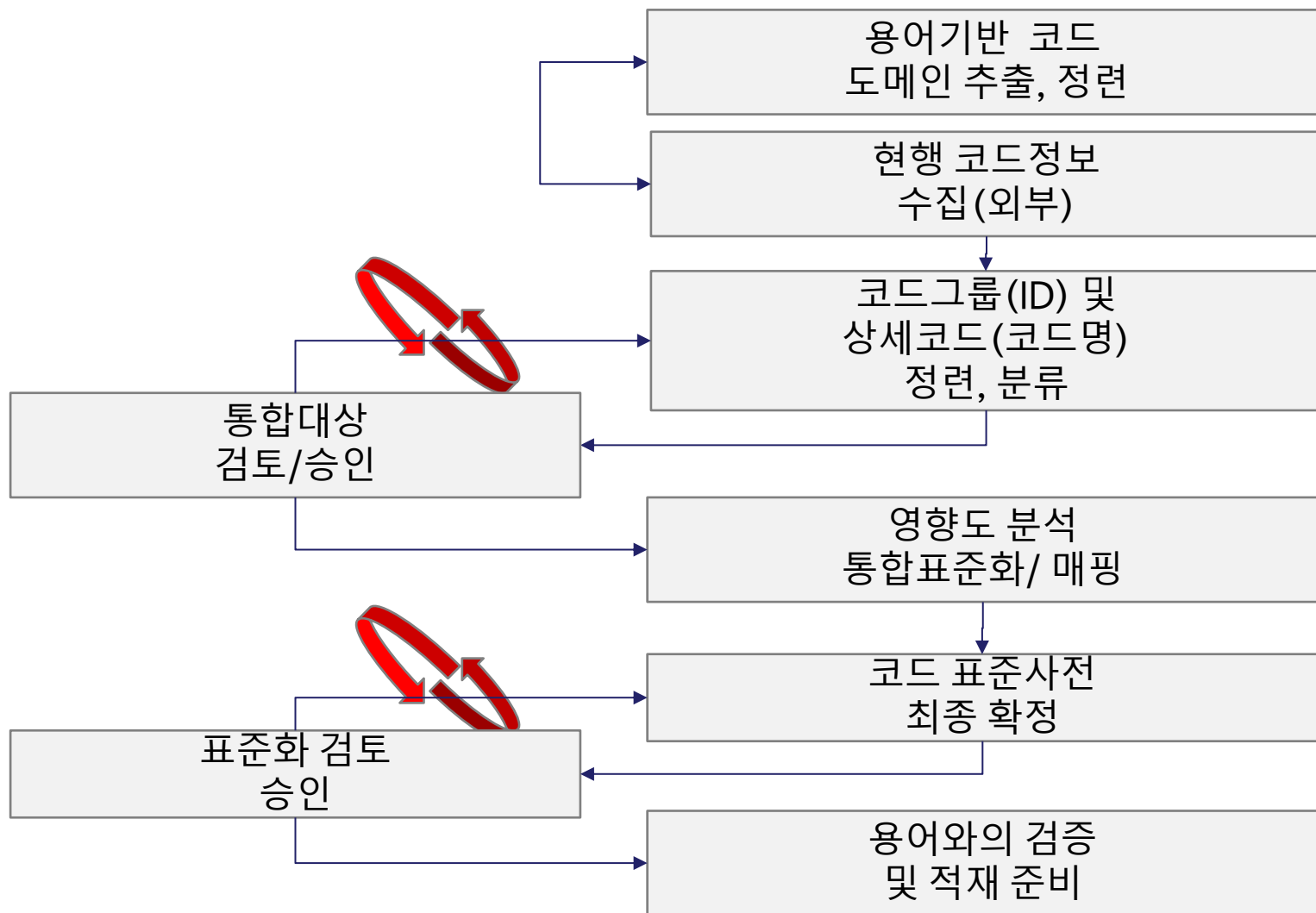
1_ 데이터 표준화 프로세스

2) 데이터사전 수집, 정련, 생성 단계 (코드 표준화 지침 수립)



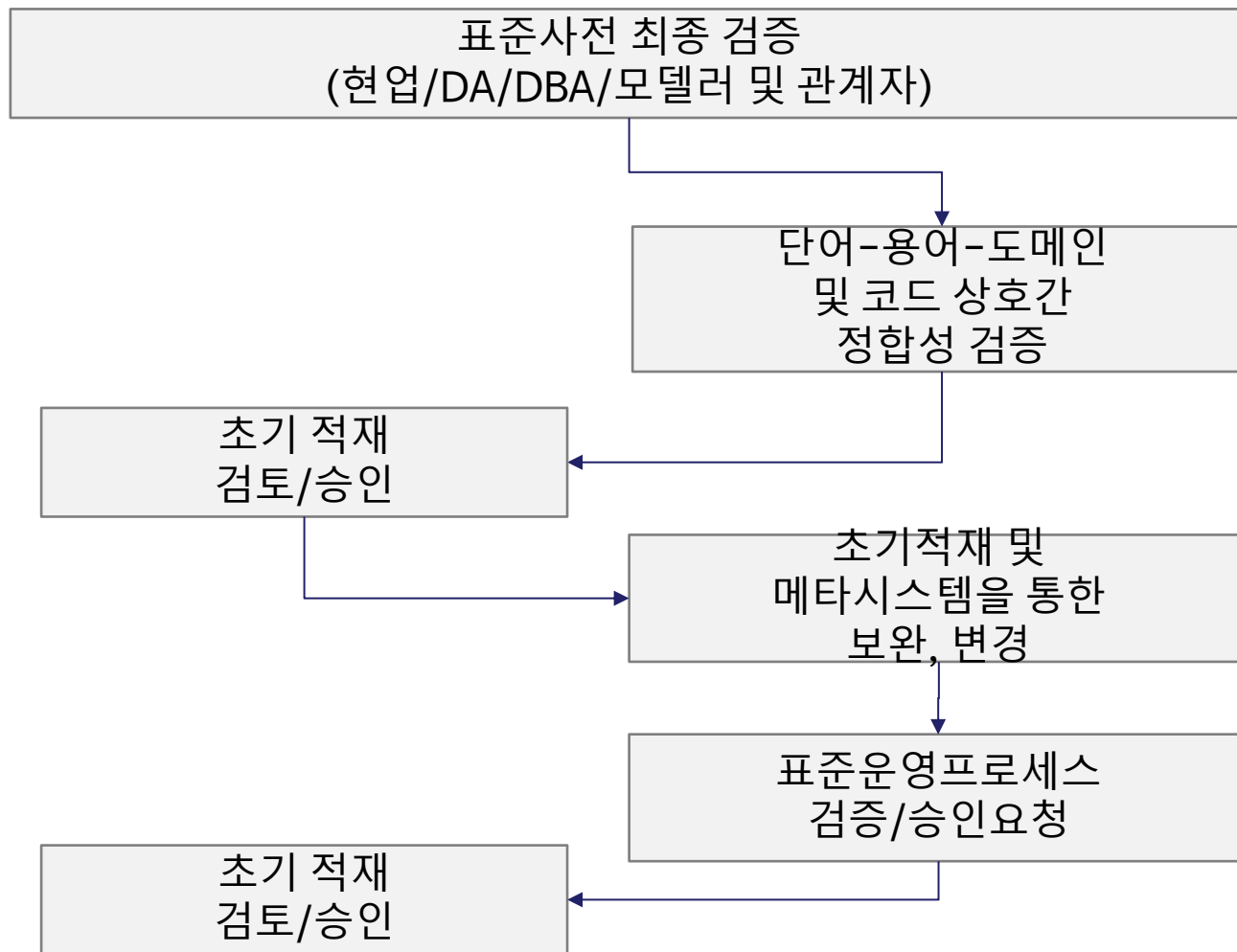
1_ 데이터 표준화 프로세스

2) 데이터사전 수집, 정련, 생성 단계 (코드 표준화 지침 수립)



1_ 데이터 표준화 프로세스

3) 데이터사전 검증 및 초기적재 (최종 검증 및 고객 승인)





데이터 품질

생각해 보기

- 데이터 처리에 있어서 왜 데이터 품질이 중요한가?



학습목차

1

데이터 품질 개념

학습목표

1 데이터 품질이 무엇인지를 설명 한다.

2 데이터 품질의 정의를 설명 한다.

3 데이터 품질의 기대 효과를 설명한다.

데이터 품질의 개념

1) 데이터 품질이란?

- 데이터 품질이란?

- “Consistently meeting all knowledge worker and end-customer expectations through data and data services to accomplish enterprise and customer objectives.”
 - Larry P. English –

“데이터를 활용하는 사용자의 다양한 활용 목적이나 만족도를 지속적으로 충족시킬 수 있는 수준”

데이터 품질의 개념

1) 데이터 품질이란?

- 데이터 품질의 필요성

- 아무리 S/W, 솔루션이 뛰어 나도 데이터가 불량이면 분석 결과를 믿을 수 없음
- 다양한 활용을 위해서는 통합이 필요함
- 검증되지 않은 데이터 아무리 많아봐야 무용지물, 통합 비용은 Loss일 뿐임
- 주기적 검증을 통해 신뢰성 있는 데이터 확보가 핵심 Key

데이터 품질의 개념

1) 데이터 품질이란?

■ 사례를 통해 본 데이터 품질의 필요성

- 데이터 품질관리는 1:10:100의 원칙에 입각하여 이익 관정보다 손실 관점으로 접근

불량을 즉각 고치면 1의 비용이 들지만 책임 소재 등의 이유로 숨겨지거나
지연되면 10의 비용이 들며 이 것이 고객의 손으로 들어가 클레임으로 이어지면
100의 원가가 든다는 법칙임

데이터 오류 사례 및 오류 손실비용		
해외사례	미 해군 안전국	■ 항법시스템의 오류 원인으로 포트로알호의 좌초로 인해 선박의 수리 비용 견적만 4,000만 달러가 나왔으며 수리기간 7개월이라는 손실, 이 사건으로 미 해군은 국제적 망신을 당함
	미국 국립의학 연구원(ION)	■ 의료진에게 제대로 전달되지 않아 잘못된 정보로 인해 엉뚱한 수혈을 받게 된 이 환자는 끝내 사망. 매년 많게는 9만 8000명 환자들이 의료정보 과오로 사망하는 것으로 추정

데이터 품질의 개념

1) 데이터 품질이란?

