1과목.빅데이터 분석 기획

(Ch_03. 데이터 수집 및 저장 계획 - SEC 01. 데이터 수집 및 전환
SEC 02. 데이터 적재 및 저장)

빅데이터 분석 기사(1과목. 빅데이터 분석 기획)

CHAPTER 1. 빅데이터의 이해

CHAPTER 2. 데이터 분석 계획

CHAPTER 3. 데이터 수집 및 저장 계획

데이터 수집 및 저장 계획

데이터 분석 계획은 총 2개의 작은 섹션으로 구성된다.

1. 데이터 수집 및 전환
2. 데이터 적재 및 저장

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

01 데이터 수집

1) 데이터 수집 과정

수집 데이터 도출 → 수집 데이터 목록화 → 데이터 소유 기관 확인 및 협의 → 데이터 유형 확인 및 분류 → 데이터 수집 기술 선정 → 수집 계획서 작성 → 수집 주기 정의 → 데이터 수집

① 수집 데이터 도출 및 목록화 : 수집할 데이터를 찾고, 이를 목록화한다.

② 데이터 소유 기관 확인 및 협의 : 데이터를 소유하고 있는 기관과 데이터 사용에 대한 업무 협의를 진행한다.

③ 데이터 유형 확인 및 분류 : 데이터를 유형별로 확인하여 분류한다.

④ 데이터 수집 기술 선정 : 데이터의 특징(내부 및 외부 데이터, 정형, 반정형, 비정형 데이터 등)을 파악하여 수집 기술을 선정한다.

⑤ 수집 계획서 작성 및 수집 주기 정의: 수집 대상 데이터 범위 및 수집 기술을 선정한 뒤, 데이터를 수집하기 위한 세부 계획을 수립하고, 수집 주기를 정의한다.

⑥ 데이터 수집 : 선정된 데이터 범위와 데이터 수집 계획을 기반으로 목적 데이터를 수집한다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 데이터 수집 기술

① ETL(Extract Transform Load)

▶ 분석을 위한 데이터를 데이터 저장소인 DW(Data Warehouse) 및 DM(Data Mart)으로 이동시키기 위해 다양한 소스 시스템으로부터 필요한 원본 데이터를 추출(Extract)하고, 변환(Transform)하여 적재(Load)하는 작업 및 기술이다.

② FTP(File Transfer Protocol)

▶ TCP/IP 프로토콜을 기반으로 서버, 클라이언트 사이에서 파일을 송,수신하기 위한 프로토콜이다.

③ 스쿱(Sqoop)

▶ 커넥터를 사용하여 MySQL 또는 Oracle, 메인 프레임과 같은 관계형 데이터베이스 시스템(RDBMS)에서 하둡 파일 시스템(HDFS)으로 데이터를 수집하거나 하둡 파일 시스템에서 관계형 데이터베이스로 데이터를 보낼 수 있는 기술이다.

④ 스크래피(Scrapy)

▶ 파이썬 언어 기반의 비정형 데이터 수집 기술로 웹 데이터를 수집하는 것을 목표로 설계되었다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 데이터 수집 기술

⑤ 아파치 카프카(Apache Kafka)

▶ 대용량 실시간 로그 처리를 위한 분산 스트리밍 플랫폼이다.

⑥ 플럼(Flume)

▶ 많은 양의 로그 데이터를 효율적으로 수집, 집계, 이용하기 위해 이벤트(Event)와 에이전트(Agent)를 활용하는 기술이다.

⑦ 스크라이브(Scribe)

▶ 다수의 서버로부터 실시간으로 스트리밍 되는 로그 데이터를 수집하여 분산 시스템에 데이터를 저장하는 대용량 실시간 로그 수집 기술이다.

⑧ 척와(Chukwa)

▶ 분산된 각 서버에서 에이전트를 실행하고, 컬렉터가 에이전트로부터 데이터를 수집하여 하둡 파일 시스템에 저장하고, 저장된 데이터에 대한 실시간 분석 기능을 제공하는 기술이다.

이벤트(Event) : 에이전트에 의해 옮겨지는 데이터의 기본 단위
에이전트(Agent) : 자바 가상 머신(JVM) 프로세스로 소스, 채널, 싱크로 구성되며 외부에서 전달된 이벤트를 다른 목적지로 이동하는 역할을 담당

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 데이터 수집 기술

⑨ CEP(Complex Event Processing)

▶ 직역하면 "복잡한 이벤트 처리"로서 여러 이벤트를 저장 전에 지속적으로 처리하여 미리 정의된 규칙에 따라 유의미한 이벤트를 식별해낼 수 있는 기술이다.

⑩ 크롤링(Crawling)

▶ 인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠 수집 기술이다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

개념 체크

1. 다음 중 ETL에 속하지 않는 작업은?

① Extract ② Transform

③ Trade ④ Load

ETL은 대표적인 데이터 수집 기술 중 하나로 Extract(추출), Transform(변환), Load(적재)의 약자이다.

2. 데이터 수집 기술 중 하나로 파이썬 언어 기반의 비정형 데이터 수집 기술로 웹 데이터를 수집하는 것을 목표로 설계된 기술은?

① CEP ② Apache Kafka

③ FTP ④ Scrapy

파이썬 언어 기반의 비정형 데이터 수집 기술로 웹 데이터를 수집하는 것을 목표로 설계된 데이터 수집 기술은 스크래피(Scrapy)이다.

3. 인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠를 수집하는 기술은?

① Chukwa ② Crawling

③ Sqoop ④ TCP/IP

인터넷 상에서 제공되는 다양한 웹 사이트로부터 소셜 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠를 수집하는 기술은 크롤링(Crawling)이다.

4. 다음 중 데이터 수집 기술이 아닌 것은?

