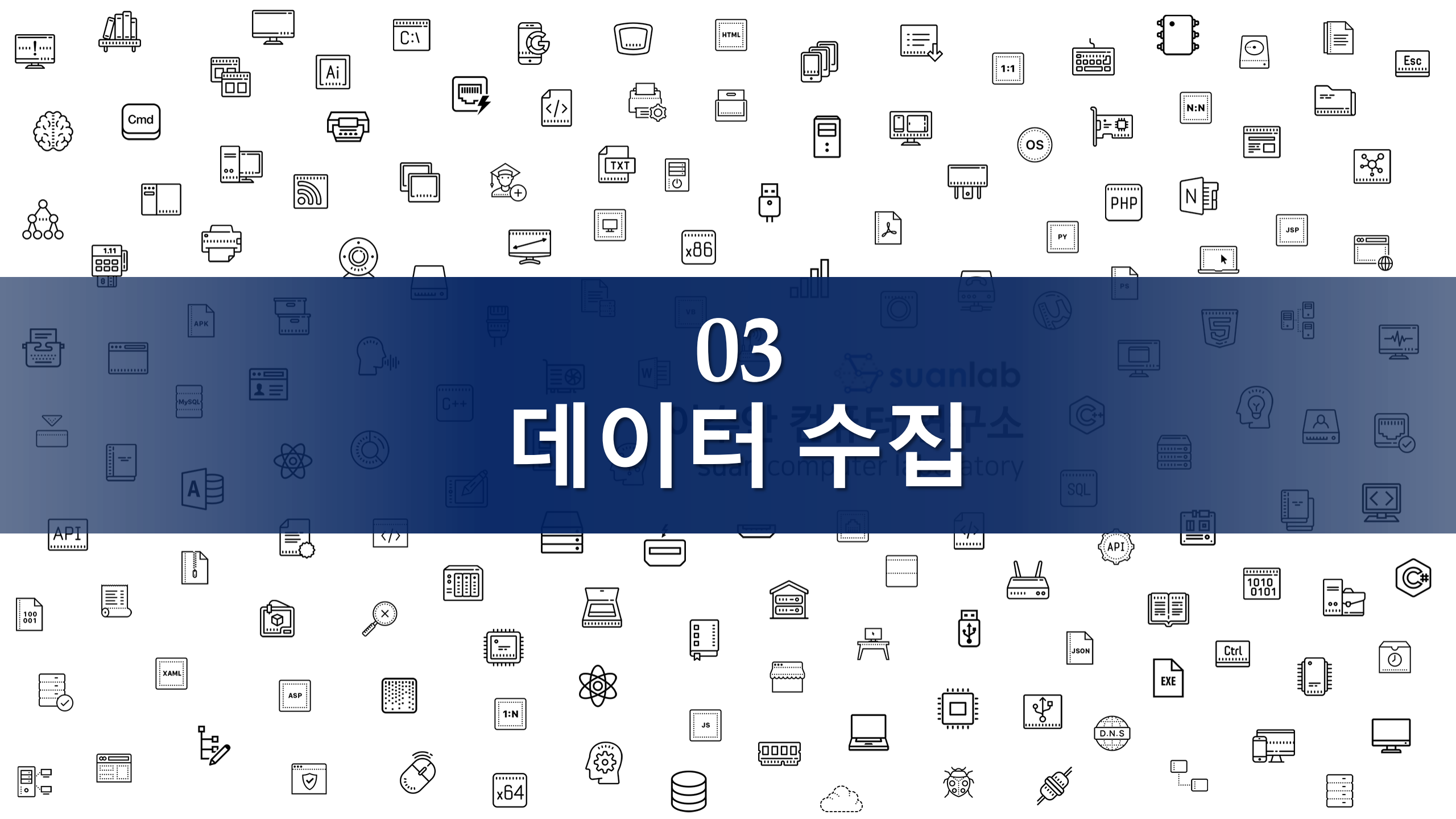
데이터 전처리
Data Preprocessing


03
데이터 수집

목차

1. 데이터 수집 개요
2. 수집 데이터 분류
3. 데이터 수집 절차
4. 데이터 수집 방법 및 기술

1. 데이터 수집 개요

데이터 특징

- 데이터가 가진 특징을 명확히 인지할 때, 데이터를 처리할 수 있는 기술적 사항을 고려한 설계가 가능
- 일반적인 데이터 특징
  - 존재론적 특징
  - 목적론적 특징
- 수집 단계에서 데이터 특징
  - 가역적 데이터
  - 불가역적 데이터