■ 사례를 통해 본 데이터 품질의 필요성

- 데이터 품질관리는 1:10:100의 원칙에 입각하여 이익 관정보다 손실 관점으로 접근

데이터 오류 사례 및 오류 손실비용		
국내 사례	W 은행	<ul style="list-style-type: none">▪ 고객정보 불일치로 인한 영업기회 상실 비용 약 156억▪ 연간 고객DM 반송 건 약 70만건 / 낭비 비용 약 3억
	금감원	<ul style="list-style-type: none">▪ IT검사과정에서 전산원장과 재무제표상 일부 계정에서 '회계불일치'사례가 발생한 내용이 금융시장에 100조원대 분식회계설로 확대되면서 파장
	국토부	<ul style="list-style-type: none">▪ 2010년 전국 3,733만 필지의 토지·임야대장과 707만동에 대한 건축물 대장 자료를 부동산 등기부와 비교 분석한 결과, 토지·임야대장의 자체 오류가 약 560만건인 것으로 분석
	감사원	<ul style="list-style-type: none">▪ 행안부의 주민전산자료와 연금공단이가입자 이력DB를 대조 확인한 결과 주민등록번호나 이름이 일치하지 않는 것이 30만 9,825건이었으며 해당 징수보험료는 690억원
	국민연금	<ul style="list-style-type: none">▪ 국민연금에 등록된 주민등록번호와 이름이 일치하지 않아 보험료를 적게 지급받아온 사례가 30만여 명에 이름
	주식시장	<ul style="list-style-type: none">▪ 국내 주식시장에서 코스피200 지수가 잘못 산출되어 발표된 사건발생, 코스피200을 구성하는 기업의 시가 총액이 잘못 적용되어 코스피200 지수가 실제보다 높게 산출

데이터 품질의 개념

2) 데이터 품질 관리의 필요성

- 데이터 품질 관리는 왜 필요한가?
 - 데이터 품질에 대한 관리는 기업의 지속적인 경쟁력 확보를 가능하게 함.
- 저 품질 데이터의 위험
 - 대외적 위협 요소와 대내적 위협 요소로 위험 요소가 발생됨

데이터 품질의 개념

2) 데이터 품질 관리의 필요성

■ 위협 요소

대외적 위협 요인

- 낮은 고객만족도
 - (예) 주문한 내역과 다른 상품발송으로 인한 고객불만
- 기업의 경쟁력 약화
 - 데이터 품질 저하로 발생된 손실 비용, 재 작업 비용 등의 증가
- 기업의 대외 이미지 추락
 - (예) 잘못된 데이터로 인한 고객의 피해로 소송이 날로 증가 함으로써 기업의 대외 이미지 감소

대내적 위협 요인

- 조직간의 불신
 - 조직마다 중복되어 관리되는 데이터간의 불일치로 인한 조직 간의 불신을 야기함.
- 의사결정 지연
 - 데이터에 대한 신뢰가 없으므로 의사결정시 데이터의 재확인 및 데이터의 정합성을 확인하는데 많은 시간이 소모됨.
- 운영비용 증가
 - 데이터의 에러를 찾아내고 수정하는데 많은 시간과 자원이 소요됨.

데이터 품질의 개념

2) 데이터 품질 관리의 필요성

- 품질 관리를 통한 위협 요소 극복
 - 고객 품질 제고 마인드형성 및 Data의 중요성에 대한 임직원 공감대 형성
 - Risk 관리 및 마케팅 Offer의 정확성 확보
 - 정보(Data)에 대한 전사 공통 기준 확보
 - 전 사원의 업무 역량 향상 (시스템 데이터의 적극적 활용)
 - 데이터 일관성으로 인한 부서간 원활한 커뮤니케이션

데이터 품질의 개념

3) 데이터 품질 정의

- 고품질의 데이터란?

- 데이터 그 자체보다는 사용하고자 하는 목적에 따라 해당 용도의 요구 사항을 만족시킬 수 있어야 함.
- 정확도는 있으나 조직에 있어 유용성,가치가 없는 데이터는 관리가 불필요함.

- 데이터 품질의 정의

- 지식 노동자와 최종 소비자의 기대에 일관성 있게 부합하는 데이터의 상태
- 지식 작업자가 데이터를 활용하여 업무목적을 달성하는 데이터의 수준
- 품질이 확보된 데이터란 사용하기에 문제가 없고, 사용에 따른 효과를 얻을 수 있는 데이터의 상태

데이터 품질의 개념

3) 데이터 품질 정의

- 데이터 품질의 대상

- 정의 (Definition)

- ▶ 데이터 사양 및 메타 데이터에 대한 품질

- ▶ 표준화 & 모델링 관점

- ▶ “데이터 구조가 얼마나 견고하고 유연하게 설계가 되었는가?”

- 값 (Content)

- ▶ 데이터 값의 정확성에 대한 품질

- ▶ 완전성, 정확성 관점

- ▶ “데이터 값이 얼마나 정확히 나타내고 작업을 효과적으로 수행하는데 필요한 정보인가?”

데이터 품질의 개념

3) 데이터 품질 정의

- 데이터 품질의 대상

- 표현 (Presentation)

- ▶ 지식 작업자(Knowledge Worker)에게 전달되는 정보 제품으로서의 품질
 - ▶ 적시성, 편리성, 활용성 관점
 - ▶ “필요할 때 즉시 제대로 된 정보를 얻을 수 있는가?”

데이터 품질의 개념

4) 데이터 품질 관리 정의

■ 데이터 품질 관리의 정의

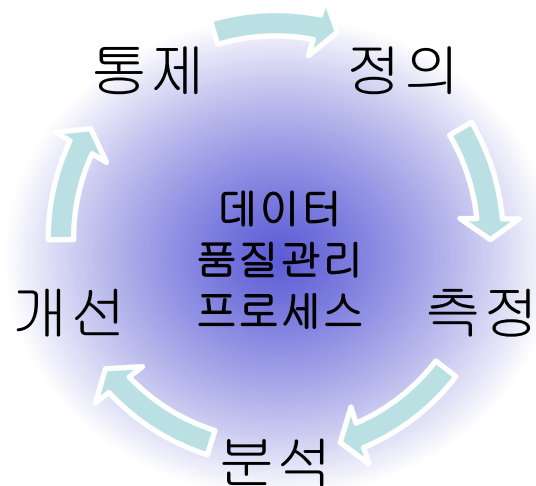
- 데이터의 품질을 지속적으로 유지하고 개선함으로써 사용자의 만족도를 극대화하기 위해 수행하는 일련의 활동
- 데이터를 사용하는 사용자의 만족도를 충족시키는 수준을 만드는 과정
- 초기 데이터 품질을 획득하고 이를 지속적 유지/개선 할 수 있는 프로세스 및 시스템

데이터 품질의 개념

5) 데이터 품질 관리 프로세스

■ 데이터 품질 관리 프로세스

- 고객 관점의 데이터 품질 관리 프로세스는 정의, 측정, 분석, 개선, 통제 단계로 이루어지며 이는 지속적이며 반복적으로 수행되어야 함.
- 측정 가능하며 가시적이고 명확해야 함.
- 특정 IT기술에 의존하여 수행되는 프로세스가 아님.



데이터 품질의 개념

5) 데이터 품질 동향

■ 데이터 품질 동향

- 저품질 데이터 생성을 예방하기 위해 국내외 각 정부 및 산하 단체 기업에서 데이터의 표준 및 가이드라인을 개발하여 배포하고 있으며 데이터 품질 활동을 의무화하고 있음.

미국 데이터 품질법 (Data Quality Act)	Public Law 106-554, Sec 515 미국 정부는 2000년 12월 데이터 품질법을 제정하여 예산 관리국(OMB)으로 하여금 정보 품질관리 가이드라인을 개발하게 하고, 각 연방 정부기관은 OMB 가이드라인에 따라 기관별 품질 가이드라인의 개발 및 적용을 의무화 함으로써, 정부기관에서 배포하는 정보의 품질확보를 도모하고 있음.
미국 국방성 데이터 품질관리 가이드라인	DOD Guidelines on Data Quality Management 미국 국방성은 조직 운영의 효율성 극대화와 시스템 통합/이행 및 정보 시스템의 자동화에 따라 대두되는 저품질 데이터 문제를 개선하기 위해, TDQM(Total Data Quality Management)방법론을 정립하고 품질 개선 프로세스에 따라 체계적이며 지속적인 데이터 품질관리를 실시하고 있음.
데이터 품질 관리 인증	데이터 품질 관리 인증(DQMC: Data Quality Management Certification)을 위해 한국 데이터베이스 진흥센터 부설로 '데이터 품질 관리 인증센터'를 설립하고 2006년 10월부터 국내 공공기관을 중심으로 데이터 품질 관리 인증 심사를 실시함.

데이터 품질의 개념

6) 기대 효과

- 데이터 품질 기대 효과

- 데이터 신뢰도 향상

- ▶ 데이터 관리절차의 체계화로 정보의 신뢰도 향상
 - ▶ 고객 만족도 향상, 신용 증대(고객 충성도, 고객 유지율 향상)
 - ▶ CRM 등의 정보기반을 다짐.

