I.빅데이터 분석 기획

01. 빅데이터의 이해

1.1 빅데이터 개요 및 활용

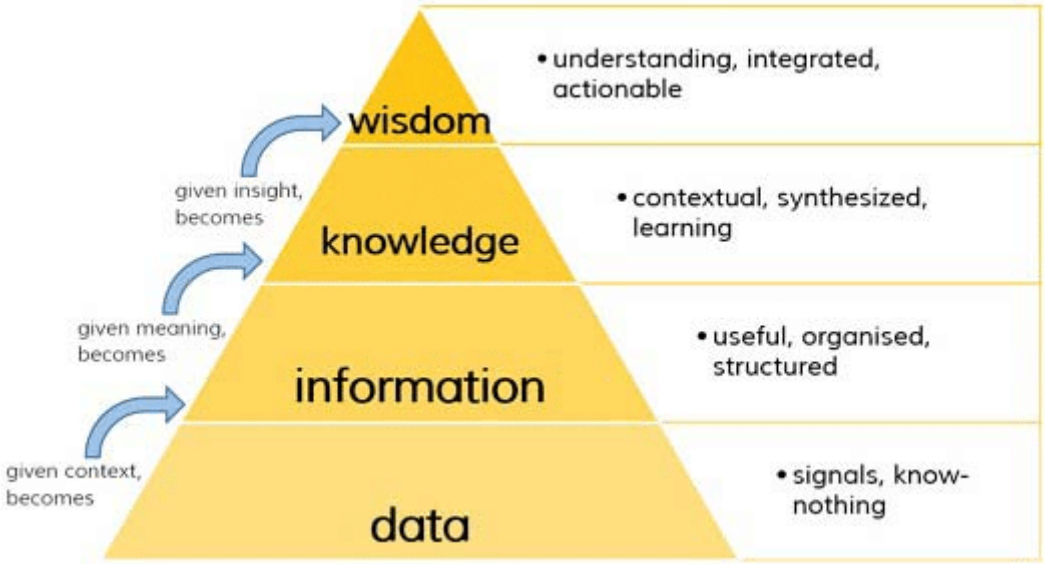
-	KeyWord
빅데이터 특징	빅데이터, DIKW피라미드, 3V, 5V, 7V, 정형, 반정형, 비정형, 암묵지, 형식지
빅데이터 가치	책임원칙
빅데이터 조직	하드스킬, 소프트스킬

==========

1) 빅데이터 특징

(1) 빅데이터 개념

- 빅데이터
  - 수십 TB의 데이터 및 데이터 분석기술
  - 데이터로부터 가치를 추출하고 결과를 분석하는 기술
- DIKW Pyramid(피라미드): 데이터->정보->지식->지혜



단계	설명
Data (데이터)	객관적 사실 / 순수한 수치나 기호
Information (정보)	데이터를 가공 및 처리 → 연관관계&의미가 도출된 데이터

단계	설명
Knowledge (지식)	정보를 구조화 → 분류&일반화시킨 결과물, 규칙
Wisdom (지혜)	근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어 상황, 맥락에 맞게 규칙을 적용하는 요소

- Byte 크기 비교(KMGT PEZY)
    - KB < MB < GB < TB < PB < EB < ZB < YB
    - 킬로 < 메가 < 기가 < 테라 < 페타 < 엑사 < 제타 < 요타
    - 테라바이트 = 10^12 바이트

(2) 빅데이터 개념

- 3V: Volume, Variety, Velocity
    - Volume (규모): 빅데이터 분석 규모
    - Variety (다양성): 자원 유형 ⇒ 정형/ 반정형/ 비정형
    - Velocity (속도): 수집/ 분석/ 활용속도 ⇒ 실시간성/ 처리속도 가속화
- 5V: 3V + Veracity, Value
    - Veracity (신뢰성): 데이터가 가지는 신뢰 및 품질
    - Value (가치): 데이터를 통해 얻을 수 있는 가치 (정확성, 시간성과 관련됨)
- 7V: 5V + Validity, Volatility
    - Validity (정확성): 데이터가 가지는 유효성 및 정확성
    - Volatility (휘발성): 데이터가 의미가 있는 기간 (장기적인 관점에서 유용한 가치를 창출해야 함)