데이터의 존재론적 특징

- 존재적 특징으로 데이터를 구분하면 정성적 데이터qualitative data와 정량적 데이터quantitative data로 구분
- 정성적 데이터는 데이터 자체가 하나의 텍스트(“환율이 내리고 있어 올해 목표한 수출 목표의 조기 달성이 가능해 보인다.”)를 이루고 있기 때문에 데이터 하나 하나가 함축된 정보를 가짐
- 정량적 데이터는 여러 속성(이름, 나이, 성별, 주소 등)이 모여 하나의 객체를 형성하고, 각 속성은 속성 하나 혹은 여러 개의 속성이 결합해 측정이나 설명이 가능

구분	정성적 데이터	정량적 데이터
형태	비정형 데이터	정형, 반정형 데이터
특징	객체 하나에 함의된 정보를 가짐	속성이 모여 객체를 이룸
구성	언어, 문자 등으로 이루어짐	수치, 도형, 기호 등으로 이루어짐
저장 형태	파일, 웹	데이터베이스, 스프레드시트
소스 위치	외부 시스템(주로 소셜 데이터)	내부 시스템(주로 DBMS)

데이터의 목적론적 특징

- 인식체계에서 어떤 사실에 대해 '데이터'라고 인식하게 되는 객체가 갖고 있는 인식주체에게 필요한 존재 목적
- 하나의 서비스 혹은 활용(데이터 분석)을 위해 데이터가 존재하고 있다는 인식에서부터 출발
- 인식의 주체가 데이터에서 목적에 맞는 특징을 찾아내는 것
- 데이터는 인식 주체의 관점에 따라 여러 종류의 데이터로 인식될 수 있음
- 그 과정에서 원본 데이터의 속성분리, 다른 데이터와의 병합이 발생해 하나의 새로운 데이터 객체가 생성

수집 데이터의 특징 정의

- 수집에서 데이터의 특징은 재생산
- 수집에서 일어나는 모든 과정은 데이터를 활용하기 위한 재생산 과정
- 재생산은 원천 데이터를 탐색・수집・정제・저장하는 과정을 거쳐 새로운 데이터를 생산하는 것을 의미
- 수집에서 데이터는 원본 데이터와의 재생산 데이터의 관계를 고려할 필요가 있음
- 재생산된 데이터는 가역적 데이터와 불가역적 데이터로 구분

가역적, 불가역적 데이터

- 가역적 데이터
  - 생산된 데이터의 원본으로 일정 수준 환원이 가능한 데이터
  - 데이터 원본의 내용이 변경되었을 경우 변경 이력에 대한 추적이 가능
  - 원본 데이터의 흔적이 재생산된 데이터에도 남아 있음
- 불가역적 데이터
  - 생산된 데이터의 원본으로 환원이 불가능한 데이터
  - 원본 데이터와는 전혀 다른 형태로 재생산되기 때문에 재생산된 데이터의 원본 데이터에 대한 추적이 불가능
  - 원본 데이터의 내용이 변경되었더라도 재생산된 데이터에는 아무런 영향을 미치지 않음

구분	가역적 데이터	불가역적 데이터
환원 가능성	가능	불가능
의존성	원본 데이터에 의존	원본 데이터에 독립적
원본 데이터와의 관계	1:1	1:N 혹은 N:1
처리 과정	탐색	병합
활용 분야	데이터 웨어하우징, 로그 수집	소셜 분석, 텍스트 마이닝