- 생산성 / 효율성 향상

- ▶ 개발,변경시스템에 대한 신뢰성 증대
 - ▶ 낭비 없는 프로세스에 의한 생산성 향상
 - ▶ System의 신뢰도 향상
 - ▶ 품질 현황을 분석하고 개선 체계를 확립

데이터 품질의 개념

6) 기대 효과

- 데이터 품질 기대 효과

- 비용절감

- ▶ 데이터 재정비에 드는 추가 시간과 비용 절약
 - ▶ 마케팅 분야에서 비용 감소(이중 메일발송 등의 추가지출 절약, 정확한 표적고객 선택)
 - ▶ 데이터 오류로 인한 예산 낭비 방지

- 데이터 활용성 증대

- ▶ 믿을 수 있는 데이터로 전사 타 시스템에서의 활용성 증대

학습목차

2 데이터 품질 관리 구성 요소 설명

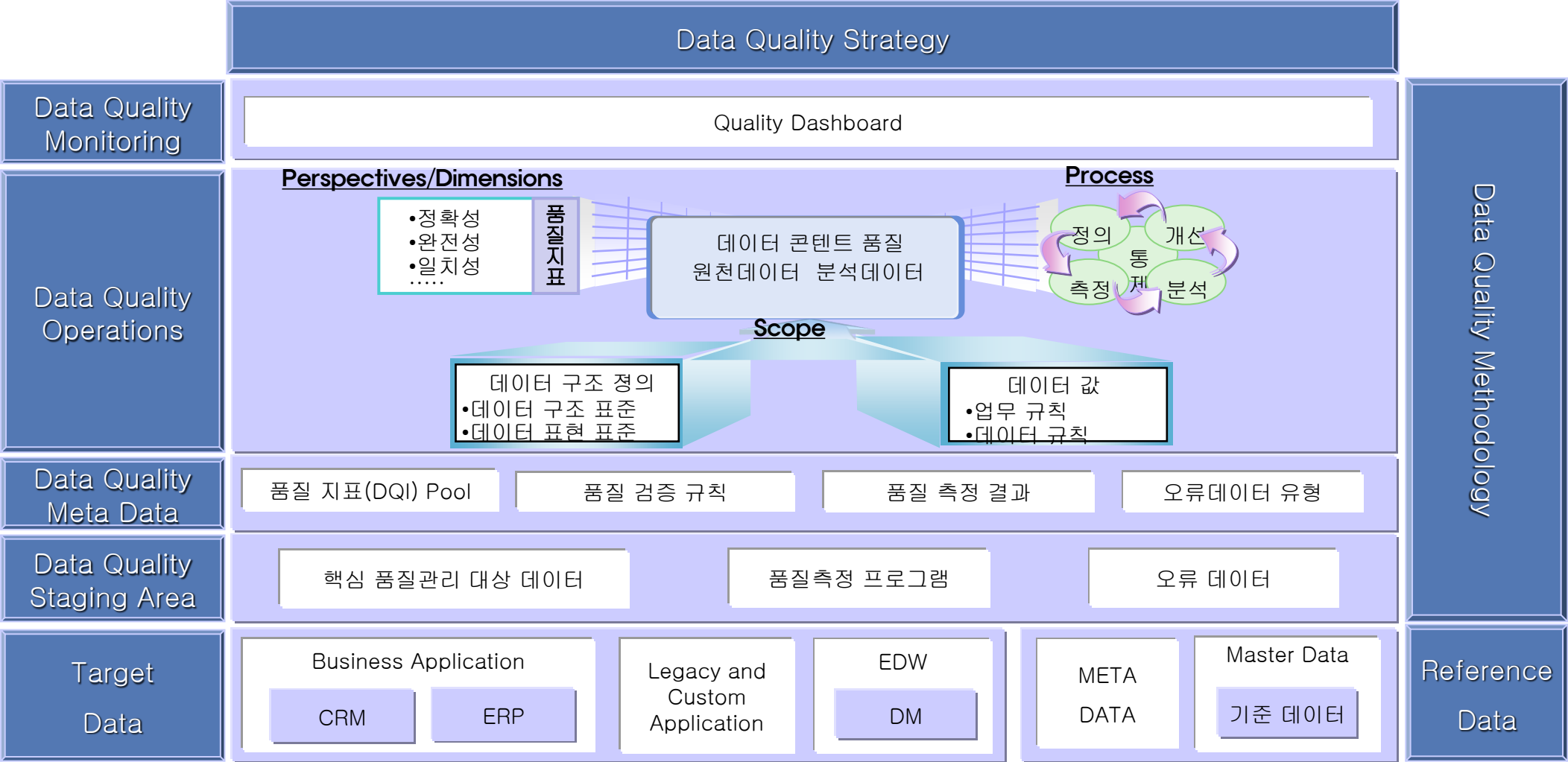
학습목표

1 데이터 품질 관리 구성 요소를 설명 한다.

2 데이터 품질 관리 프로세스를 설명한다.

데이터 품질 관리 구성 요소

1) 데이터 품질 관리 구성 요소 아키텍처



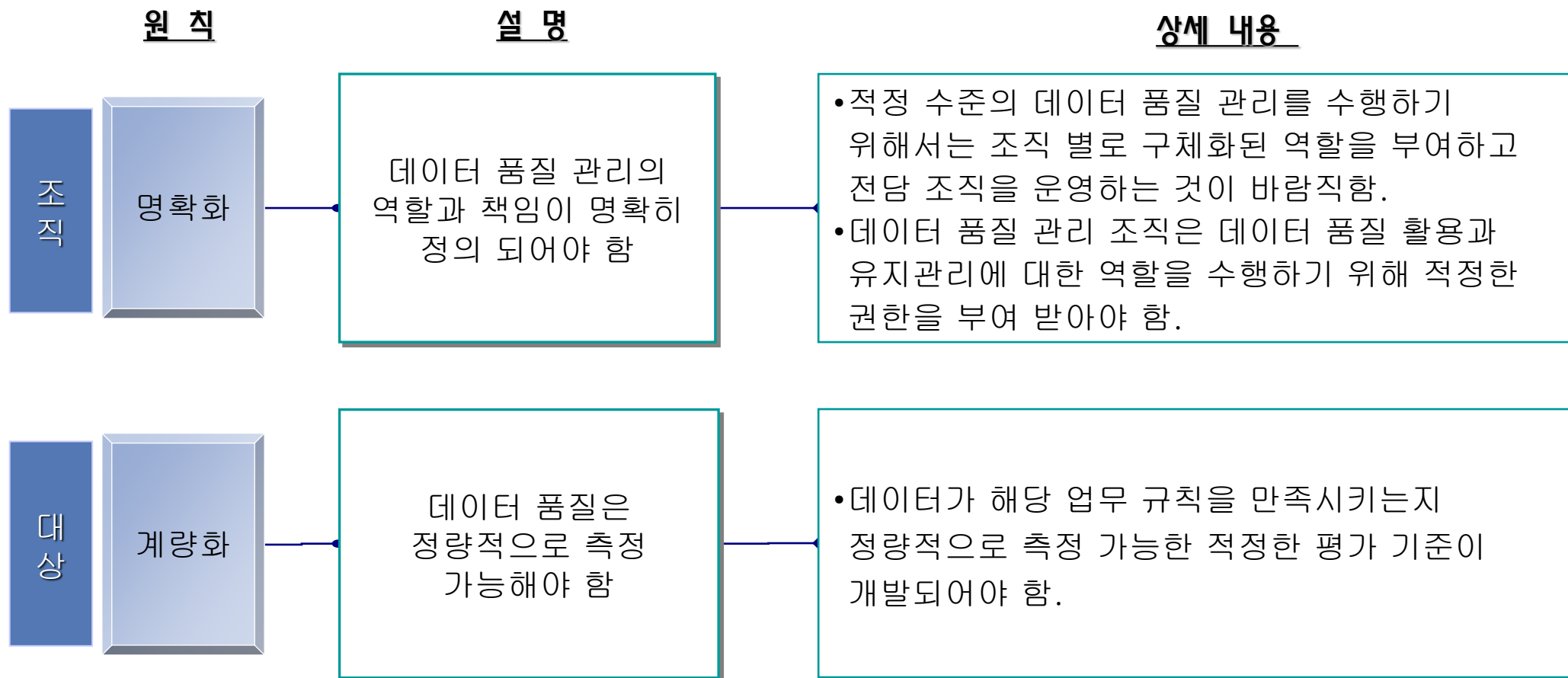
데이터 품질 관리 구성 요소

1) 데이터 품질 관리 구성 요소 아키텍처

구성 요소		설 명
Data Quality Strategy	Policy	전사 데이터 품질관리를 위한 비전 및 전략
	Organization	데이터 품질관리를 위한 조직 및 해당 조직 구성원의 역할과 책임에 대한 정의
Data Quality Methodology		체계적인 데이터 품질관리를 위한 절차 및 가이드라인
Data Quality Monitoring		다양한 관점으로 계량화된 데이터 품질 수준을 Dashboard 등을 통해 지속적으로 모니터링
Data Quality Operations	데이터 콘텐츠 품질	데이터 품질관리의 대상으로 CTQ 항목을 선정하여 관리 CTQ (Critical To Quality) = 데이터의 품질이 기업 경영에 중요한 영향을 미치는 우선 선정 대상이 되는 정보 항목
	품질 지표(DQI)	데이터의 가치와 신뢰성을 평가하는 기준이자 데이터를 바라보는 관점
	프로세스	데이터 품질관리를 위한 운영 프로세스 (정의-측정-분석-개선-통제)
	Scope	데이터 품질관리 대상에 대한 품질 측정 항목/범위(데이터 구조,데이터 값) - 데이터 구조 정의 : 데이터 구조 정의 값 (예:컬럼명칭) 및 표현 형태 (예:저장 패턴) - 데이터 값 : 데이터 구조에 저장된 실제 값(업무 데이터)
Data Quality Meta Data		데이터 품질관리 활동을 위해 필요한 메타성 기초 정보들과 그 결과로 생성된 다양한 데이터들
Data Quality Staging Area		데이터 품질 측정 및 분석을 위한 작업 영역으로 Data Source로부터 품질 관리 대상 데이터를 추출하여 적재
Target Data		데이터 품질관리가 필요한 대상 데이터
Reference Data		데이터 품질관리에 참조되는 데이터 (측정 대상의 기초 정보 및 측정 시 기준 데이터 등)