① FTP ② Scrapy

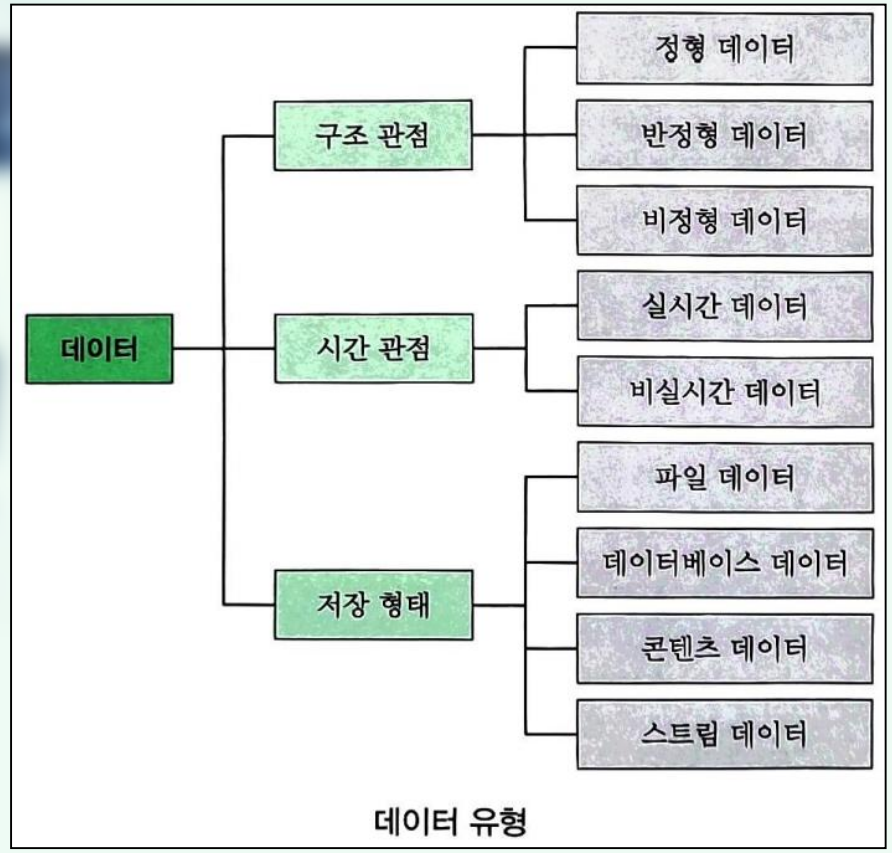
③ Flume ④ DW

데이터 수집 기술에는 ETL, FTP, 스쿱, 스크래피, 플럼, 스크라이브, CEP, Apache Kafka, 크롤링, 척와 등이 있다. DW는 데이터 저장 기술 중 하나인 DataWarehouse의 약자이다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

02 데이터 유형 및 속성 파악

; 데이터 유형은 구조적 관점에서는 정형, 반정형, 비정형 데이터로 나누고, 시간적 관점으로는 실시간, 비실시간 데이터로 나뉜다. 또한, 저장 형태의 기준으로는 파일 데이터, 데이터베이스 데이터, 콘텐츠 데이터, 스트림 데이터로 나뉜다.



데이터 유형

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

02 데이터 유형 및 속성 파악

1) 구조 관점의 데이터 유형

명칭	설명
정형 데이터	정형화된 스키마(schema, 형태) 구조 기반의 형태를 갖고, 고정된 필드에 저장되며 형식의 일관성을 갖는 데이터로 컬럼(column, 열)과 로우(row, 행) 구조를 가진다. 예) 관계형 데이터베이스(RDBMS), 스프레드시트
반정형 데이터	스키마(schema, 형태) 구조 형태를 갖고 메타 데이터를 포함하며, 값과 형식에서 일관성을 갖지 않는 데이터이다. 예) XML, HTML, 웹 로그, 알람, 시스템 로그, JSON, RSS, 센서 데이터
비정형 데이터	스키마(schema, 형태) 구조를 가지지 않고, 고정된 필드에 저장되지 않은 데이터 예) SNS, 웹 게시판, 텍스트, 이미지, 오디오, 비디오 데이터

XML(eXtensible Markup Language) : W3C에서 개발된, 다른 특수한 목적을 갖는 마크업 언어를 만드는데 사용하도록 권장하는 다목적 마크업 언어이다
Markup Language : 문서가 화면에 표시되는 형식을 나타내거나 데이터의 논리적인 구조를 명시하기 위한 규칙들을 정의한 언어 의 일종이다.
JSON (JavaScript Object Notation) : 데이터를 저장하거나 전송할 때 많이 사용되는 경량의 DATA 교환 형식을 말한다.
RSS(Really Simple Syndication) : 같은 "사이트 피드"란, 새 기사들의 제목만, 또는 새 기사들 전체를 뽑아서 하나의 파일로 만들어 놓은 것이다.
센서 데이터 : 사용자가 차세대 컴퓨터를 손에 쥔 상태로 모션을 취할 때, 차세대 컴퓨터에 내장된 센서에 의해 측정된 데이터 의 집합이다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 시간 관점에서의 데이터 유형

명칭	설명
실시간 데이터	생성된 이후 수초 ~ 수분 이내에 처리되어야 유의미한 데이터 예) 센서 데이터, 시스템 로그, 네트워크 장비 로그, 알람 등
비실시간 데이터	생성된 데이터가 수 시간 또는 수 주 이후에 처리되어야 유의미한 결과를 얻을 수 있는 데이터 예) 통계, 웹 로그, 구매 정보, 서비스 로그, 디지털 헬스 케어 정보

3) 저장 형태 관점에서의 데이터 유형

명칭	설명
파일 데이터	시스템 로그, 스프레드시트 등 파일 형식으로 저장되는 데이터
데이터베이스 데이터	관계형 데이터베이스, NoSQL 등에 의해 데이터베이스 테이블에 저장된 데이터
콘텐츠 데이터	텍스트, 이미지, 오디오, 비디오 등과 같이 개별적으로 데이터 객체로 구분되는 미디어 데이터
스트림 데이터	센서 데이터, HTTP 트랜잭션(Transaction) 등과 같이 네트워크를 통해 실시간으로 전송되는 데이터

NoSQL : 단어 뜻 그 자체를 따지자면 "Not only SQL"로, SQL만을 사용하지 않는 데이터베이스 관리 시스템(DBMS)을 지칭하는 단어이다. 관계형 데이터베이스를 사용하지 않는다는 의미가 아닌, 여러 유형의 데이터베이스를 사용하는 것이다.
트랜잭션(transaction) : 데이터베이스의 상태를 변화시키기 위해 수행하는 작업의 단위

1. 데이터 분석 계획 – 분석 방안 수립

개념 체크

1. 다음 중 구조 관점의 데이터 유형이 아닌 것은?

① 정형 데이터 ② 마이 데이터

③ 반정형 데이터 ④ 비정형 데이터

구조 관점의 데이터 유형은 정형, 반정형, 비정형 데이터이다.

2. 데이터 유형 중 저장 형태의 관점으로 분류되는 데이터 유형이 아닌 것은?

① 스트림 데이터 ② 콘텐츠 데이터

③ 파일 데이터 ④ 실시간 데이터

데이터 유형 중 저장 형태의 관점으로 분류되는 데이터 유형은 파일 데이터, 데이터베이스 데이터, 콘텐츠 데이터, 스트림 데이터이다.

3. 다음 중 생성된 데이터가 수 시간 또는 수 주 이후에 처리되어야 유의미한 결과를 얻을 수 있는 데이터는?

① 실시간 데이터 ② 비실시간 데이터

③ 파일 데이터 ④ 공유 데이터

생성된 데이터가 수 시간 또는 수 주 이후에 처리되어야만 유의미한 결과를 얻을 수 있는 데이터는 비실시간 데이터이다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

03 데이터 변환

- 데이터 변환은 데이터를 특정 규칙에 맞게 바꾸어 주는 작업을 의미한다.
- 대표적인 데이터 변환 기술에는 평활화, 집계, 일반화, 정규화, 속성 생성이 있다.
- 데이터 변환 기술

명칭	설명
평활화(smoothing)	데이터의 노이즈를 구간과 군집화 등으로 다듬는 기법
집계(Aggregation)	다양한 차원으로 데이터를 요약하는 기법
일반화(Generalization)	특정 구간으로 값을 스케일링 하는 기법
정규화(Normalization)	데이터를 정해진 구간(0~1)으로 전환하는 기법
속성 생성 (Feature Construction)	여러 데이터를 대표할 수 있는 새로운 속성값을 생성하는 기법

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

04 데이터 비식별화

- 데이터 비식별화는 개인을 특정할 수 없도록 개인정보의 일부 혹은 전체를 변환하는 기술을 의미한다.
- 데이터 비식별화 기술을 통해 민감 데이터 활용에 대한 보안성이 향상될 수 있다.