빅데이터 처리에서 수집 과정의 역할

- 전형적인 프로젝트는 제안요청에 따른 분석→ 설계→ 구현→ 테스트→ 이행 과정에 따라 진행
- 빅데이터 관련 프로젝트와 전형적인 프로젝트의 차이점
  - 데이터의 생산 방식과 데이터의 생산 주체
  - 빅데이터 관련 프로젝트는 데이터의 생산 주체와 방식이 데이터의 수집이라는 측면에서 이뤄짐
  - 전형적인 프로젝트에서 데이터 생산의 주체는 프로젝트의 요구사항에 이미 정의되어 있기 때문에 다른 주체로의 교체가 불가능
  - 빅데이터 프로젝트에서는 생산의 주체가 아닌 이미 생산된 데이터를 가져오는 것이 중요하기 때문에 데이터의 생산 과정과 생산의 주체는 크게 중요하지 않음
  - 어떤 데이터를 가져올 것인가와 생산된 데이터를 수집하는 과정의 안정성이 가장 큰 고려사항
  - 수집 가능 여부, 수집된 데이터를 테스트하기 위해서는 사전에 원천 데이터를 탐색하는 과정이 필요
  - 데이터의 생산주체와 재생산된 데이터의 관리 주체가 다르기 때문에 데이터의 안정적 수집을 위해서는 반드시 수집 대상에 대한 관리정책이 필요

빅데이터 프로젝트 vs. 일반 프로젝트

	빅데이터 프로젝트	일반 프로젝트
생산 주체의 결정	상대적으로 중요하지 않고 프로젝트 실행 단계에서 누가 무엇을 생산했는지에 대한 검토만 수행	중요한 요소이며 프로젝트 실행 이전에 먼저 결정
생산 주체의 종류	로그 데이터 같은 경우 머신이 데이터의 생산 주체이듯 다양한 생산주체가 있음	일반적으로 생산 주체는 정보 서비스를 사용하는 사용자
구현 형태	수집 기술로 구현	입력 UI로 구현
분석 과정	기술검토 및 적용 필요	업무 프로세스 정립 필요
설계 과정	다른 처리 과정의 아키텍처에 영향을 미침	다른 처리 과정의 아키텍처의 일부분

데이터 수집 정의

- 서비스 활용에 필요한 데이터를 시스템의 내부 혹은 외부에서 주기성을 갖고 필요한 형태로 수집하는 활동
- 데이터 수집 정의 요소
  - 서비스 활용
  - 데이터 위치
  - 주기성
  - 수집 데이터의 저장 형태

서비스 활용

- 서비스 활용에서 데이터 수집의 역할은 서비스의 품질을 결정할 뿐 아니라 서비스의 생명주기에도 영향을 미침
- 어떤 서비스를 할 것인지 결정했으면 먼저 수집할 원천 데이터 탐색 필요
- 데이터 탐색 과정에서 고려해야 할 사항
  - 수집의 난이도적 측면과 비용적 측면
  - 데이터 수집의 안정성

데이터 위치

- 수집 데이터의 위치에 따라 구축하려는 시스템 내부 혹은 외부의 데이터로 구분
- 내부시스템에서 데이터를 가져올 경우
  - 대부분 정형 데이터인 경우가 많음
  - 외부 데이터 수집과 비교하면 수집에 드는 비용이나 난이도 측면에서 훨씬 유리
  - 서비스의 생명주기 관리에 용이
- 외부 데이터일 경우
  - 정형 데이터는 협약된 곳에서 수집하는 데이터를 제외하고는 대부분이 비정형 데이터
  - 수집에 드는 비용이나 난이도가 내부 데이터 수집과 비교하면 상대적으로 어려움
  - 외부 환경은 내부 시스템과는 다르게 통제가 불가능하기 때문에 서비스의 수명주기 관리가 원천 데이터에 의존할 경우 수집이 불가능
  - 원천 데이터를 수집할 수 없을 경우에 대비한 서비스 관리 정책이 별도로 필요

주기성

- 데이터 수집의 주기성을 나눌 때 일반적으로 배치(일괄) 혹은 실시간이냐에 따라 주기를 결정 (서비스의 활용적인 측면이지 원본 데이터의 생명주기에 관한 것이 아님)
- 데이터의 수집은 다른 곳에서 데이터를 생성한 것을 복제하는 행위
- 주기를 갖고 계속 수집하는 행위인지 아니면 한 번의 수집으로 끝나는 과정인지를 먼저 고려해야 함
- 일회성 수집이 아니고 일정한 주기를 기준으로 수집 활동이 진행되어야 한다면, 서비스의 종류 및 환경에 따라 실시간으로 수집할 것인지 아니면 배치 작업을 통해 수집할 것인가를 결정해야 함