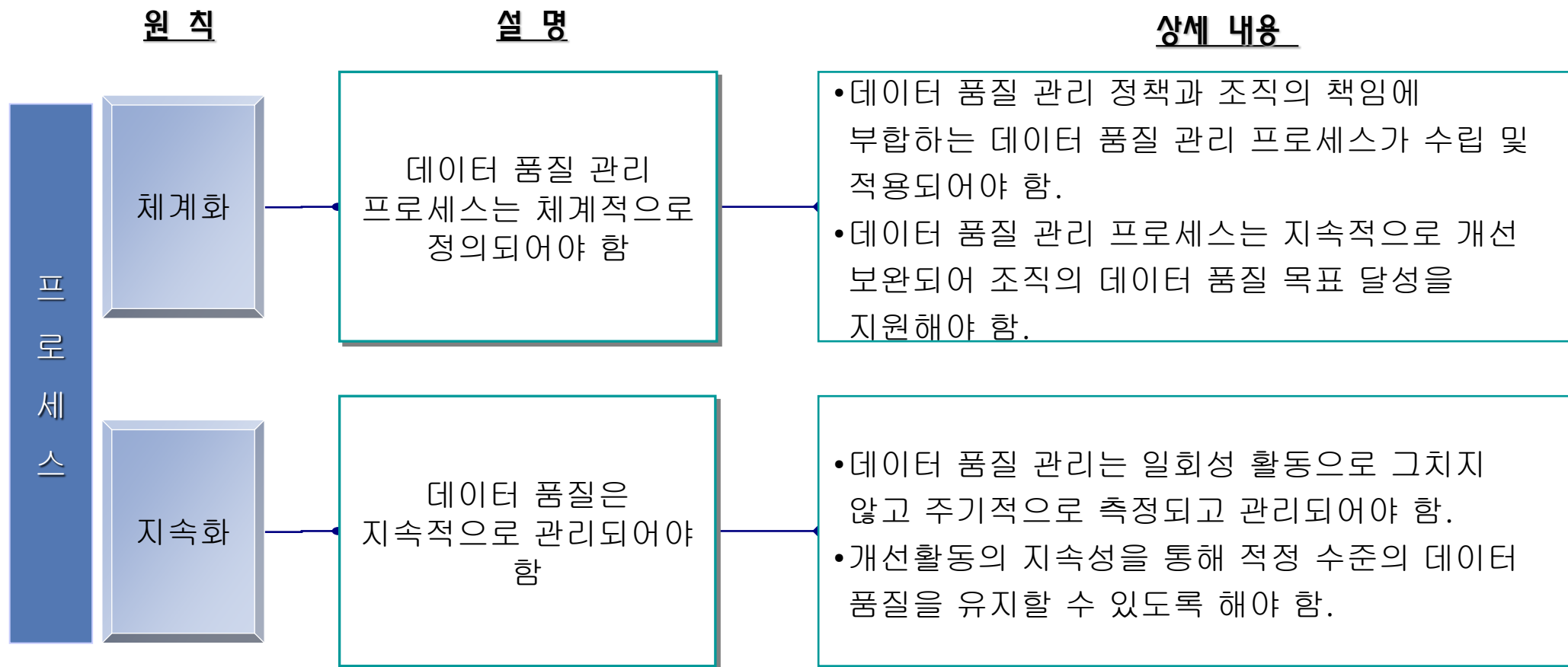
데이터 품질 관리 구성 요소

2) 데이터 품질 관리 원칙



데이터 품질 관리 구성 요소

2) 데이터 품질 관리 원칙



데이터 품질 관리 구성 요소

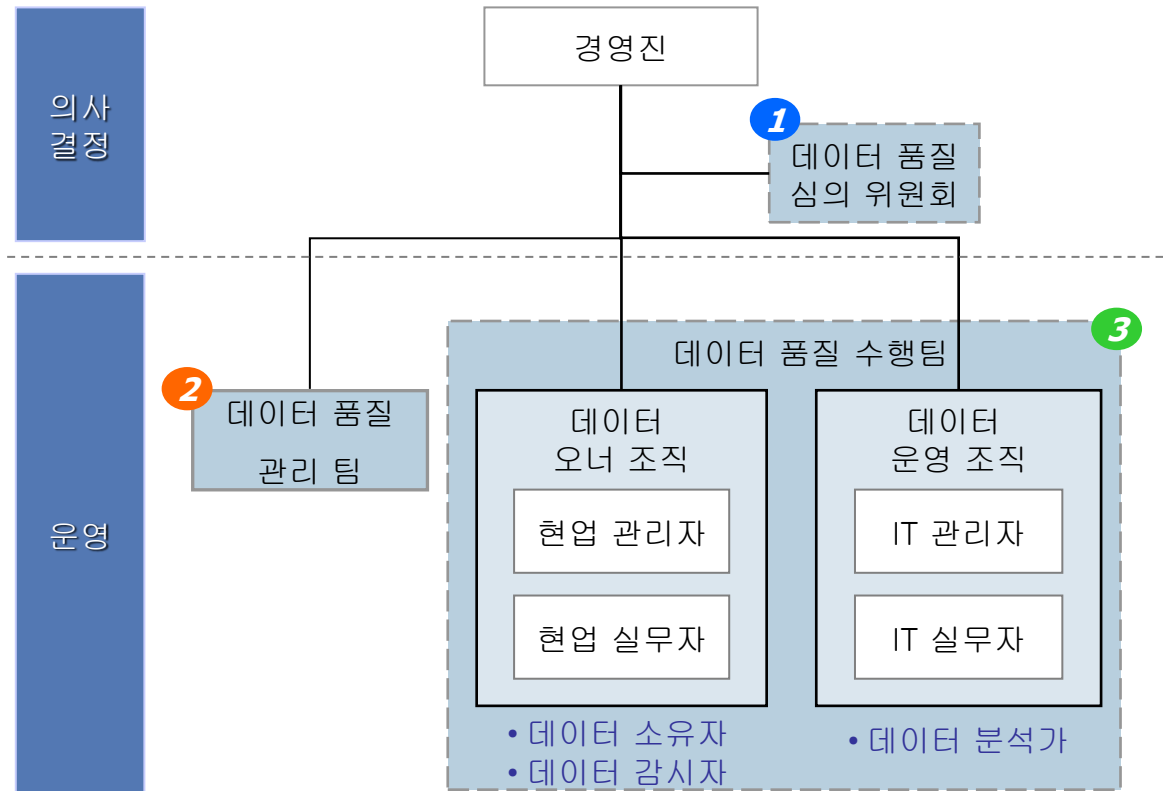
2) 데이터 품질 관리 원칙

■ 핵심 성공 요소

- 데이터 품질 관리는 기술만으로 성공할 수 없는 영역으로 IT문제가 아닌 비즈니스 문제라는 것을 인식하도록 함.
- 해당 기업의 데이터와 데이터 품질에 대한 명확한 정의를 내리고, 조직의 전략상 중요한 데이터에 자원을 집중하도록 함.
- 공식적이고 정규화된 데이터 품질 관리 체계를 만들도록 함.

데이터 품질 관리 구성 요소

3) 데이터 품질 관리 조직



 : DQ 상시 조직  : DQ 가상 조직

1

데이터 품질 심의위원회

- 각 업무 영역에 대한 의사 결정 권한을 가진 인원으로 구성
- 가상 조직으로, 필요 시 안건에 따라 관련 인원 소집

2

데이터 품질 관리 팀

- 업무와 IT에 대해 전반적인 지식을 가진 사람들로 구성
- 상시 조직으로, 전사의 데이터 품질 체계에 대한 운영 및 관리의 주체

3

데이터 품질 수행팀

- 현업 부서와 IT 부서원들로 혼합 구성되며, 현존하는 조직도 위에 가상으로 구성
- DQ 프로젝트 수행 시: 해당 업무의 영역별 전문가로 구성되어 실 프로젝트 수행
- DQ 프로젝트 완료 후: 해당 영역의 데이터 품질에 대한 관리의 책임

데이터 품질 관리 구성 요소

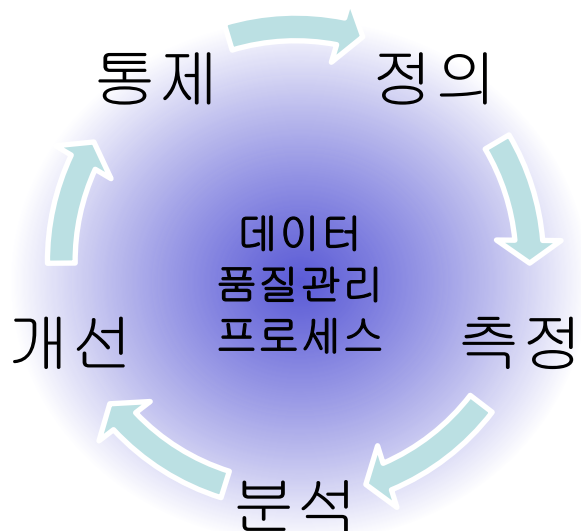
3) 데이터 품질 관리 조직

■ 역할과 책임

조직	역할	책임
데이터 품질 심의 위원회	의사 결정	<ul style="list-style-type: none">• 데이터 품질에 대한 전략적 자문 제공• 데이터 품질 활동 지원을 위한 방침 마련 및 정책 실행• 전사 목표를 선정하고 프로젝트를 선택하며 이를 조율하는 역할 수행
데이터 품질 관리팀	데이터 품질 관리	<ul style="list-style-type: none">• DQM 체계 유지 관리• 전사 데이터 문제 해결 시 코디네이터 역할• 전사 데이터 품질 수준에 대한 모니터링 및 통합 운영 관리 책임
데이터 품질 수행팀	데이터 소유	<ul style="list-style-type: none">• 담당 업무영역에서의 업무 관점의 품질 요건 정의• 담당 업무영역에서의 데이터 품질 문제/이슈 해결
	데이터 감시	<ul style="list-style-type: none">• 담당 업무영역의 데이터 품질 현황 모니터링• 담당 업무영역의 품질 문제에 대한 원인 분석 및 개선 방안 모색• 개선 방안 실시
	데이터 분석	<ul style="list-style-type: none">• 데이터 관점의 품질 요건 정의• 데이터 품질 수준 측정을 위한 개발 및 측정 프로세스 운영• 오류 데이터에 대한 원인 분석 및 개선 방안 지원• 개선 방안 실시

데이터 품질 관리 프로세스 (1/2)

1) 데이터 품질 관리 프로세스란?



정의

- 핵심 품질 항목(CTQ)를 선정하고 데이터 품질 측정을 위한 기준 지표(DQI)를 정의한다.
- 품질 관리 대상 데이터에 대한 정보를 수집하여 품질 검증 규칙을 정의한다.

측정

- 데이터 품질을 측정하기 위한 아키텍처를 정의하고 데이터 품질 평가 환경을 구축한다.
- 품질 측정 대상에 대해 정의된 검증 규칙에 따라 현재 품질 수준을 평가한다.

분석

- 데이터 품질 측정 결과를 분석하여 발생한 문제의 유형과 해당 문제로 인한 업무적 영향을 분석하고 근본 원인을 도출한다.