(3) 빅데이터의 유형

- 정형
    - 스키마 구조/ 고정필드(속성)/ DBMS에 저장
    - 스키마: DB에서 자료의 구조, 표현방법, 자료 간 관계를 형식언어로 정의한 구조
    - Oracle, MS-SQL 등의 관계형 데이터베이스
- 반정형
    - 고정필드X / 메타 데이터 or 스키마 정보 포함
    - XML, HTML, JSON 등
- 비정형
    - 고정필드X / 메타 데이터X / 스키마X
    - 데이터 각각이 객체로 구분됨
    - 텍스트, 문서, 이진 파일, 이미지, 동영상 등

(4) 데이터 지식경영

- 데이터 기반 지식격영의 핵심 이슈는 암묵지와 형식지의 상호작용에 있음
- 지식구분: 암묵지, 형식지
    - 암묵지: 학습 및 경험으로 개인에게 체화되어 있음/ 겉으로 드러나지 않음/ 공유되기 어려움
    - 형식지: 문서 및 매뉴얼/ 형상화된 지식/ 전달 및 공유하기 쉬움

- 상호작용: 내면화, 공통화, 표출화, 연결화
    - 공통화: 암묵지 → 암묵지/ 다른사람과 대화 등의 상호작용(인수인계..)
    - 내면화: 형식지 → 암묵지/ 교육 등을 통해 체화(공부..)
    - 표출화: 암묵지 → 형식지/ 내재된 경험을 문서화 및 매체화(논문 퍼블리시..)
    - 연결화: 형식지가 상호결합하여 새로운 형식지 창출(후속 연구..)

==========

2) 빅데이터의 가치

(1) 빅데이터의 가치

- 경제적 자산/ 불확실성 제거/ 리스크 감소/ 스마트한 경쟁력/ 타 분야 융합

(2) 빅데이터 가치 산정이 어려운 이유

- 데이터 활용 방식의 다양화
    - 특정 데이터를 언제, 어디서, 누가 활용할 지 알 수 없음
    - 기존에 풀 수 없던 문제해결
- 새로운 가치 창출
    - 기존에 없던 가치를 창출
- 분석기술의 급속한 발전
    - 분석 비용이 저렴해지면서 활용도가 증가함

(3) 빅데이터 위기 요인 및 통제 방안

- 빅데이터 위기 요인
    - 사생활 침해: 인스타에 여행 간다고 자랑함 → 집에 강도 침입
    - 책임원칙(사용자 책임)훼손: 민주주의 국가 원리는 잠재적 위협이 아니라, 명확한 결과에 대한 책임을 물음
    - 데이터 오용: 언제나 맞을 수는 없다는 오류, 잘못된 지표를 사용하는 오용
- 위기 요인에 대한 통제 방안
    - 알고리즘에 대한 접근 허용: 알고리즘을 통해 불이익 당한 사람들을 위해 "알고리즈미스트"라는 전문가 필요
    - 책임의 강조: 개인정보를 사용하는 "사용자"의 책임을 강조
    - 결과 기반의 책임 적용

==========

3) 빅데이터 산업의 이해

(1) 빅데이터 산업 개요

- 클라우드 컴퓨팅 기술의 발전 → 데이터 처리 비용 급감 → 빅데이터 발전
    - 클라우드 컴퓨팅: 인터넷을 통해 다수의 사용자들에게 가상화된 컴퓨터의 시스템 리소스를 요구하는 즉시 '서비스'로 제공하는 컴퓨팅 기술
- 주요국, 글로벌 기업: 산업 육성 및 "활용"에 주력
- 우리나라: 데이터 생산량은 많음/ "활용"은 저조

(2) 산업별 빅데이터 활용

- 의료 및 건강/ 과학기술/ 정보보안/ 제조 및 공정/ 소비 및 거래/ 교통 및 물류 등

==========

4) 빅데이터 조직 및 인력

(1) 빅데이터 조직 설계

- 빅데이터 업무 프로세스
  - 빅데이터 도입 -> 구축 -> 운영
- 조직 구조 설계 요소
  - 업무 활동/ 부서화/ 보고 체계
  - 수직 업무 활동: 우선순위 결정
  - 수평 업무 활동: 업무 프로세스 절차별로 배분
- 조직 구조 유형
  - 집중 구조(별도): 전사의 분석 업무를 별도 조직에서 담당
  - 기능 구조(각자): 해당 부서에서 각자 분석 수행(전사적인 분석 어려움)
  - 분산 구조(배치): 분석 조직 인력들을 현업 부서로 배치(업무 과다, 베스트 프랙티스 공유 가능)
- 조직 구조의 설계 특성
  - 공식화(기준설정) / 분업화 / 직무 전문화 / 통제 범위(인원수) / 의사소통 및 조정