수집 데이터의 저장 형태

- 서비스 활용에 대해 수집 대상 데이터의 위치・주기・수집방법이 결정됐으면, 수집한 데이터를 어떠한 형태로 저장할 것인지를 고려해야 됨
- 빅데이터 처리에서 일반적으로 사용되는 저장소는 분산 파일 시스템이지만, 수집된 데이터를 가공・처리하기 위해서 DBMS가 사용될 수도 있고 서비스를 DBMS를 통해 제공할 수도 있으므로 서비스 환경에 맞는 아키텍처를 설계해야 함
- 원본 데이터의 형태에 따라 사용되는 수집 기술이 다르고 수집되는 형태도 다르기 때문에 데이터 저장소의 아키텍처 설계 시 파일 시스템, 관계형 데이터베이스, 분산처리 DBMS 등을 고려해야 함

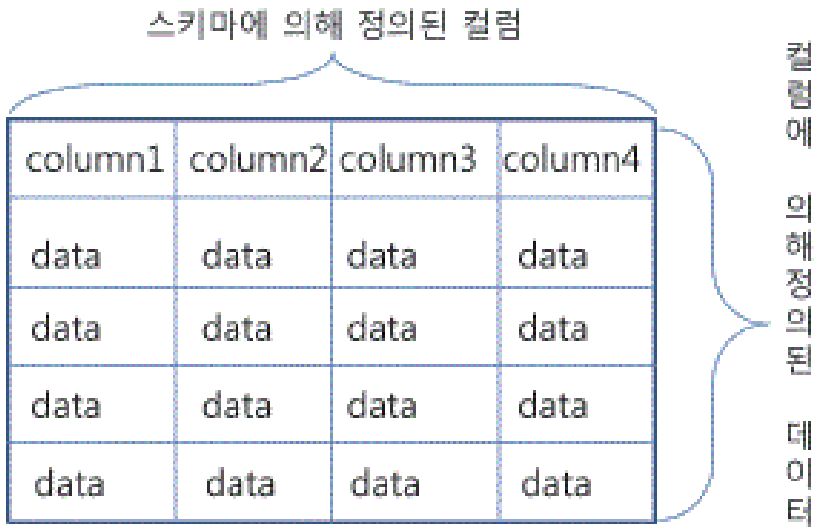
2. 수집 데이터 분류

수집 대상 데이터의 종류

- 수집 데이터의 저장・관리되는 형태에 따른 분류
  - 정형 데이터
  - 반정형 데이터
  - 비정형 데이터
- 수집 데이터의 저장 위치에 따른 분류
  - 내부 데이터
  - 외부 데이터
- 수집 데이터의 생산 주체에 따른 분류

정형 데이터Structured Data

- 관계형 데이터베이스 시스템의 테이블과 같이 고정된 컬럼에 저장되는 데이터와 파일
- 지정된 행과 열에 의해 데이터의 속성이 구별되는 스프레드시트 형태의 데이터
- 관계형 데이터베이스 시스템의 정형 데이터를 비정형 데이터Unstructured Data와 비교할 때 가장 큰 차이점은 데이터의 스키마를 지원하는 것
- 데이터의 스키마 정보를 관리하는 DBMS와 데이터 내용이 저장되는 데이터 저장소로 구분