1) 개인정보 비식별 조치 가이드라인

; 2016년 7월 행정자치부와 관계부처의 합동으로 발표한 개인정보보호를 위한 개인정보 비식별 조치에 대한 가이드라인이다. 이는 데이터 처리 과정에서 개인을 특정할 수 없도록 하는 관리 지침이다.

① 추진 배경

- ▶ 정부 3.0 및 빅데이터 활용 확산에 따른 데이터 활용가치 증대
- ▶ 개인정보보호 강화에 대한 사회적 요구 지속
- ▶ '보호와 활용'을 동시에 모색하는 세계적 정책변화에 적극 대응

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

1) 개인정보 비식별 조치 가이드라인

② 단계

사전 검토 -> 비식별 조치 -> 적정성 평가 -> 사후관리

▶ 사전 검토 : 개인정보보호에 해당하는지 여부를 검토 후, 개인정보가 아닌 것이 명백한 경우 법적 규제 없이 자유롭게 활용한다.

▶ 비식별 조치 : 정보 집합물(데이터셋)에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용, 개인을 알아볼 수 없도록 하는 조치이다.

▶ 적정성 평가 : 다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지를 비식별조치 적정성 평가단을 통해 평가한다.

▶ 사후관리 : 비식별 정보 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용 과정에서 재식별 방지를 위해 필요한 조치를 수행한다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

1) 개인정보 비식별 조치 가이드라인

③ 비식별 조치방법

처리 기법	설명
가명처리(Pseudonymization)	•개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 방법 •세부 기술 : 휴리스틱 가명화, 암호화, 교환방법 예) 김나나, 27세, 서울 거주, A대 재학 ->이연이, 20대, 서울 거주, B대 재학
총계처리(Aggregation)	•통계값을 적용하여 특정 개인을 식별할 수 없도록 하는 방법 •세부 기술 : 총계처리, 부분총계, 라운딩, 재배열 예) 김나나 170cm, 이나리 165cm, 김철수 180cm, 박철우 175cm ->통계학과 학생 합 : 690cm, 평균 키 : 172.5cm
데이터 삭제(Data Reduction)	•민감 데이터 일부 혹은 전체를 삭제하여 개인을 식별할 수 없도록 하는 방법 •세부 기술 : 식별자 삭제, 식별자 부분 삭제, 레코드 삭제, 식별 요소 전부 삭제 예) 주민등록번호 800510-2111111 ->80년대생 여자
데이터 범주화(Data Suppression)	•특정 정보를 해당 그룹의 대푯값으로 변환하거나 구간값으로 변환하여 특정 개인을 식별할 수 없도록 하는 방법 •세부 기술 : 감추기, 랜덤 라운딩, 범위 방법, 제어 라운딩 예) 김나나, 27세 -> 김씨, 20~30세
데이터 마스킹(Data Masking)	•민감 정보 일부를 *와 같은 기호로 표기하는 방법 •세부 기술 : 임의 잡음 추가, 공백화 대체 예) 김나나, 27세, 서울 거주, A대 재학 ->김ㅇㅇ 27세, 서울 거주, ㅇㅇ대 재학

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

1) 개인정보 비식별 조치 가이드라인

④ 적정성 평가

▶ 비식별 조치가 충분하지 않은 경우 공개 정보 등 다른 정보와의 결합, 다양한 추론 기법 등을 통해 개인이 식별될 우려가 있다.

▶ 개인정보보호책임자 책임 하에 외부전문가가 참여하는 비식별조치 적정성 평가단을 구성, 개인식별 가능성에 대한 엄격한 평가가 필요하다.

▶ 적정성 평가 시 프라이버시 보호 모델 중 k-익명성을 활용한다.

[적정성 평가 절차]

기초자료 작성 → 평가단 구성 → 평가 수행 → 추가 비식별 조치 → 데이터 활용

⑤ 사후 관리

▶ 비식별 정보 안전조치

▶ 재식별 가능성 모니터링

▶ 비식별 정보 제공 및 위탁계약 시 준수사항

▶ 재식별 조치요령

1. 데이터 분석 계획 – 분석 방안 수립

개념 체크

1. 다음 중 데이터 변환 기술에 속하지 않는 것은?

① 섭동 ② 평활화

③ 집계 ④ 정규화

데이터 변환 기술은 평활화, 집계, 일반화, 정규화, 속성 생성이 있다. 섭동은 개인정보 익명 처리 기법 중 하나이다.

2. 데이터 변환 기술 중 하나로 특정 구간의 값을 스케일링 하는 기법은?

① 평활화 ② 정규화

③ 속성 생성 ④ 일반화

데이터 변환 기술 중 하나로 특정 구간의 값을 스케일링 하는 기법은 일반화이다.

3. 다음과 같은 개인정보 비식별 조치 방법은?

주민등록번호 950305-2345678 → 95년생 여자

① 가명처리 ② 총계처리

③ 데이터 범주화 ④ 데이터 삭제

위의 내용은 개인정보 비식별 조치 방법 중 데이터 삭제에 대한 내용이다.

4. 다음 중 개인정보 비식별 조치에 대한 적정성 평가 절차 에 속하지 않는 것은?

① 데이터 응용 ② 기초자료 작성

③ 평가단 구성 ④ 추가 비식별 조치

개인정보 비식별 조치에 대한 적정성 평가 절차

기초자료 작성 -> 평가단 구성 -> 평가 수행 -> 추가 비식별 조치 -> 데이터 활용 순이다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 재현 데이터(Synthetic Data)

- 재현 데이터는 실제 데이터와 특성이 유사하여 실제 데이터를 분석한 결과와 유사한 결과를 얻을 수 있도록 인공적으로 재현하여 생성한 가상의 데이터이다.
- 재현 데이터는 개인정보보호 등을 이유로 실제 데이터에 접근하기 어려운 경우 혹은 학습에 사용될 실제 데이터가 현저히 적은 경우 사용한다.
- 재현 데이터는 재구성된 가상의 데이터이기 때문에 실제 데이터와 달리 법적인 제약에서 자유롭고 다양한 형태로 데이터를 재구성할 수 있다.
- 재현 데이터의 유형에는 완전 재현 데이터, 부분 재현 데이터, 복합 재현 데이터가 있다.

명칭	설명
완전 재현 데이터 (Fully Synthetic Data)	원본 데이터 중 전체 데이터를 재현 데이터로 대체한 데이터로 정보보호 측면에서 가장 강력한 보안성을 가진다.
부분 재현 데이터 (Partially Synthetic Data)	원본 데이터 중 일부 데이터(민감 데이터)만 재현 데이터로 대체한 데이터
복합 재현 데이터 (Hybrid Synthetic Data)	일부 변수들의 값을 재현 데이터로 생성하고, 생성된 재현 데이터를 실제 데이터 모두 이용하여 또 다른 일부 변수들의 값을 다시 도출하는 방식으로 생성한 데이터

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

3) 개인정보 익명처리 기법

; 대표적인 개인정보 익명처리 기법에는 가명, 일반화, 섭동, 치환 등이 있다.