(2) 조직 역량

- 지속적인 경영과 성과 달성을 위해 중요한 요소
- 역량 모델링
  - 목표 달성을 위해 우수 성과자의 기여가 중요함
  - 직무별 역량 모델: 우수 성과자의 직무 역량 요소들을 도출하여 만듦
- 데이터 사이언티스트의 요구역량
  - 하드 스킬(Hard skill): 이론적 지식(기법, 방법론 습득) / 분석기술의 숙련도(노하우)
  - 소프트 스킬(Soft skill): 통찰력(논리적 비판, 호기심 등) / 협력(커뮤니케이션) / 전달력(스토리텔링, 비주얼라이제이션)
  - 가트너(Gartner): 분석 모델링, 데이터 관리, 소프트 스킬, 비즈니스 분석
- 데이터 사이언티스트
  - 복잡한 비즈니스 문제를 모델링, 인사이트를 도출하여 통계학, 알고리즘, 데이터 마이닝, 시각화 기법 등을 통해 가치를 찾아내는 사람
- 역량 모델 개발 절차
  - 조직의 미션,성과목표,핵심성공요인 검토 → 조직 구성원의 행동특성 도출 → 역량 도출 → 역량 모델 확정
- 역량 교육 체계 설계 절차
  - 요구사항 분석 → 직무별 역량모델 검토 → 역량 차이 분석 → 직무 역량 매트릭스 → 교육 체계 설계

(3) 조직성과 평가

- 개인성과 관리가 중요 → 목표설정 위한 핵심성공요인(CSF), 목표달성 위한 핵심성과지표(KPI) 정의
- 조직성과 평가 절차

    - 목표 설정 → 모니터링 → 목표 조정 → 평가 실시 → 결과의 피드백
- 균형 성과표(BSC; Balanced Score Card) 4가지 관점
- 재무/ 고객/ 내부 프로세스/ 학습 및 성장

1.2 빅데이터 기술 및 제도

-	**KeyWord**
빅데이터 플랫폼	빅데이터 플랫폼, 하둡 에코시스템, R, 우지, 플럼, HBase, 스쿱, 맵리듀스, 얀, 스파크, HDFS, 척와, 스크라이브, 히호, 피그, 하이브, 머하웃, 임팔라, 주키퍼
인공지능의 개념	인공지능, 빅데이터
개인정보보호법·제도	개인정보보호
개인정보 활용	개인정보 비식별화

==========

1) 빅데이터 플랫폼

(1) 빅데이터 플랫폼의 개념

- 빅데이터에서 가치를 추출하기 위해 일련의 과정을 규격화한 기술
- 일련의 과정은: 수집 → 저장 → 처리 → 분석 → 시각화
- 의료, 환경, 범죄, 자동차 등 특화된 분석을 지원하는 플랫폼이 발전 추세

(2) 빅데이터 플랫폼 구성요소

- 데이터 수집 → 저장 → 분석 → 활용
- 수집: ETL(Extract Transform Load), 크롤러(Crawler), EAI(Enterprise Architecture Integration) 등
- 저장: RDBMS(Relational DBMS, 관계형 데이터베이스), NoSQL(Not Only SQL) 등
    - NoSQL: 전통적인 RDBMS와 다른 DBMS를 지칭하기 위한 용어(고정된 테이블 스키마 X 조인 연산 X 수평적 확장 O)
- 분석: 텍스트 마이닝, 머신러닝, 통계, 데이터 마이닝, SNS 분석, 예측 분석 등
- 활용: 데이터 가시화, 비즈니스 인텔리전스(BI), Open API 연계, 히스토그램, 인포그래픽 등

(3) 빅데이터 플랫폼 데이터 형식

- HTML: 웹페이지 만들 때 사용/ 텍스트, 태그, 스크립트로 구성
- XML: 다목적 마크업 언어/ 데이터 표현을 위해 태그 사용
- CSV: 필드를 쉼표로 구분한 텍스트 데이터, 텍스트 파일
- JSON: Key-Value로 이루어진 데이터 오브젝트를 전달하기 위해, 텍스트를 사용하는 개방형 표준 포맷