개선

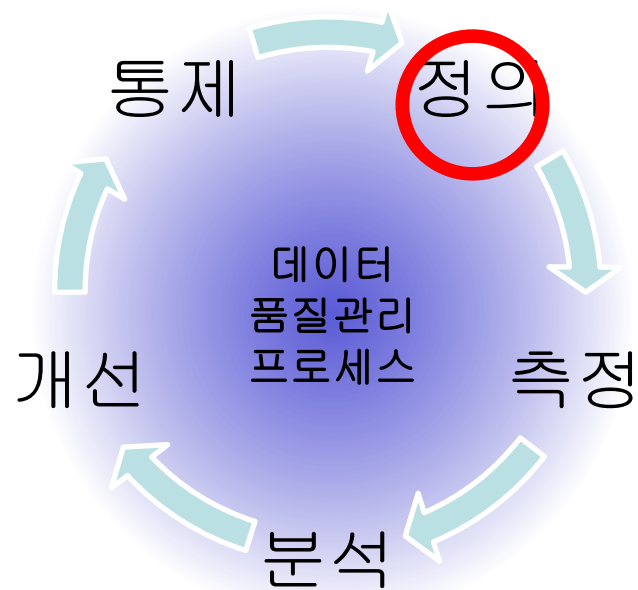
- 근본 원인별로 가능한 개선 방안을 도출하고 개선의 우선 순위를 선정한다.
- 개선안에 대해 일정 계획을 수립하여 개선 작업을 수행한다.

통제

- 개선 수행 결과를 통하여 개선 전/후를 평가하고 그 효과를 분석한다.
- 품질 관리 체계를 수립하고 데이터 품질 관리 프로세스는 지속적으로 개선 보완되도록 한다.

데이터 품질 관리 프로세스 (1/2)

2) 정의



품질 관리 대상 선정

핵심 품질 항목(CTQ)를 선정

DQI 정의

데이터 품질 관리에 필요한 DQI와 세부 항목을 정의

품질 평가 기준 정의

품질 평가 기준 및 관리 방법 정의

품질 검증 규칙 도출

품질 관리 영역에 대한 품질 검증 규칙을 도출

DQI 매핑

품질 검증 규칙과 DQI를 매핑함

※ CTQ (Critical to Quality): 데이터의 신뢰도(품질)가 기업 경영에 중요한 영향을 미치는 데이터 품질측정 및 관리의 우선 선정 대상이 되는 정보 항목

데이터 품질 관리 프로세스 (1/2)

2) 정의

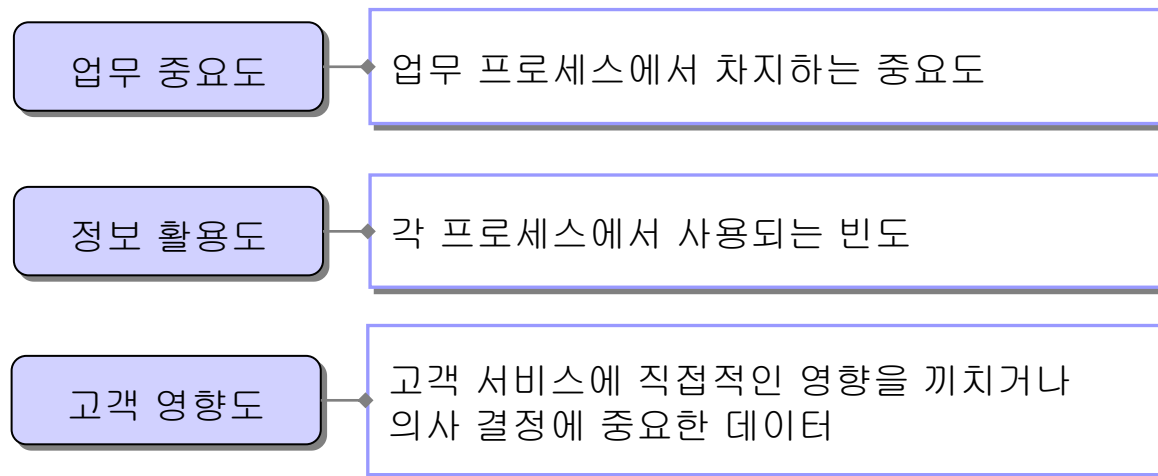
■ 정의 및 선정 기준

– CTQ (Critical To Quality) 란 ?

▶ 데이터의 신뢰도가 고객, 프로세스 및 시장 환경 등 기업 경영에 중요한 영향을 미치는 정보 항목

▶ 데이터 품질 관리의 우선 대상이 되는 정보 항목

■ 선정 기준 [예시]



- CTQ의 단위는 상황에 따라 다를 수 있으며, 이해를 돕기 위해 계층 구조를 가질 수 있음

데이터 품질 관리 프로세스 (1/2)

2) 정의

- DQI (데이터품질지표) 정의

- － 품질 측정 기준

- ▶ 지속적으로 품질 점검을 통해 관리되어야 할 평가 기준이자 데이터 품질을 바라보는 관점
 - ▶ DQI는 조직의 상황에 따라 다양하게 재정의될 수 있으나,
 - ▶ 조직에 적합한 DQI를 선택한 후에는 반드시 각 DQI에 대해 명확히 정의해야 함.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ DQI (데이터 품질 지표) Pool 일부 발췌

DQI	상세 유형	정의
완전성	컬럼 완전성	필수 입력 컬럼은 값이 채워져 있어야 함.
	레코드 완전성	업무적으로 발생해야 하는 레코드는 누락되지 않아야 함.
정확성	데이터 타입 정확성	컬럼의 값은 정의된 데이터 타입과 맞아야 함.
	데이터 허용값 정확성	정의된 도메인 및 허용된 범위 내의 값이 발생해야 함.
	참조 무결성	테이블간 논리적 상호 참조 관계를 준수해야 함.
일관성	계산 집계 일관성	결과 값은 계산규칙을 준수해야 함.
	데이터 일치성	중복된 값(비정규화)은 서로 동일한 값을 가져야 함.
	조건 일관성	다른 컬럼에 의해 입력 값의 범위가 제한되어야 함.
표준 준수성	도메인 준수성	컬럼은 정의된 도메인에 맞게 정의 되어야 함.
	표준 용어 준수성	테이블 및 컬럼의 명칭은 명명규칙상에 사용하도록 정의된 표준 용어를 사용해야 함.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ DQI (데이터품질지표) DQI 속성

- DQI는 측정하여 정량화 할 수 있는 것과 정량화 할 수 없는 것으로 구분됨.
- 정량적 측정 방법
 - ▶ 측정 대상 대비 (규칙 위반 또는 준수 건수)를 비율로 나타내는 방법
 - ▶ 측정 대상의 규칙 위반 여부만을 나타내는 방법으로 구분될 수 있음.

DQI Dimension	Measurable	Ratio (주로 값에 대한 품질 측정)	<ul style="list-style-type: none">• 규칙 예시 - 데이터 허용값 정확성 : 입금취소구분의 값은 NULL,0~9,A 만 올 수 있다.• 점수화 예시 - 90점 = (90건/100건) *100
		Alternative (주로 구조 정의에 대한 품질 측정)	<ul style="list-style-type: none">• 규칙 예시 - 도메인 준수성 : 대출 이자율 컬럼 정의는 이자율 도메인 정의를 준수해야 한다.• 점수화 예시 - Y (준수함)
	Non-Measurable		<ul style="list-style-type: none">• 규칙 예시 - 검색 용이성 : 데이터를 추출하여 활용할 수 있도록 검색 기능과 검색 조건이 편리하며, 출력 방식이 적절해야 한다.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 프로파일링의 정의

- 데이터의 저장 구조 및 분포 현황에 대한 분석 데이터를 제공함
- 데이터의 현황을 이해하고 데이터 규칙을 도출하는데 활용할 수 있음.
- 품질 측정대상 DB의 데이터를 읽어 데이터의 저장 구조 및 값의 분포에 대한 정보를 통계적으로 제공함.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 프로파일링의 목적

- DBMS Dictionary 정보에 들어있는 데이터 저장 구조 및 실제 발생한 데이터 값을 통계적으로 분석한 자료를 통하여 데이터의 현황 및 특성을 파악할 수 있도록 함 .
- 데이터 현황에 대한 분석 및 특성 파악을 토대로 데이터 규칙의 후보를 도출할 수 있도록 함.
- 프로파일링을 통해 정확한 메타 데이터를 판별해 내고 정확한 메타 데이터에 위배되는 값 오류, 구조 오류 등 부정확한 데이터 현황을 발견할 수 있도록 함.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 프로파일링의 유형

- 컬럼 분석 : 컬럼에 저장된 값에 대한 분포 및 값의 포맷별 분포를 분석함.

예) 계약 일자 컬럼의 값의 분포 : Null 100건, 000101 10건, 1990101 120 건.....

일자 컬럼의 포맷 분포 : 9999999 150건, 999X999 100건.....

- PK 분석 : 테이블 내에 컬럼 및 컬럼 조합의 유일성을 조사하여 Primary Key 후보를 분석함.

예) 계약 테이블의 주민번호+계약일자의 중복 데이터 분포율 2%

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 프로파일링의 유형

- 테이블간 데이터 분석 : 테이블 간에 동일 데이터 도메인을 가지는 컬럼군을 분석하여 데이터의 중복성을 분석함.

예) 계약 테이블의 고객명과 고객 테이블의 고객명의 일치율 97%

- 테이블간 FK 관계분석 : 테이블 간에 상호 Foreign Key 관계를 가지는 컬럼이 존재하는 분석함.