● 개인정보 익명처리 기법

명칭	설명
가명(Pseudonym)	개인을 식별할 수 있는 값을 다른 값으로 대체하는 기법
일반화(Generalization)	구체적인 값을 일반화된 값으로 대체하는 기법
섭동(Perturbation)	동일한 확률적 정보를 가지는 변형된 값을 원래 데이터로 대체하는 기법
치환(순열) (Permutation)	분석 시 가치가 적고 식별성이 높은 열 항목에 대해 대상 열 항목의 모든 값을 열 항목 내에서 무작위로 순서를 변경하여 식별성을 낮추는 기법

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

05 데이터 품질 검증

1) 데이터 품질 특성

- 양질의 데이터 분석 결과를 얻기 위해서는 수집되는 데이터의 품질이 보장되어야 한다.
- 고품질 데이터의 품질 요소는 정확성, 완전성, 적시성, 일관성이 있다.
- 고품질 데이터 품질 요소

지표	설명
정확성(Accuracy)	제공되는 데이터는 사용 목적에 맞게 정확해야 한다.
완전성(Completeness)	제공되는 데이터의 완전한 형태로 제공되야 한다.
적시성(Timeliness, 시의성)	제공되는 데이터는 사용되는 목적에 맞게 활용 시점이 자유로워야 한다.
일관성(Consistency)	데이터 사용 목적에 따라 일관된 데이터 활용 기준이 제시되어야 한다.

- 데이터의 형태(정형 데이터, 비정형 데이터)에 따라 품질 기준이 다르다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

1) 데이터 품질 특성

① 정형 데이터의 품질 기준

▶ 정형 데이터의 품질 기준에는 완전성, 유일성, 유효성, 일관성, 정확성이 있다.

▶ 정형 데이터의 품질 기준

품질 기준	설명
완전성(Completeness)	필수 데이터 누락 없이 완전한 형태로 데이터가 존재해야 한다. 예) 쇼핑몰 회원 데이터에서 고객 주소지 정보가 누락되면 안 된다.
유일성(Uniqueness)	데이터 항목은 유일하게 존재해야 하고 중복되면 안 된다. 예) 입력된 고객의 핸드폰 번호는 유일해야 한다.
유효성(Validity)	데이터 항목은 유효 범위 및 도메인을 충족해야 한다. 예) 주민번호 형식은 000000-0000000 형식을 맞춰야 한다.
일관성(Consistency)	데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되며 서로 일치해야 한다. 예) 고객의 주문번호와 해당 주문 고객 아이디는 일치해야 한다.
정확성(Accuracy)	데이터는 실세계에 존재하는 객체의 표현값을 정확히 반영해야 한다. 예) 주문 부분 취소로 수정된 고객 주문 정보와 입력된 구매 정보가 동일해야 한다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

1) 데이터 품질 특성

② 비정형 데이터의 품질 기준

▶ 비정형 데이터의 품질 기준에는 기능성, 신뢰성, 사용성, 효율성, 이식성이 있다.

▶ 비정형 데이터의 품질 기준

품질 기준	설명
기능성(Functionality)	해당 콘텐츠가 특정 조건에서 사용될 때 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 성질 [품질 세부 기준] 적절성, 정확성, 상호 운용성, 기능 순응성
신뢰성(Reliability)	해당 콘텐츠가 규정된 조건에서 사용될 때 규정된 신뢰 수준을 유지하거나 사용자로 하여금 오류를 방지할 수 있도록 하는 성질 [품질 세부 기준] 성숙성, 신뢰 순응성
사용성(Usability)	해당 콘텐츠가 규정된 조건에서 사용될 때, 사용자에 의해 이해되고 선호될 수 있게 하는 성질 [품질 세부 기준] 이해성, 친밀성, 사용 순응성
효율성(Efficiency)	해당 콘텐츠가 규정된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 성질 [품질 세부 기준] 시간 효율성, 자원 효율성, 효율 순응성
이식성(Portability)	해당 콘텐츠가 다양한 환경과 상황에서 실행될 가능성 [품질 세부 기준] 적응성, 공존성, 이식 순응성

1. 데이터 분석 계획 – 분석 방안 수립

개념 체크

1. 다음 중 개인정보 비식별 조치 가이드 라인 단계의 순서 로 옳은 것은?

① 비식별 조치 → 사전 검토 → 사후관리 → 적정성 평가

② 적정성 평가 → 사전 검토 → 비식별 조치 → 사후관리

③ 사전 검토 → 비식별 조치 → 적정성 평가 → 사후관리

④ 사전 검토 → 사후관리 → 비식별 조치 → 적정성 평가

개인정보 비식별 조치 가이드 라인 단계는 사전 검토 -> 비식별 조치 -> 적정성 평가 -> 사후관리 순이다.

2. 다음 중 개인정보 비식별 조치 가이드 라인의 사후관리에 포함되지 않는 것은?

① 재식별 가능성 모니터링

② 비식별 정보 안전조치

③ 데이터 활용

④ 비식별 정보 제공 및 위탁계약 시 준수사항

사후관리

1. 비식별 정보 안전조치

2. 재식별 가능성 모니터링

3. 개인정보 익명 처리 기법 중 하나로 개인을 식별할 수 있는 다른 값으로 대체하는 기법은?

① 가명 ② 일반화

③ 섭동 ④ 치환

가명 : 개인을 식별할 수 있는 값을 다른 값으로 대체하는 기법

일반화 : 구체적인 값을 일반화된 값으로 대체하는 기법

섭동 : 동일한 확률적 정보를 가지는 변형된 값을 원래 데이터로 대체하는 기법

치환(순열) : 분석 시 가치가 적고 식별성이 높은 열 항목에 대해 대상 열 항목의 모든 값을 열 항목 내에서 무작위로 순서를 변경하여 식별성을 낮추는 기법

4. 다음 중, 고품질 데이터의 품질 요소에 해당하지 않는 것은?

① 정확성 ② 완전성

③ 적시성 ④ 관계성

정확성 : 제공되는 데이터는 사용 목적에 맞게 정확해야 한다.

완전성 : 제공되는 데이터의 완전한 형태로 제공되어야 한다.

1. 데이터 분석 계획 – 분석 방안 수립

개념 체크

5. 재현 데이터(Synthetic Data)에 대한 설명으로 틀린 것은?

① 재현 데이터는 실제 데이터와 특성이 유사하다.

② 인공적으로 재현하여 생성한 가상의 데이터이다.

③ 재현 데이터는 개인정보보호 등을 이유로 실제 데이터에 접근하기 어려운 경우 혹은 학습에 사용될 실제 데이터가 많은 경우 사용한다.

④ 재현 데이터는 재구성된 가상의 데이터이기 때문에 실제 데이터와 달리 법적인 제약에서 자유롭고 다양한 형태로 데이터를 재구성할 수 있다.

재현 데이터는 개인정보보호 등을 이유로 실제 데이터에 접근하기 어려운 경우 혹은 학습에 사용될 실제 데이터가 현저히 적은 경우 사용한다.

6. 다음 중, 비정형 데이터의 품질 기준에 속하지 않는 것은?

① 기능성 ② 신뢰성

③ 사용성 ④ 정확성

비정형 데이터의 품질 기준에는 기능성, 신뢰성, 사용성, 효율성, 이식성이 있다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

2) 데이터 변환 후 품질 검증

① 메타데이터 활용 : 메타데이터는 구조화된 데이터로 다른 데이터를 설명해주는 데이터이다.