(4) 빅데이터 플랫폼 구축 소프트웨어

- R: 빅데이터 분석
  - S언어를 기반으로 만들어짐/ 강력한 시각화 기능
  - r-project.org: R is a free software environment for statistical computing and graphics.
- 우지(Oozie): 워크플로우 관리
  - 하둡 작업(job) 관리/ 워크플로우 및 코디네이터 시스템/ 스케줄링 및 모니터링
  - oozie.apache.org: Oozie is a workflow scheduler system to manage Apache Hadoop jobs.
- 플럼(Flume): 데이터 수집
  - Event, Agent 활용/ 대량 로그데이터를 수집, 집계, 이동
  - 여러 서버에서 생산된 대용량 로그 데이터를 수집하여 원격 목적지에 데이터를 전송하는 기능
  - flume.apache.org: service for collecting, aggregating, and moving large amounts of log data.
- HBase: 분산 데이터베이스
  - 컬럼 기반 저장소/ HDFS, 인터페이스 제공
  - 큰 테이블에 대한 빠른 조회 가능/ HDFS 위에 구축되어, HDFS에 있는 데이터에 랜덤 엑세스 및 읽기
  - hbase.apache.org: Hadoop database. Random, realtime read/write access to bigdata.
- 스쿱(Sqoop): 정형 데이터 수집
  - SQL to Hadoop/ SQL ↔ HDFS/ Connector를 사용
  - 동작 2가지 import(SQL → HDFS), export(HDFS → SQL)
  - sqoop.apache.org: tool designed for transferring bulk data between Hadoop and structured datastores.

(5) 분산 컴퓨팅 환경 소프트웨어 구성요소

- 맵리듀스(Map Reduce)
  - 맵 → 셔플 → 리듀스 순서대로 데이터 처리
  - 맵: Key-Value로 데이터 취합(입력된 데이터를 가공하여 Key-Value 쌍으로 변환)
  - 셔플: 데이터 통합 처리
  - 리듀스: 맵 처리된 데이터 정리 (Key를 기준으로 결과물을 모아서 집계)- 대용량 데이터를 위한 분산 병렬 처리 소프트웨어 프레임워크
- 얀(YARN)
  - 자원 관리 플랫폼
  - Master(리소스매니저)-Slave(노드매니저)
  - 리소스 매니저: 스케줄러 / 클러스터 이용률 최적화 수행
  - 노드 매니저: 노드 내 자원 관리/ 리소스 매니저에 보고
  - 애플리케이션 마스터: 자원 교섭/ 컨테이너 실행
  - 컨테이너: 프로그램 구동을 위한 격리 환경 지원
- 아파치 스파크(Apache Spark) -대규모 데이터 분산처리시스템
  - 실시간 데이터 처리(스트리밍 데이터, 온라인 머신러닝 등)
  - 저장이 아니라 데이터 프로세싱하는 역할
- 하둡 분산 파일 시스템(HDFS)
  - 대용량 파일을 분산된 서버에 저장, 처리
  - Master(네임노드)-Slave(데이터노드)
  - 네임 노드: 속성 기록 (파일 이름, 권한 등) / 메타 데이터 관리 / 데이터 노드 모니터링
  - 데이터 노드: 데이터 저장 / 일정한 크기로 나눈 블럭 형태로 저장함
- 아파치 하둡(Apache Hadoop)
  - HDFS, 맵리듀스를 중심으로 하둡 에코시스템을 가짐
  - 클라우드 플랫폼 상에서 클러스터를 구성하여 데이터 분석

(6) 하둡 에코시스템 (Hadoop Ecosystem)