예) 고객 테이블의 고객번호를 기준으로 계약 테이블의 고객번호와 데이터 일치율 98%

데이터 품질 관리 프로세스 (1/2)

2) 정의

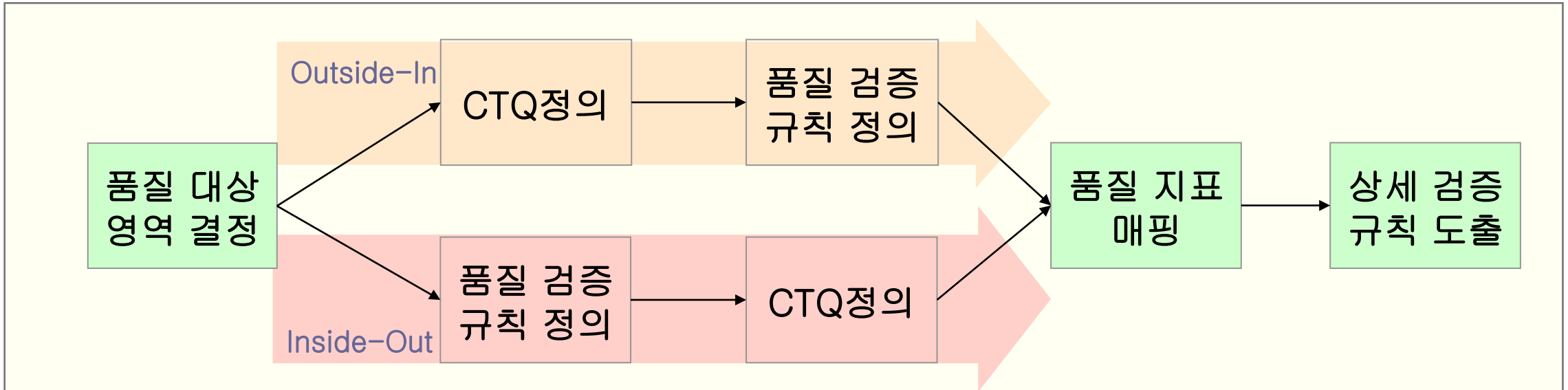
■ 품질 검증 규칙 정의

접근 방식	Inside-Out 방식	Outside-In 방식
정의	<ul style="list-style-type: none">• 데이터 분석으로부터 시작(프로파일링 결과 활용)• 데이터의 구조나 허용값, 다른 데이터와의 관계 등의 정보를 파악하기 위해 메타 데이터 활용• 추출된 부정확한 데이터로부터도 품질 검증 규칙을 도출 가능	<ul style="list-style-type: none">• 업무 관점의 이슈나 현업 담당자의 인터뷰 등을 통해 업무 수행에 핵심이 되는 요건으로부터 시작• 재작업, 반품, 고객 불만 등의 현황 및 핵심 업무 프로세스로부터 규칙 도출
장단점	<ul style="list-style-type: none">• 규칙 조사 착수 방법으로 용이하게 접근 할 수 있음.• 데이터 자체를 통해 접근하므로 잠재된 문제 발견의 기회가 있음.• 정확하지만 유효하지 않은 값은 발견할 수 없음.	<ul style="list-style-type: none">• 업무 수행에 중요한 항목으로 집중할 수 있음.• 타 조직의 사람들과 많은 인터뷰,조사 등을 필요로 하므로 높은 참여율을 필요로 함.
예시	<ul style="list-style-type: none">• 고객 레코드는 나이 필드값을 반드시 가져야 한다.• 주문 레코드의 상품코드는 반드시 전사 상품 코드에 정의된 값을 가져야 한다.• 대출 이자율 컬럼의 정의는 이자율 도메인을 준수하여야 한다.	<ul style="list-style-type: none">• 보증에 연결된 여신이 하나면 A,둘이상이면 B,없으면 C값을 가져야 한다.• 담보금액은 담보충당금보다 커야 한다.• 캠페인에 활용되는 고객 데이터는 6개월내에 최신화된 것이어야 한다.

데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 규칙 도출 경로



데이터 품질 관리 프로세스 (1/2)

2) 정의

■ 규칙 조사

[규칙 조사 Case 1] Inside-Out 방식

테이블명	컬럼명	도메인	FORMAT	DQI 상세유형	품질 검증 규칙	CHILD 컬럼	참조 테이블	참조 컬럼
AAA	-			참조무결성		고객번호	CCC	고객번호
AAA	xx번호	xx번호	Number(10)					
AAA	고객번호	고객번호	...					
AAA	대출여부	여부	...	조건 일관성	평점 >0 THEN 'Y' ELSE 'N'			
BBB	총금액	총금액	Number(17,3)	조건 일관성	총잔액보다 크거나 같아야 한다			
BBB	총수신잔액	총금액	...	계산집계 일관성	a잔액,b잔액의 합보다 크거나 같아야 한다.			

[규칙 조사 Case 2] Outside-In 방식

CTQ	품질 검증 규칙	DQI	DQI 상세유형	상세 검증 규칙
사원 정보	한 명의 관리자는 하나의 부서만 관리할 수 있음.	유일성	데이터 유일성	Dept table에서 dept_mgr_id 의 Unique 검증
	부서 관리자는 정규직 사원이어야 함.	정확성	참조 무결성	Dept table의 dept_mgr_id는 emp table의 emp_id에 존재

학습목차

3 데이터 품질 관리 프로세스를 설명

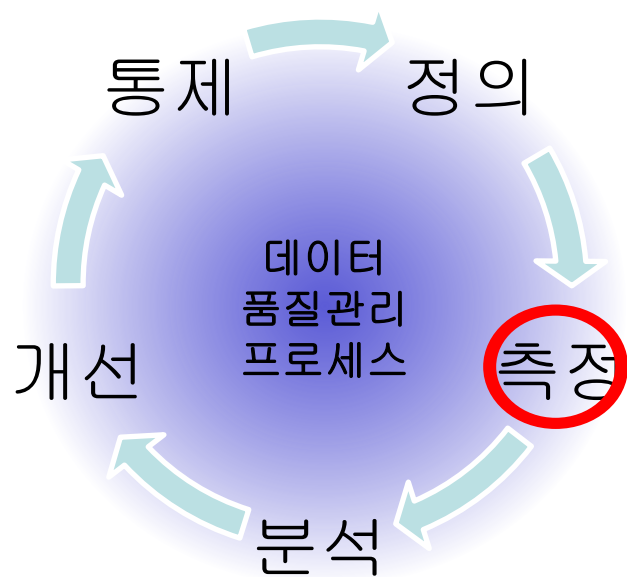
학습목표

1 데이터 품질 관리 프로세스를 설명한다.

2 연관 분야와의 관계를 설명한다.

데이터 품질 관리 프로세스 (2/2)

1) 측정



측정 환경 구성

데이터 품질 측정을 위한 시스템 환경을 구성

측정 방안 개발

품질 점검 규칙을 상용 도구 혹은 프로그램을 통해 측정 가능하도록 개발

데이터 품질 측정

정의된 품질 검증 규칙에 따라 데이터 품질 측정

데이터 품질 관리 프로세스 (2/2)

1) 측정

■ 품질 점수 산정

– 품질 측정 점수

▶ Ratio 측정 (주로 값에 대한 품질 측정)

측정 점수 = (정상 건수 / 전체 대상 건수) = 100 X [1 - (오류 건수 / 전체 대상 건수)]

※ 오류 및 전체 대상 건수에 대한 정의 : DQI 별로 정의되고, 품질 검증 규칙별로 확정됨.

▶ Alternative 측정 (주로 구조 정의에 대한 품질 측정)

측정 점수 = 규칙 준수 여부(1 또는 0)

– 분석 기준별(DQI, 시스템, 조직 등)로 품질 점수를 산정할 경우, 관련 품질 검증 규칙의 품질 점수에 대해 평균값,최소값 등의 대표 값을 적용함.

▶ 분석 기준별 대표 값 산정 시 성격이 다른 Ratio 측정과 Alternative 측정의 품질 점수는 분리하여 표시하는 것을 고려하도록 함.

데이터 품질 관리 프로세스 (2/2)

1) 측정

■ 품질 점수 산정

– 품질 측정 점수

▶ Ratio 측정 (주로 값에 대한 품질 측정)

측정 점수 = (정상 건수 / 전체 대상 건수) = 100 X [1 - (오류 건수 / 전체 대상 건수)]

※ 오류 및 전체 대상 건수에 대한 정의 : DQI 별로 정의되고, 품질 검증 규칙별로 확정됨.

▶ Alternative 측정 (주로 구조 정의에 대한 품질 측정)

측정 점수 = 규칙 준수 여부(1 또는 0)

– 분석 기준별(DQI, 시스템, 조직 등)로 품질 점수를 산정할 경우, 관련 품질 검증 규칙의 품질 점수에 대해 평균값,최소값 등의 대표 값을 적용함.

▶ 분석 기준별 대표 값 산정 시 성격이 다른 Ratio 측정과 Alternative 측정의 품질 점수는 분리하여 표시하는 것을 고려하도록 함. 예)

테이블	Ratio 점수	Alternative 점수
고객 주문	93.1	30

데이터 품질 관리 프로세스 (2/2)

1) 측정

- DQI 계량화 방안 예시 [평균 적용](Ratio)

DQI 항목 (상세 유형)		정확성			...	Column별 평균
		참조 무결성	데이터 허용값 정확성	데이터 타입 정확성	...	
AAA	고객 번호	90	80	100		90
	고객 구분	N/A	100	90		95
	...	80	70	80		77
DQI 유형별 평균		85	83	90		-

* 필요 시 가중치를 반영한 점수를 구할 수 있음

데이터 품질 관리 프로세스 (2/2)

1) 측정

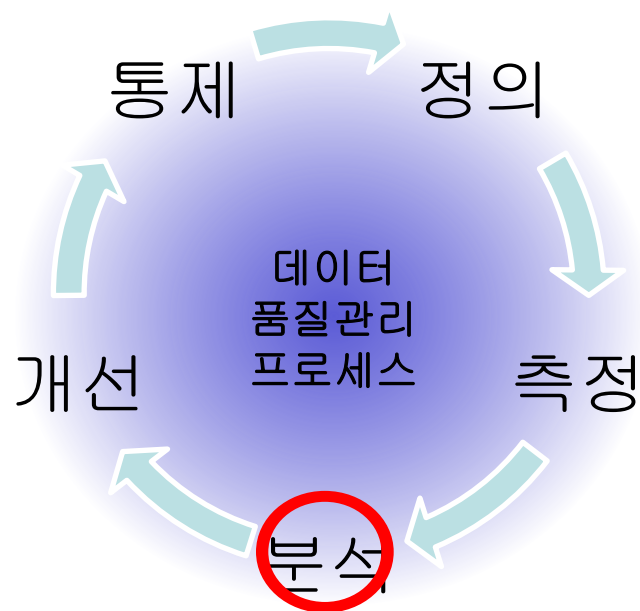
- [DQI 계량화 방안 예시 평균 적용](Alternative)

DQI 항목 (상세 유형)		구조정의 일관성			...	Column별 평균
		도메인 준수	표준 용어 준수	정의 일관성	...	
AAA	고객 번호	Y	Y	Y	...	100
	고객 구분	Y	Y	N	...	67