[메타데이터를 통한 유효성 분석 과정]

메타데이터 수집 -> 메타데이터 분석 -> 데이터 속성 분석

② 정규 표현식 활용 : 정규 표현식은 일정한 규칙을 갖는 문자열의 집합을 표현하는 데 사용되는 언어이다.

예) ₩ : 특수문자 표기, | : OR(또는), $: 종료 문자열

③ 데이터 프로파일링 활용 : 데이터 프로파일링은 데이터 현황 분석을 위한 자료 수집을 통해 잠재적 오류 징후를 발견하는 방법이다.

[데이터 프로파일링 절차]

메타데이터 수집 및 분석 -> 대상 및 유형 선정 -> 프로파일링 수행 -> 프로파일링 결과 리뷰 -> 프로파일링 결과 총합

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

3) 데이터 품질 진단 절차

품질 진단 계획수립 → 품질 기준 및 진단 대상 정의 → 데이터 품질 측정 → 품질 측정 결과 분석 → 데이터 품질 개선

① 품질 진단 계획수립 : 프로젝트 정의, 조직 정의 및 편성, 품질 진단 절차 정의, 세부시행계획 확정, 품질 기준 및 진단 대상 정의 순으로 품질 진단 계획을 수립한다.

② 품질 기준 및 진단 대상 정의 : 품질 기준 선정, 품질 이슈 조사, 데이터 관리 문서 수집, 진단 대상 중요도 평가, 진단 대상 선정, 핵심 데이터 항목 정의, 데이터 프로파일링, 업무규칙 정의 순으로 품질 기준 및 진단 대상을 정의한다.

③ 데이터 품질 측정 : 품질 측정 계획수립, 품질 측정 체크리스트 준비, 데이터 품질 측정 수행, 데이터 품질 측정 결과 보고 순으로 데이터 품질을 측정한다.

④ 품질 측정 결과분석 : 오류가 발견된 컬럼 또는 측정항목에 대해 품질 기준별, 발생 유형별 오류 원인을 분석하고, 주요 발생 사례를 정리한다. 주요 오류 원인별 개선방안을 도출한다.

⑤ 데이터 품질 개선 : 도출된 개선안과 우선순위에 따라 세부 수행 일정과 책임 소재, 관련조직 및 업무 관련자에 대한 공지 계획 등이 포함된 품질 개선을 계획 수립한다. 수립된 품질 개선 계획에 따라 개선 활동을 수행하며, 품질담당자는 개선 진행 상황을 모니터링 하여 전체적인 조율 및 진행률을 관리한다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

개념 체크

1. 다음 중 구조화된 데이터로 다른 데이터를 설명해 주는 데이터를 무엇이라고 하는가?

① 메타 데이터 ② 시스템 데이터

③ 집계 데이터 ④ 로그 데이터

메타 데이터는 구조화된 데이터로 다른 데이터를 설명해 주는 데이터이다.

2. 다음 중, 데이터 프로파일링 절차로 알맞은 것은?

① 메타 데이터 수집 및 분석 -> 대상 및 유형 선정 -> 프로파일링 수행 -> 프로파일링 결과 리뷰 -> 프로파일링 결과 총합

② 메타 데이터 수집 및 분석 -> 대상 및 유형 선정 -> 프로파일링 수행 -> 프로파일링 결과 총합 -> 프로파일링 결과 리뷰

③ 대상 및 유형 선정 -> 프로파일링 수행 -> 프로파일링 결과 총합 -> 프로파일링 결과 리뷰 -> 메타 데이터 수집 및 분석

④ 대상 및 유형 선정 -> 프로파일링 결과 총합 -> 프로파일링 수행 -> 메타 데이터 수집 및 분석 -> 프로파일링 결과 리뷰

3. 다음 중, 데이터 품질 진단 절차로 알맞은 것은?

① 품질 진단 계획수립 → 품질 기준 및 진단 대상 정의 → 데이터 품질 측정 → 품질 측정 결과 분석 → 데이터 품질 개선

② 품질 기준 및 진단 대상 정의 → 품질 진단 계획수립 → 데이터 품질 측정 → 품질 측정 결과 분석 → 데이터 품질 개선

③ 품질 진단 계획수립 → 품질 기준 및 진단 대상 정의 → 데이터 품질 측정 → 데이터 품질 개선 → 품질 측정 결과 분석

④ 품질 기준 및 진단 대상 정의 → 데이터 품질 측정 → 품질 진단 계획수립 → 품질 측정 결과 분석 → 데이터 품질 개선

데이터 품질 진단 절차는 품질 진단 계획수립 -> 품질 기준 및 진단 대상 정의 -> 데이터 품질 측정 -> 품질 측정 결과 분석 -> 데이터 품질 개선 순이다.

1. 데이터 수집 및 저장 계획 – 데이터 적재 및 저장

01 데이터 적재

- 데이터가 수집되면 데이터의 전처리 작업 전에 해당 데이터를 빅데이터 시스템에 적재(Load)해야 한다.
- 수집된 데이터를 기반으로 데이터를 추출(Extract), 변환(Transform), 적재(Load)의 과정을 거치는 작업이 ETL이다.
- 데이터 적재 도구

종류	설명
플루언티드(Fluented)	트레저 데이터(Treasure Data)에서 개발된 크로스 플랫폼 오픈소스 데이터 수집 소프트웨어
플럼(Flume)	많은 양의 로그 데이터를 효율적으로 수집, 집계, 이용하기 위해 이벤트(Event)와 에이전트(Agent)를 활용하는 기술
스크라이브(Scribe)	다수의 서버로부터 실시간으로 스트리밍 되는 로그 데이터를 수집하여 분산 시스템에 데이터를 저장하는 대용량 실시간 로그 수집 기술
로그스태시(Logstash)	실시간 파이프라인 기능을 갖는 오픈소스 데이터 수집 엔진

파이프라인(pipeline) : 데이터 처리 단계의 출력이 다음 단계의 입력으로 이어지는 형태로 연결된 구조

1. 데이터 수집 및 저장 계획 – 데이터 적재 및 저장

02 데이터 저장

- 데이터가 수집되면 데이터 전처리 작업 후에 해당 데이터를 활용할 수 있도록 저장해야 한다.
- 정형 데이터는 관계형 데이터베이스(RDBMS), 반정형 데이터는 NoSQL, 비정형 데이터는 분산 파일 시스템(DFS)에 저장된다.

① 관계형 데이터베이스(RDBMS) : 전형적인 데이터베이스 형태의 테이블로 이루어져 있고, 테이블은 키(key)와 값(value)의 관계를 나타낸다.

② 비관계형 데이터베이스(NoSQL, Not Only SQL)

- ▶ 대부분의 전형적인 데이터베이스 시스템에서 찾을 수 있는 행과 열로 이루어진 테이블 형식 스키마를 사용하지 않은 데이터베이스이다.
- ▶ 대규모 데이터를 저장하기 위한 DBMS(Database Management System, 데이터베이스 관리 시스템)이다.
- ▶ 고정된 테이블 스키마가 없고, 조인(JOIN) 연산을 사용할 수 없으며, 수평적 확장이 가능하다.

1. 데이터 수집 및 저장 계획 – 데이터 적재 및 저장

02 데이터 저장

② 비관계형 데이터베이스(NoSQL, Not Only SQL)

▶ NoSQL유형에는 Key-Value Store, Column Family Data Store, Document Store, Graph Store가 있다.