- 수집, 저장, 처리 기술
  - 비정형 데이터 수집: 척와/ 플럼/ 스크라이브
    - 척와 (Chukwa): 분산된 서버에서 에이전트 실행 → 컬렉터가 데이터 받아서 HDFS 저장
    - 플럼 (Flume): 대량 로그데이터 수집, 집계, 이동 / 이벤트, 에이전트를 활용하는 기술
    - 스크라이브 (Scribe): 대용량 실시간 스트리밍 로그 데이터 수집 기술
  - 정형 데이터 수집: 스쿱/ 히호
    - 스쿱 (Sqoop): 대용량 데이터 전송 솔루션 / 커넥터를 사용하여 RDBMS ↔ HDFS
    - 히호 (Hiho): 대용량 데이터 전송 솔루션 / 깃허브에 공개되어 있음
  - 분산 데이터 저장: HDFS
    - 대용량 파일을 분산된 서버에 저장, 저장된 데이터를 빠르게 처리할 수 있게 하는 시스템
    - 범용 하드웨어, 서버 기반 / 데이터 접근 패턴을 스트리밍 방식으로 지원/ 자동복구
  - 분산 데이터 처리: 맵리듀스
  - 분산 데이터베이스: HBase
- 데이터 가공, 분석, 관리를 위한 주요 기술
  - 데이터 가공: 피그 / 하이브
    - 피그 (Pig): 대용량 데이터 집합을 분석하기 위한 플랫폼 / 맵리듀스 API 매우 단순화 / SQL과 유사한 형태
    - 하이브 (Hive): 하둡 기반 DW 솔루션 / SQL과 유사한 HiveQL 쿼리 제공
  - 데이터마이닝: 머하웃(Mahout)
    - 하둡 기반 데이터 마이닝 알고리즘을 구현한 오픈 소스 (분류, 클러스터링, 추천 및 협업 필터링 등)
    - 확장성을 가진 머신러닝용 라이브러리 (mahout.apache.org)
  - 실시간 SQL 질의: 임팔라(Impala)
    - 하둡 기반 실시간 SQL 질의 시스템/ 인터페이스로 HiveQL 사용/ 수초 내에 결과 확인 가능
    - 오픈소스 대규모 병렬 처리 SQL 쿼리 엔진(impala.apache.org)
  - 워크플로우 관리: 우지(Oozie)
    - 하둡 잡 관리용 워크플로우 및 코디네이터 시스템 / 자바 웹 애플리케이션 서버
  - 분산 코디네이션: 주키퍼(Zookeeper)
    - 분산 환경에서 서버 간 상호조정이 필요한 다양한 서비스를 제공하는 시스템
    - 한 서버에만 서비스가 분산되지 않도록 분산, 한 서버에서 처리한 결과를 다른 서버들과 동기화 (zookeeper.apache.org)
- 데이터 웨어하우스(DW; Data Warehouse)
  - 데이터를 공통 형식으로 변환하여 관리하는 데이터베이스 - 사용자 의사결정에 도움을 주기 위해, 기간시스템의 DB에 축적된 데이터를 효율적으로 분석 가능한 형태로 변환해놓은 저장소

빅데이터 플랫폼 요약

빅데이터 플랫폼 구축 소프트웨어		분산컴퓨팅 환경 소프트웨어 구성요소		하둡 에코시스템			
				수집/ 저장/ 처리		가공/ 분석/ 관리	
R	분석 & 시각화	맵리듀스	맵→셔플→리듀스 (맵: Key-Value)	비정형 데이터 수집	- 척와 - 플럼 - 스크라이브	가공	- 피그 - 하이브
우지	워크플로우 코디네이터	얀	자원관리 플랫폼 (리소스, 노드매니저)	정형 데이터 수집	- 스쿱 - 히호	데이터 마이닝	머하웃
플럼	대량 로그 수집 (이벤트, 에이전트)	아파치 스파크	대규모 분산처리 실시간 프로세싱	분산 데이터 저장	HDFS	실시간 SQL질의	임팔라
HBase	분산 DB HDFS, 인터페이스	HDFS	대용량 분산저장 (네임, 데이터노드)	분산 데이터 처리	맵리듀스	워크플로우 관리	우지
스쿱	SQL to Hadoop (커넥터)	아파치 하둡	클라우드 플랫폼 (HDFS, 맵리듀스)	분산 데이터베이스	HBase	분산 코디네이션	주키퍼

==========

2) 인공지능의 개념

(1) 인공지능의 개념

- 인공지능
  - 인간의 지적능력을 인공적으로 구현하여 컴퓨터가 인간의 지능적인 행동, 사고를 모방할 수 있도록 하는 소프트웨어
- 인공지능 ⊃ 머신러닝 ⊃ 딥러닝