	AAA의 DQI 유형별 평균	100	80	50	...	90
...
DQI 유형별 평균		95	90	75	...	-

데이터 품질 관리 프로세스 (2/2)

2) 분석



품질 측정 결과 분석

분석 기준별(DQI별, 시스템별 등)로 데이터 품질 점수를 분석하여 품질 현황을 파악

데이터 오류 유형 분석

오류 데이터 분석을 통하여 오류 유형을 도출

데이터 오류 영향 분석

데이터 품질 문제로 발생하는 영향을 분석하여, 원인 분석의 우선순위를 결정

원인 분석

해당 데이터 관련자와 함께 데이터 품질 문제가 발생한 근본 원인을 분석

데이터 품질 관리 프로세스 (2/2)

2) 분석

■ 원인 분석 방안

명칭	설명	절차 및 요소
오류 연관 데이터 분석	오류를 발생시키는 원인의 범위를 구체화하기 위해 사용함.	<p>□ 오류 데이터 식별 → 분석 데이터 집합 구성 → 특성 파악 (분석 데이터 집합이 전체 데이터 분포와 비교하여 의미 있는 차이점이 없는지 확인) 예) Rule1: 고객주소는 반드시 입력되어야 함. Rule2: 고객 전화번호는 반드시 입력되어야 함. → Rule1과 2의 union 데이터 분석 결과 “내부 고객” 인 경우에는 오류 없음. → 고객 등록 프로그램 중 “내부 고객등록” 을 제외한 부분에 대해 원인 분석</p>
데이터 발생 이벤트 분석	데이터 획득 경로를 파악하고, 경로별로 데이터 오류가 발생할 수 있는 요인을 분석함.	<p>□ 데이터 생성방식 파악 (사용자입력, Batch작업에 의한 생성, 외부 기관 데이터) □ 데이터 입력경로 파악 (입력 주체나 관련 프로그램)</p>
데이터 관리 현황 파악	데이터 노후화, 데이터 값의 오류를 발생시킬 수 있는 관리요소를 파악함.	<ul style="list-style-type: none">• 데이터 노후화 관리 현황 : 주기적인 모니터링 & 클리닝 존재 여부• 데이터 운영자의 작업 : 운영 관리자에 의한 DB 접근 & 수정 가능성 확인• 데이터 구조 관리 : 데이터 구조에 대한 문서, 모델, 메타에 대한 관리 방식 (설계자/개발자가 데이터 구조를 잘못 이해하여, 컬럼 등을 다른 용도로 사용했을 가능성 확인)

데이터 품질 관리 프로세스 (2/2)

2) 분석

■ DQI별 오류현황 예

DQI	DQI 상세유형	RULE수	오류건수 합계	평균 오류율(%)
완전성	컬럼 완전성	10개	1,111,796	7.46%
	레코드 완전성	1개	36	0.01%
유일성	데이터 유일성	1개	12	0.00%
정확성	참조 무결성	5개	31,232	0.08%
	데이터 허용값 정확성	xx개	xxx	0.20%
일관성	계산집계 일관성	x개	xxx	0.02%
	데이터 일치성	x개	xxx	xxx%
	조건 일관성	X개	Xxx	xxx%
표준성	도메인 준수	x개	10	1%
	표준용어 준수	X개	30	3%
	구조정의 일관성	X개	8	0.8%
평균		xx개	xxx	1.07%

데이터 품질 관리 프로세스 (2/2)

2) 분석

■ 원인 분석 예

테이블명	영수증()		
컬럼명	취소구분()		
점검규칙	입금취소구분의 값은 NULL,0~9,A 만 올 수 있음.		
DQI 유형	정확성 / 데이터허용값 정확성		
측정결과	검증대상 전체건수	71,941,302	
	오류발생건수	15,399	
	오류율	0.021%	
	오류 데이터 상세	취소구분	건수
		-----	-----
			71,142,790
		0	430,889
		01	14,041 <- 오류
		02	72
오류 원인 유형	운영자일괄변경오류		
오류 원인 분석	자동갱신 일괄변경 배치작업처리시 입금취소구분을 두자리로 잘못 설정		
점검의견/시사점	운영자에 의한 데이터 일괄 변경은 운영팀 내에서 사전에 충분한 검증이 이루어져야 함.		

데이터 품질 관리 프로세스 (2/2)

3) 개선



개선 방안 정의

오류 데이터에 대한 개선의 방향성을 확인하고, 이에 따라 개선 방안을 정의

솔루션 정의

개선의 범위와 규모에 따라 개선 작업 수행을 위한 솔루션 정의

개선 실행 계획 수립

도출된 개선안에 대해 목표 수준을 정의하고, 일정 계획을 수립

개선 실행

일정 계획에 따른 개선 작업 수행

데이터 품질 관리 프로세스 (2/2)

3) 개선

■ 개선 방안 예

개선 방향	개선 분야	개선 방안	고려 사항
오류 데이터 정제	데이터	<ul style="list-style-type: none">• Cleansing• Correcting• Enrichment• Standardization	<ul style="list-style-type: none">• 전체 오류 데이터를 정제하기 힘든 경우, 신규 발생하는 데이터에 대해서만 적용하거나, 데이터를 사용하는 쪽(DW 등)에서 필요한 데이터만 정제할 수 있음.• 즉각적으로 수행할 수는 있으나, 근본적인 해결책이 아니므로 오류 데이터가 반복 생성될 수 있음.
오류 데이터 발생 원인 해결	시스템	<ul style="list-style-type: none">• 응용 프로그램 개선 (화면 개선, 점검로직 추가 등)• 데이터 구조 변경	<ul style="list-style-type: none">• 비교적 단순한 개선은 시스템 유지보수 업무의 일부로 포함 가능하지만, 복잡하거나 업무에 지장을 초래하는 경우에는 해당 시스템의 개선 또는 재개발 시점에서 고려될 수 있음.
	표준화	<ul style="list-style-type: none">• 코드 체계 표준화• 데이터 설계 표준 적용	<ul style="list-style-type: none">• 내부 조직간의 이해관계를 조정할 수 있는 고객 내부 담당자 또는 조직이 필요함.• 표준화 개선 사항은 데이터 수정 및 응용 프로그램 수정, 데이터 구조 변경이 뒤따르므로 단기 개선 과제로는 어려움.
	프로세스	<ul style="list-style-type: none">• 사용자 교육• 데이터 관리프로세스 개선• 업무 프로세스 개선	<ul style="list-style-type: none">• 업무 프로세스 개선의 경우, 전사 관점의 접근이 필요하므로 장기적인 전략이 필요할 수 있음.

데이터 품질 관리 프로세스 (2/2)

3) 개선

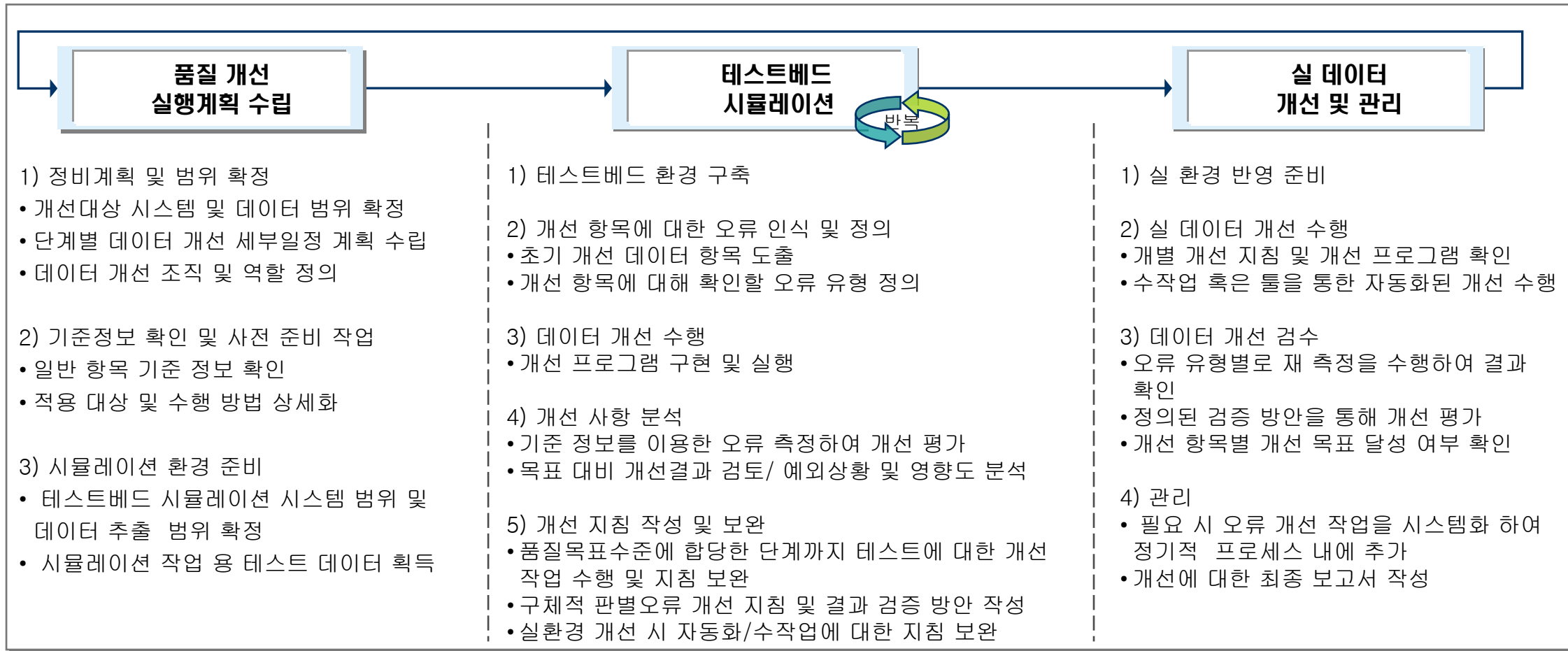
■ 개선 과제

<u>개선 분야</u>	<u>개선 과제</u>	<u>상세 내역</u>
데이터	데이터 보강 및 정비	<ul style="list-style-type: none">오류 데이터에 대한 클린징 및 누락 데이터 보강사용되지 않는 데이터 항목 삭제
시스템/표준화	메타 데이터 관리 시스템 구축	<ul style="list-style-type: none">표준 용어/도메인/데이터 사전/기준코드 정의메타성 정보 통합 관리를 위한 시스템 구축DB 오브젝트와 컬럼 및 상세 속성에 대한 정보 관리
데이터/시스템 /표준화	마스터 데이터 관리 체계 수립	<ul style="list-style-type: none">기업 비즈니스의 핵심 데이터 정의마스터 데이터에 대한 통합 및 전사 Single View 정의 및 관리
프로세스	데이터 관리 프로세스 개선	<ul style="list-style-type: none">데이터 표준/구조 관리 프로세스데이터 변경 관리 프로세스
시스템/표준화	통합 모델 관리	<ul style="list-style-type: none">표준에 따른 데이터 모델 통합통합 레포지토리를 활용한 전사 데이터 모델 공유