유형	설명
Key-Value Store	유일한 key와 하나의 value를 갖는 데이터베이스 예) DynamoDB, Redis
Column Family Data Store	key 안에 (Column, Value) 조합으로 된 여러 개의 필드를 갖는 데이터베이스 예) Cassandra, HBase
Document Store	Value의 데이터 타입이 Document 타입을 사용하는 데이터베이스 예) Couchbase, MongoDB
Graph Store	시맨틱 웹(Sementic Web)과 온톨로지(Ontology) 분야에서 활용되는 그래프로 데이터를 표현하는 데이터베이스 예) AllegroGraph, Neo4j

▶ NoSQL은 CAP 이론을 기반으로 한다.

시맨틱 웹 : 인터넷과 같은 분산 환경에서 리소스에 대한 정보와 자원 사이의 관계-의미 정보를 기계가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 자동화된 기계가 처리하도록 하는 기술
온톨로지 : 사람들이 세상에 대하여 보고 듣고 느끼고 생각하는 것에 대하여 서로 간의 토론을 통하여 합의를 이룬 바를 개념적이고 컴퓨터에서 다룰 수 있는 형태로 표현한 모델
CAP 이론 : 시스템은 일관성(Consistency), 가용성(Availability), 분단 허용성(Partition Tolerance) 세 가지 속성 중에서, 두 가지만 가질 수 있다는 것이다.

1. 데이터 수집 및 저장 계획 – 데이터 적재 및 저장

02 데이터 저장

③ 네트워크를 기반으로 파일을 수집, 저장, 공유할 수 있는 시스템이다.

▶ 데이터 저장 기술

명칭	설명
데이터 웨어하우스 (DW, Data Warehouse)	•사용자의 의사결정에 도움을 주기 위해 다양한 소스(Source)에서 수집된 대량의 원시 데이터를 주제별로 공통의 형식으로 변환하여 장기간 저장하는 데이터 저장소 •보통의 경우 구조화된 정형 데이터를 보관 •데이터 웨어하우스(DW)의 데이터는 향후 데이터 마트(DM)에 제공된다.
데이터 마트 (DM, Data Mart)	•특정 부서가 필요로 하는 분석 목적에 맞는 데이터를 다루기 위해 구축된 데이터 저장소 •데이터 웨어하우스(DW)보다 적은 소스(source)로부터 데이터를 수집
데이터 레이크 (Data Lake)	•가공되지 않은 다양한 종류의 데이터(Raw Data)를 저장할 수 있는 데이터 저장소 •데이터 레이크는 데이터를 기본 형식으로 저장할 수 있고, 스키마(schema)와 상관없이 저장 가능
데이터 댐 (Data Dam)	•데이터 수집 기술을 활용하여 다양한 산업 의 방대한 원시 데이터(Raw Data)들을 한 곳에 모아 둔 댐(Dam)과 같은 데이터 저장소 •수집된 데이터들은 향후 목적에 맞게 사용할 수 있도록 가공되어 네트워크 기술을 활용하여 다양한 산업 분야로 제공될 수 있다.

스키마(schema) : 데이터베이스의 구조와 제약 조건에 관한 전반적인 명세를 기술한 메타 데이터(다른 데이터를 설명해주는 데이터)의 집합

1. 데이터 수집 및 저장 계획 – 데이터 적재 및 저장

개념 체크

1. 다음 중 데이터 적재 도구가 아닌 것은?

① 플루언티드　　② 플럼

③ 스크라이브　　④ 데이터 댐

데이터 적재 도구에는 플루언티드, 플럼, 스크라이브, 로그스태시 가 있다. 데이터 댐은 데이터 저장 기술 중 하나이다.

2. 다음 중 NoSQL유형에 속하지 않는 것은?

① Key-Value Store　　② Column Family Data Store

③ Document Store　　④ Row Store

NoSQL 유형 비관계형 데이터베이스에는 Key-Value Store, Column Family Data Store, Document Store, Graph Store 가 있다.

3. 다음 중, 데이터 저장 기술이 아닌 것은?

① Data Rounge　　② DW

③ DM　　④ Data Lake

데이터 저장 기술에는 DW(데이터 웨어 하우스), DM(데이터 마트), 데이터 레이크, 데이터 댐 이 있다.

1. 데이터 수집 및 저장 계획 예상 문제

예상 문제

1. 다음 중 데이터 수집 순서로 옳은 것은?

① 수집 데이터 도출 → 수집 데이터 목록화 → 데이터 소유 기관 확인 및 협의 → 데이터 유형 확인 및 분류 → 데이터 수집 기술 선정 → 수집 계획서 작성 → 수집 주기 정의 → 데이터 수집

② 수집 데이터 목록화 → 데이터 소유 기관 확인 및 협의 → 데이터 유형 확인 및 분류 → 데이터 수집 기술 선정 → 수집 계획서 작성 → 수집 주기 정의 → 수집 데이터 도출 → 데이터 수집

③ 데이터 수집 수집 데이터 도출 → 수집 데이터 목록화 → 데이터 소유 기관 확인 및 협의 → 데이터 유형 확인 및 분류 → 데이터 수집 기술 선정 → 수집 계획서 작성 → 수집 주기 정의

④ 수집 주기 정의 → 수집 데이터 도출 → 수집 데이터 목록화 → 데이터 소유 기관 확인 및 협의 → 데이터 유형 확인 및 분류 → 데이터 수집 기술 선정 → 수집 계획서 작성 → 데이터 수집

2. 다음 중 데이터 수집 기술이 아닌 것은?

① FTP ② Scrapy

③ Flume ④ DW

데이터 수집 기술에는 ETL, FTP, 스쿱, 스크래피, 플럼, 스크라이브 등이 있다. DW는 데이터 저장 기술 중 하나인 데이터 웨어 하우스의 약자이다.

3. 다음 중 ETL에 속하지 않는 작업은?

① Extract ② Transform

③ Trade ④ Load

ETL은 대표적인 데이터 수집 기술 중 하나로 Extract(추출), Transform(변환), Load(적재)의 약자이다.

4. 데이터 수집 기술 중 하나로 파이썬 언어 기반의 비정형 데이터 수집 기술로 웹 데이터를 수집하는 것을 목표로 설계된 기술은?

① CEP ② Apache Kafka

③ FTP ④ Scrapy

1. 데이터 수집 및 저장 계획 예상 문제

예상 문제

5. 인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠를 수집하는 기술은?

① Chukwa ② Crawling

③ Sqoop ④ TCP/IP

6. 다음 중 구조 관점의 데이터 유형이 아닌 것은?

① 정형 데이터 ② 마이 데이터

③ 반정형 데이터 ④ 비정형 데이터

구조 관점의 데이터 유형은 정형, 반정형, 비정형 데이터가 있다.

7. 데이터 유형 중 저장 형태의 관점으로 분류되는 데이터 유형이 아닌 것은?

① 스트림 데이터② 콘텐츠 데이터

③ 파일 데이터 ④ 실시간 데이터

데이터 유형 중 저장 형태의 관점으로 분류되는 데이터 유형에는 파일 데이터, 데이터베이스 데이터, 콘텐츠 데이터, 스트림 데이터가 있다. 실시간 데이터는 시간적 관점으로 분류되는 데이터이다.