(2) 빅데이터와 인공지능의 관계 ⇒ 상호보완 관계

- (인공지능의 분석력, 예측력) + (빅데이터의 신뢰성, 현실성) ⇒ 의미있는 결과 도출
- 빅데이터로 말미암아 비정형 데이터 고속 분석이 가능해짐
  -> 1950년대에 등장한 인공지능을 최신트렌드로 끌고 올 수 있게 됨
  -> 자체 알고리즘으로 스스로 학습하는 딥러닝 기술
  -> 특정 분야에서 인간의 지능을 뛰어넘는 능력
- 상호보완 관계
  - 빅데이터는 인공지능의 구현완성도 높여줌
  - 인공지능은 빅데이터의 문제해결 완성도를 높여줌
  - 빅데이터 기술이 주목받는 이유
    - 우수한 정보처리를 바탕으로 의미있는 결과를 도출
  - 빅데이터와 인공지능의 목표가 부합
    -> 빅데이터는 인공지능을 위한 기술이 될 가능성이 큼

==========

3) 개인정보보호법·제도

(1) 개인정보보호

- 개인정보보호: 정보 주체(개인)의 개인정보 자기 결정권을 철저히 보장하는 활동

- 개인정보 자기 결정권: 자신에 관한 정보가 언제, 어떻게, 어느 범위까지 타인에게 전달 및 이용될 수 있는지를 그 정보 주체가 스스로 결정할 수 있는 권리
- 개인정보: 살아있는 개인에 관한 정보/ 개인을 알아볼 수 있는 정보

(2) 개인정보보호의 필요성

- 개인정보는 정보사회의 핵심 인프라
- 유출 시 피해 심각함 / 정보사회 핵심 인프라 / 개인정보 자기 통제권

(3) 빅데이터 개인정보보호 가이드라인

- 개인정보 비식별화 / 재식별 시 즉시 조치 / 민감정보 처리 금지
- 처리 사실, 목적 등을 공개해 투명성 확보 / 수집정보의 보호조치

(4) 개인정보보호 관련 법령

- 개인정보 보호법 / 정보통신망법 / 신용정보법 -> "개·망·신"
- 위치정보법 / 개인정보의 안전성 확보조치 기준

(5) 개인정보보호 내규

- 데이터 수집 시 개인정보보호를 위한 가이드라인
- 정보보호 업무처리 지침 / 개발 보안 가이드 / 개인정보 암호화 매뉴얼 소프트웨어 개발 보안 구조 / 기술적, 관리적 보호

==========

4) 개인정보 활용

(1) 개인정보 비식별화

- 개인정보 일부/전부를 삭제/대체
  ->다른 정보와 쉽게 결합해도 특정 개인을 식별할 수 없도록 하는 조치
- 데이터값 삭제 / 가명처리 / 총계처리 / 범주화 / 데이터 마스킹 등

(2) 개인정보 비식별화 절차

- 사전검토 → 비식별 조치 → 적정성 평가 → 사후관리

(3) 개인정보 비식별 조치 방법

- 가명처리: 식별할 수 없는 다른 값으로 대체
  - 휴리스틱 익명화 / 암호화 / 교환방법
- 총계처리: 통곗값 적용
  - 총계처리 기본방식 / 부분집계 / 라운딩 / 데이터 재배열
- 데이터 삭제: 특정 데이터값 삭제
  - 속성값 삭제 / 부분 삭제 / 준 식별자 제거를 통한 단순 익명화
- 데이터 범주화

    - 범주화(해당 그룹의 대표값으로 변환)
    - 범주화 기본방식/ 랜덤 올림/ 범위 방법/ 세분 정보 제한 방법/ 제어 올림
    - 범위화(구간 값으로 변환)
- 데이터 마스킹: 전체 또는 부분적으로 대체값(공백, *, 노이즈)으로 변환
    - 임의 잡음 추가 / 공백과 대체 방법

(4) 재식별 가능성 모니터링

- 정기적으로 모니터링
  -> 점검 항목 중 어느 하나에 해당하면 추가적인 비식별 조치 강구
- 내부 요인의 변화: 추가적인 정보 수집 / 이용과정에서 생긴 새로운 정보 / 신규 or 추가 구축되는 시스템 등
- 외부 환경의 변화: 새로운 재식별 사례, 기술, 연계가능한 정보가 출현하거나 공개된 것으로 알려진 경우