데이터 품질 관리 프로세스 (2/2)

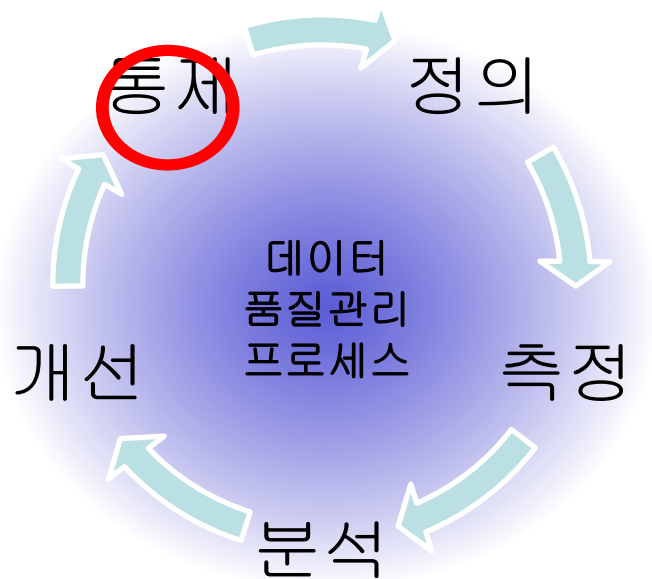
3) 개선

■ 데이터 보강 및 정비



데이터 품질 관리 프로세스 (2/2)

4) 통제



개선 결과 평가

개선 전/후 품질 점수를 비교 평가하여 그 효과를 분석

품질관리 체계 수립/관리

품질 관리의 지속성 강화를 위한 체계를 만들고 관리

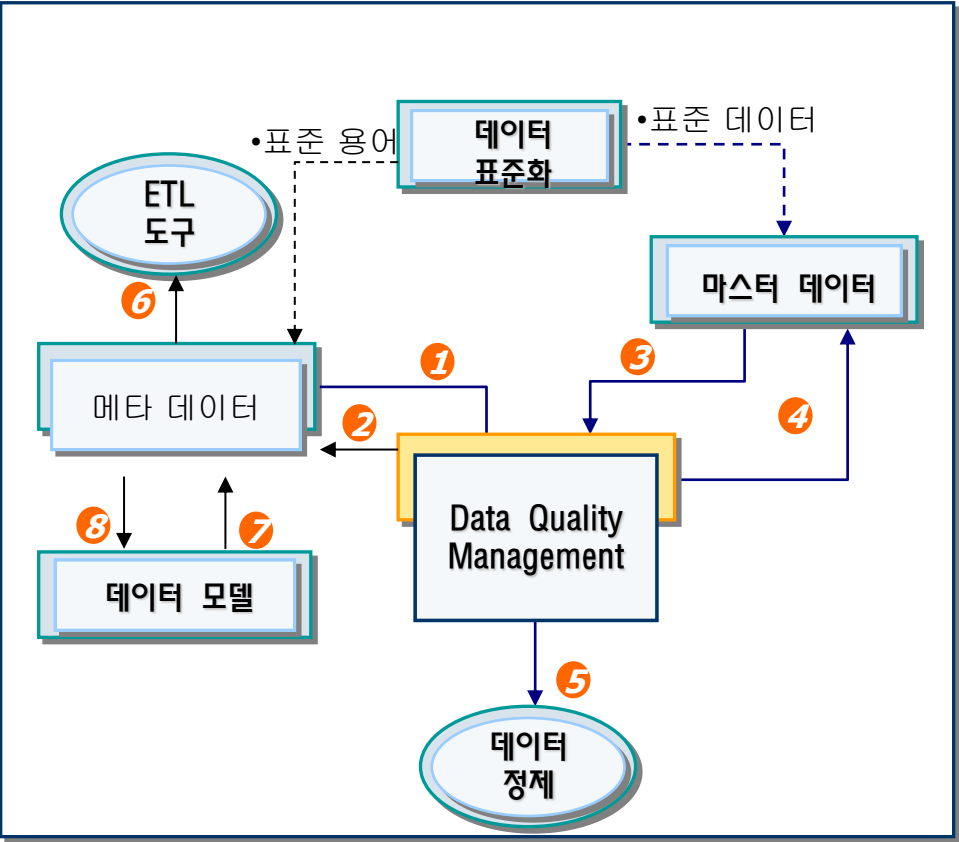
품질관리 교육 및 확산

관련 직원들의 데이터 품질 마인드 제고를 위한 교육과 정보 공유 및 확산 작업

연관 분야와의 관계

1) 데이터 인터페이스

■ 데이터 인터페이스



1	<ul style="list-style-type: none">• 품질 점검 규칙 설정을 위한 기초 정보 (테이블 구조, 컬럼 속성, 이동 주기 등)• 주요 엔터티 식별, 데이터 사용 현황 정보를 통한 오류 데이터에 대한 원인 추적 및 영향도 분석 자료
2	<ul style="list-style-type: none">• 데이터 품질 관리에서 생성된 메타 데이터 (품질 점검 규칙, DQI 지표 등)• 허용값, 제약 조건 등 정제된 메타 정보
3	<ul style="list-style-type: none">• 품질 측정 시 기준 값으로 활용
4	<ul style="list-style-type: none">• 품질 관리 대상에 대한 오류 데이터로 정제 대상 데이터 제공
5	<ul style="list-style-type: none">• 데이터 흐름 정보 활용 (매핑 정보 등)
6	<ul style="list-style-type: none">• 스키마 기본 정보, 컬럼 속성 정보
7	<ul style="list-style-type: none">• 비표준 항목 등에 대한 수정 요구 사항
8	

연관 분야와의 관계

1) 데이터 품질과 연관된 메타 데이터

- 데이터 품질과 연관된 메타 데이터
 - 타 시스템에서 관리되는 데이터는 전사의 표준으로서 품질 관리를 위한 기준 정보를 제시해주며,
 - 데이터 품질 관리를 위해 필요한 품질 지표 등도 Metadata로 저장하여 전사의 뷰를 통일함.

연관 분야와의 관계

1) 데이터 품질과 연관된 메타 데이터

- 데이터 품질과 연관된 메타 데이터

Meta 구분	사용되는 메타 데이터	데이터 품질 관리에서의 역할
표준정보	<ul style="list-style-type: none">▪ 데이터 사전: 컬럼명▪ 표준 도메인: 도메인명, 데이터 타입, 데이터 포맷▪ 기준 코드: 코드 값 명, 코드 값	<ul style="list-style-type: none">▪ 데이터 구조 정의 값의 표준 준수도 측정 시 기준 정보로 사용▪ 기준 코드성 컬럼의 ‘데이터 허용값’ 측정 시 기준 정보로 사용▪ 데이터 프로파일링 시 데이터 패턴 정보 분석을 위해 사용
Data Model 정보	<ul style="list-style-type: none">▪ 스키마 기본 정보: 시스템명, 주제영역명, 테이블명, 컬럼명, 컬럼 타입, 관계 정보▪ 컬럼 상세 속성: 허용값, 제약 조건, 도메인명▪ 변경 관리 정보: 테이블/컬럼 생성, 수정, 삭제 정보	<ul style="list-style-type: none">▪ 품질 측정 대상에 대한 품질 점검 규칙 등록 등을 위한 기초 정보로 수집하여 사용▪ 데이터 규칙 선정을 위한 기본 정보로 활용▪ 스키마 변경에 따른 기등록된 품질 점검 규칙의 영향도 파악 시 사용
Data Usage 정보	<ul style="list-style-type: none">▪ 데이터 사용 현황: 엔터티/프로세스 CRUD matrix (or 엔터티/프로그램 매트릭스)	<ul style="list-style-type: none">▪ 데이터 사용 현황 파악, CTQ 선정 시 주요 엔터티 식별▪ 오류 데이터에 대한 원인 추적 및 데이터 개선 시 응용 영향도 파악
Data Flow 정보	<ul style="list-style-type: none">▪ 데이터 흐름 정보: 시스템간 테이블/컬럼 매핑 정보▪ 데이터 작업 정보: 품질 관리 대상 데이터의 이동 주기 및 기준 정보	<ul style="list-style-type: none">▪ 오류 데이터에 대한 원인 추적 및 데이터 개선 사항 도출 시 사용▪ 품질 측정 대상 데이터의 범위 및 주기 결정을 위한 기초 정보로 활용

연관 분야와의 관계

2) 마스터 데이터

■ 마스터 데이터 관리

- 마스터 데이터: 기업 비즈니스의 핵심 요소 데이터(예:고객,계좌,공급자,제품,직원 등)로 여러 프로세스에서 자주 사용되는 공통의 데이터
- 마스터 데이터 관리: 다양한 소스로부터 생성된 핵심 데이터를 통합하고 표준화하여 적시에 제공하기 위한 프로세스,기술들의 집합
- 고객 및 제품 정보에 대한 정비 차원에서 확장되어 전사 차원의 핵심 데이터에 대한 고품질의 Single View 제공이 요구되고 있음

연관 분야와의 관계

2) 마스터 데이터

- 데이터 품질과 마스터 데이터 관리
 - 데이터 품질 측정 및 개선을 위한 기준 데이터로서의 역할
 - 전사의 핵심 데이터에 대한 고품질 관리 정책 및 실행안으로서의 역할이 기대됨.
 - 데이터 품질은 핵심 데이터에 대한 전사의 표준화된 단일 데이터를 제공하고 관리하기 위해 마스터 데이터 관리의 필수 구성 요소임.

연관 분야와의 관계

2) 마스터 데이터

■ Flow