8. 다음과 같은 특징을 갖는 데이터는?

- 정형화된 스키마의 구조를 갖고, 고정된 필드에 데이터가 저장된다.
- 데이터가 형식의 일관성을 갖고 컬럼(Column)과 로우(Row)의 구조를 갖는다.
- 대표적인 예로 관계형 데이터베이스, 스프레드시트가 있다.

① 정형 데이터 ② 반정형 데이터

③ 비정형 데이터 ④ 비실시간 데이터

1. 데이터 수집 및 저장 계획 예상 문제

예상 문제

9. 구조 관점의 데이터 유형에 대한 설명 중 틀린 것은?

① 구조 관점의 데이터는 정형, 반정형, 비정형 데이터로 나뉜다.

② 정형 데이터는 형식의 일관성을 갖고, 열과 행의 구조를 갖는다.

③ 반정형 데이터는 스키마 구조를 갖지 않는다.

④ 비정형 데이터의 예로는 텍스트, 이미지, 오디오 데이터 가 있다.

반정형 데이터는 스키마 구조 형태를 가지고, 메타 데이터를 포함하며, 값과 형식에서 일관성을 가지지 않는 데이터이다.

10. 다음 중 생성된 데이터가 수 시간 또는 수 주 이후에 처리되어야 유의미한 결과를 얻을 수 있는 데이터는?

① 실시간 데이터② 비실시간 데이터

③ 파일 데이터 ④ 공유 데이터

11. 다음 중 텍스트, 이미지, 오디오, 비디오 등과 같이 개별적으로 데이터 객체로 구분되는 미디어 데이터는?

① 파일 데이터

② 데이터베이스 데이터

③ 콘텐츠 데이터

④ 스트림 데이터

12. 데이터에 관한 구조화된 데이터로서 다른 데이터를 설명해주는 데이터는?

① 메타 데이터 ② 마이 데이터

③ 공유 데이터 ④ 분석 데이터

1. 데이터 수집 및 저장 계획 예상 문제

예상 문제

13. 다음 중 데이터 변환 기술에 속하지 않는 것은?

① 섭동 ② 평활화

③ 집계 ④ 정규화

데이터 변환 기술은 평활화, 집계, 일반화, 정규화, 속성 생성이 있다. 섭동은 개인정보 익명 처리 기법 중 하나이다.

14. 데이터 변환 기술 중 하나로 특정 구간의 값을 스케일링하는 기법은?

① 평활화 ② 정규화

③ 속성 생성 ④ 일반화

데이터 변환 기술 중 하나로 특정 구간의 값을 스케일링하는 기법은 일반화이다.

15. 다음 중 개인정보 비식별 조치 가이드 라인 단계의 순서로 옳은 것은?

① 비식별 조치 → 사전 검토 → 사후관리 → 적정성 평가

② 적정성 평가 → 사전 검토 → 비식별 조치 → 사후관리

③ 사전 검토 → 비식별 조치 → 적정성 평가 → 사후관리

④ 사전 검토 → 사후관리 → 비식별 조치 → 적정성 평가

16. 다음과 같은 개인정보 비식별 조치 방법은?

주민등록번호 810305-1345678 → 81년생 남자

① 가명처리 ② 총계처리

③ 데이터 범주화 ④ 데이터 삭제

17. 다음 중 개인정보 비식별 조치에 대한 적정성 평가 절차에 속하지 않는 것은?

① 데이터 응용 ② 기초자료 작성

③ 평가단 구성 ④ 추가 비식별 조치

개인정보 비식별 조치에 대한 적정성 평가 절차

기초자료 작성 -> 평가단 구성 -> 추가 비식별 조치 -> 데이터 활용 순이다.

18. 다음 중 개인정보 비식별 조치 가이드 라인의 사후관리 에 포함되지 않는 것은?

① 재식별 가능성 모니터링

② 비식별 정보 안전조치

③ 데이터 활용

④ 비식별 정보 제공 및 위탁계약 시 준수사항

1. 데이터 수집 및 저장 계획 예상 문제

예상 문제

19. 개인정보 익명 처리 기법 중 하나로 개인을 식별할 수 없는 다른 값으로 대체하는 기법은?

① 가명 ② 일반화

③ 섭동 ④ 치환

20. 다음과 같은 특징을 갖는 데이터 저장 기술은?

- 다양한 자원에서 수집된 대량의 데이터를 주제별로 장기간 저장하는 데이터 저장소
- 보통의 경우 구조화된 정형 데이터를 보관

① 데이터 마트 ② 데이터 웨어하우스

③ 데이터 레이크 ④ 데이터 스토어

1. 빅데이터 분석 기획 – 마무리 문제

1. 다음 설명 중 성질이 다른 것은?

① 작년 한국의 출산율은 0.81이다.

② A 회사의 개발팀 직원 수는 30명이고, B 회사의 개발팀 직원 수는 40명이다.

③ C 서점에서 판매되는 Apple책은 10,000원이다.

④ D 회사가 제공하는 웹 서비스는 총 5개이다.

①은 데이터를 가공, 처리하여 데이터의 의미가 도출된 정보(Information)에 대한 설명이다. ②, ③, ④은 가공 전에 순수한 데이터(Data)에 대한 설명이다.

2. 다음 중 3V에 해당하는 요소를 고른 것은?

ⓐ Value ⓑ Veracity ⓒ Volume

ⓓ Variety ⓔ Velocity

① ⓐ, ⓑ, ⓒ ② ⓒ, ⓓ, ⓔ

③ ⓐ, ⓒ, ⓓ ④ ⓑ, ⓒ, ⓓ

3V는 Volume(규모), Variety(다양성), Velocity(속도)이다.

3. 다음 중 PB의 크기는?

① 2^{40}Bytes

② 2^{50}Bytes

③ 2^{60}Bytes

④ 2^{70}Bytes

KB = 2의 10승, MB = 2의 20승, GB = 2의 30승, TB = 2의 40승, PB = 2의 50승, EB = 2의 60승, ZB = 2의 70승, YB = 2의 80승이다.

4. 다음 중 빅데이터 분석기술이 경제적 효율성을 갖고 발전할 수 있었던 주된 요인은?

① 정보통신 기술의 발달

② 4차 산업 시대의 시작

③ 클라우드 컴퓨팅

④ 머신러닝

빅데이터 분석기술이 경제적 효율성을 갖고 발전할 수 있었던 가장 큰 주된 요인은 클라우드 컴퓨팅 기술이다.

1. 빅데이터 분석 기획 – 마무리 문제

5. 다음 내용에 해당하는 용어는?

- DIKW 피라미드의 가장 상위 단계에 속한다.
- 데이터에 대한 누적된 이해를 바탕으로 도출되는 창의적인 판단이다.

① 지식 ② 데이터

③ 정보 ④ 지혜

DIKW 피라미드

지혜(4단계) : 데이터에 누적된 이해를 바탕으로 도출되는 창의 판단

지식(3단계) : 다양한 정보를 체계화하여 유의미한 정보로 분류시킨 대상

정보(2단계) : 사용자의 분석으로 데이터 간의 연관 관계 및 의미가 생성된 데이터

데이터(1단계) : 측정된 값으로 다른 데이터와 상관관계가 없는 가공 전에 원본 데이터

6. 데이터 지식경영의 상호작용에서 개인이 암묵지 경험을 공유함으로써 타인이 암묵지를 습득하는 과정을 일컫는 말은?

① 공통화 ② 표출화

7. 다음 중 빅데이터에 대한 설명으로 틀린 것은?

① 빅데이터는 많은 양(수십 TB)의 정형 및 비정형 데이터 를 의미한다.

② 빅데이터는 수집되고 분석되는 과정을 통해 유의미한 가치를 얻게 된다.

③ 빅데이터는 개인, 기업, 정부 등 다양한 곳에서 활용이 가능하다.

④ 빅데이터의 가치를 산정하는 것은 비교적 쉽다.

빅데이터는 데이터 활용방식, 새로운 가치 창출, 분석기술 발전 등의 이유로 정확한 가치를 산정하기에는 어려움이 있다.

8. 다음 중 빅데이터 위기 요인에 속하지 않는 것은?

① 사생활 침해

② 책임 원칙 훼손

③ 데이터 남용

④ 데이터 오용

빅데이터 위기 요인에는 사생활 침해, 책임 원칙 훼손, 데이터 오용이 있다.

1. 빅데이터 분석 기획 – 마무리 문제

9. 분석 가치 에스컬레이터에서 현재 무슨 일이 일어나고 있는지에 대한 분석을 하는 단계는?

① 묘사 분석
② 진단 분석
③ 예측 분석
④ 처방 분석

묘사 분석
- 분석의 가장 기본적인 지표를 확인하는 단계이며, 과거에 어떤 일이 일어났고, 현재 어떤 일이 일어나고 있는지를 확인하는 것

진단 분석
- 묘사 단계에서 찾아낸 분석의 원인을 이해하는 단계이며, 데이터를 기반으로 왜 발생했는지 이유를 확인하는 것

예측 분석
- 데이터를 통해 기업 혹은 조직의 미래, 고객의 행동 등을 예측하는 단계이며, 무슨 일이 일어날 것인지를 예측하는 것

처방 분석
- 예측을 바탕으로 최적화하는 단계이며, 무엇을 해야 할

11. 빅데이터 조직 구조 유형 중 다음과 같은 특징을 갖는 구조는?

- 분산조직 인력들을 현업 부서로 직접 배치하여 분석 업무를 수행한다.
- 분석 결과에 따라 신속한 피드백이 나오고 Best Practice 공유가 가능하다
- 업무 과다, 이원화의 가능성이 존재할 수 있기 때문에 부서 분석 업무와 역할 분담이 명확해야 한다.

① 분산 구조　② 집중 구조
③ 분리 구조　④ 기능 구조

분산 구조
- 분석 전문 인력을 현업 부서에 배치하여 분석 업무 수행
- 전사 차원에서 분석과제의 우선순위를 선정하고 수행
- 분석 결과를 현업에 빠르게 적용 가능

집중 구조
- 전사 분석 업무를 별도의 전담조직에서 수행
- 내부에서 전사 분석과제의 전략적 중요도에 따라 우선순위를 정함

1. 빅데이터 분석 기획 – 마무리 문제

13. 다음 중 조직평가를 위한 성숙도 단계에 속하지 않는 것은?

① 최적화 단계　② 확산 단계

③ 활용 단계　④ 응용 단계

성숙도 단계

도입 단계 – 분석을 시작해 환경과 시스템을 구축하는 단계

활용 단계 – 분석 결과를 실제 업무에 적용하는 단계

확산 단계 – 전사 차원에서 분석을 관리하고 공유하는 단계

최적화 단계 – 분석을 진화시켜서 혁신 및 성과 향상에 기여하는 단계

14. 다음 중 빅데이터 플랫폼 구성 요소에 대한 설명으로 틀린 것은?

① 데이터 수집의 대표적인 예시로 ETL, 크롤러 등이 있다.

② 데이터 저장의 대표적인 예시로 Sqoop, Hiho 등이 있다.

③ 데이터 분석의 대표적인 예시로 자연어 처리, 예측, 분석 등이 있다.

④ 데이터 활용의 대표적인 예시로 박스플롯, 인포그래픽 등이 있다.

데이터 저장의 대표적인 예시로는 RDBMS, NoSQL 등이 있다.

15. 다음 중 맵리듀스의 과정으로 옳은 것은?

① Input -> Splitting -> Mapping -> Shuffling -> Reducing -> Final Result

② Input -> Mapping -> Splitting -> Shuffling -> Reducing -> Final Result

③ Input -> Reducing -> Mapping -> Splitting -> Shuffling -> Final Result

④ Input -> Shuffling -> Splitting -> Mapping -> Reducing -> Final Result

16. 하둡 에코 시스템 기술에 대한 설명 중 틀린 것은?

① HDFS : 대용량 파일들을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리할 수 있도록 설계된 하둡 분산 파일 시스템

② Apache Spark : SQL, 스트리밍, 머신러닝 및 그래프 처리를 위한 기본 제공 모듈이 있는 대규모 데이터 처리용 통합 분석 엔진

③ HBase : 리소스 관리와 컴포넌트 처리를 분리한 아파치 소프트웨어 재단의 서브 프로젝트

1. 빅데이터 분석 기획 – 마무리 문제

17. 다음 중 인공지능에 대한 설명으로 옳지 않은 것은?

① 인공지능이란 인간의 학습능력, 인지능력을 인공적으로 학습시켜 일정 수준의 능력을 갖출 수 있도록 만든 소프트웨어이다.

② 인간과 비슷한 수준의 지능을 구사하기 위해서는 많은 양의 데이터가 수집, 학습되어야 한다.

③ 인공지능 관련 기술로 머신러닝, 딥러닝 이 있다.

④ 인공지능 분석용 데이터로 유료 데이터가 선호된다.

인공지능 분석용 데이터로 유료 데이터는 선호되지 않는다. 이는 많은 양의 양질의 데이터를 활용해서 유의미한 결과를 얻게 되는 인공지능의 특성과 맞지 않다.

18. 다음 중 개인정보보호에 대한 설명으로 옳지 않은 것은?

① 개인정보보호는 정보주체자의 개인정보 자기결정권을 철저히 보장하는 활동을 의미한다.

② 개인정보보호는 정보주체자만이 직접 보호 활동을 할 수 있다.

③ 개인을 식별할 수 있는 정보가 대부분 개인정보로 구분되어 활용되기 때문에 개인정보가 유출되면 그 피해가 막심

19. 다음 중 분석 대상과 분석 방법을 모두 알고 있는 경우 선택할 수 있는 분석 유형은?

① 최적화 ② 솔루션

③ 통찰 ④ 발견

20. CRISP-DM 분석 방법론의 분석 절차에 포함되지 않는 것은?

① 데이터 이해 ② 모델링

③ 업무 이해 ④ 샘플링

CRISP-DM의 분석 방법론의 분석 절차는 업무 이해 -> 데이터 이해 -> 데이터 준비 -> 모델링 -> 평가 -> 전개 순이다.

비즈니스 이해를 바탕으로 데이터 분석 목적의 6단계로 진행되는 데이터 마이닝 방법론이다.

KDD 분석 방법론

1996년 Fayyad가 프로파일링 기술을 기반으로 통계적 패턴이나 지식을 찾기 위해 체계적으로 정리한 방법론이다.

분석 절차

감사합니다.