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

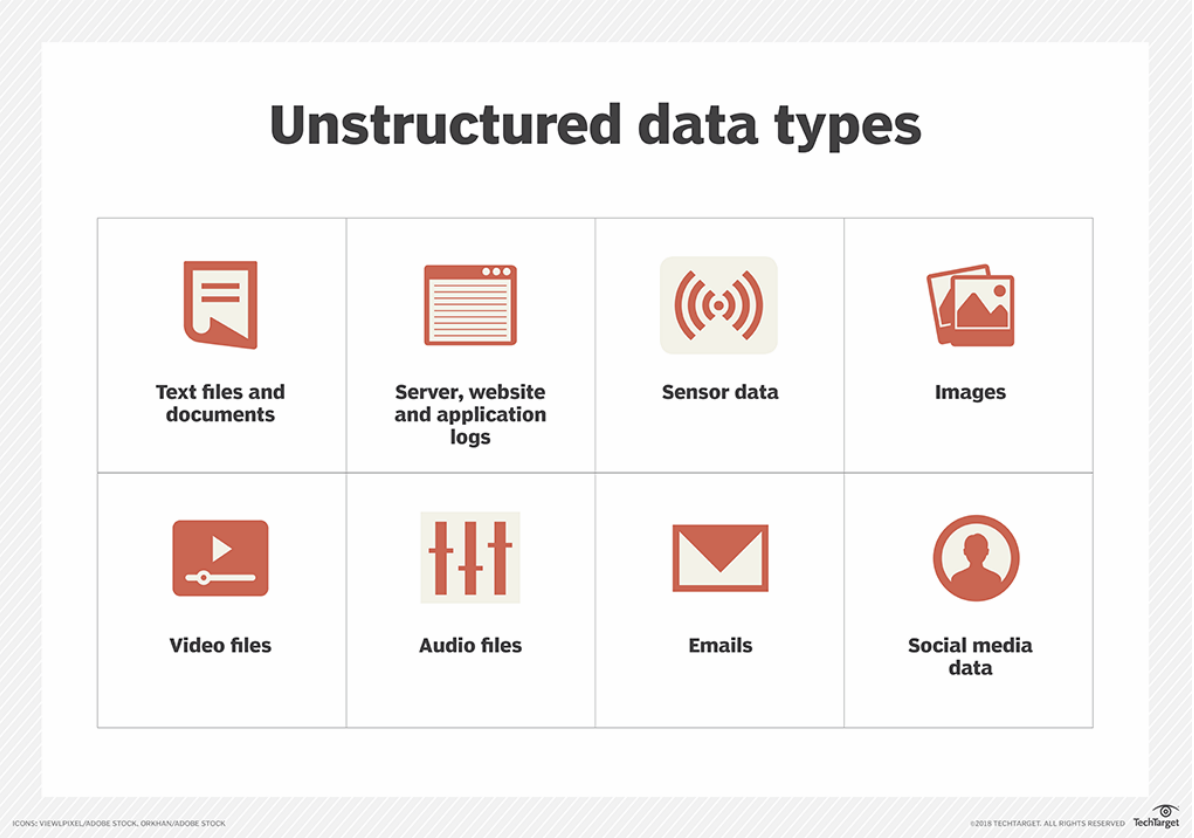
반정형 데이터Semi-Structured Data

- 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 갖고 있음
- 일반적으로 파일 형태로 저장
- 데이터 내부에 데이터 구조에 대한 메터 정보를 갖고 있기 때문에 어떤 형태를 가진 데이터인지를 파악하는 것이 필요
- 데이터 내부에 있는 규칙성을 파악해 데이터를 파싱할 수 있는 파싱 규칙을 적용
- 반정형 데이터 예
  - URL 형태로 존재 – HTML
  - 오픈 API 형태로 제공 - XML, JSON
  - 로그형태 - 웹로그, IOT에서 제공하는 센서 데이터


```
[{  "Sepal.Length" : 6.8,
    "Sepal.Width" : 3.2,
    "Petal.Length" : 5.9,
    "Petal.Width" : 2.3,
    "Species" :  "virginica"  },
   {  "Sepal.Length" : 6.7,
      "Sepal.Width" : 3.3,
      "Petal.Length" : 5.7,
      "Petal.Width" : 2.5,
      "Species" :  "virginica"  }]
```

비정형 데이터Unstructured Data

- 데이터 세트가 아닌 하나의 데이터가 수집 데이터로 객체화
- 언어 분석이 가능한 텍스트 데이터나 이미지, 동영상 같은 멀티미디어 데이터가 대표적인 비정형 데이터
- 웹에 존재하는 데이터의 경우 html 형태로 존재하여 반정형 데이터로 구분할 수도 있지만, 특정한 경우 텍스트 마이닝을 통해 데이터를 수집하는 경우도 존재하므로 명확한 구분은 어려움
- 비정형 데이터 예
  - 이진 파일 형태: 동영상, 이미지
  - 스크립트 파일 형태: 소셜 데이터의 텍스트



https://searchenterpriseai.techtarget.com/feature/Convert-unstructured-data-to-structured-data-with-machine-learning

데이터 형태별 수집 및 아키텍처 구성 난이도

형태	특징	난이도
정형 데이터	▪ 내부 시스템인 경우가 대부분이라 수집이 쉬움 ▪ 파일 형태의 스프레드시트라도 내부에 형식을 가지고 있어 처리가 쉬움 ▪ CRUD가 일어나는 일반적인 아키텍처 구조로 구성	하
반정형 데이터	▪ 보통 API 형태로 제공되기 때문에 데이터 처리 기술이 요구 ▪ 데이터의 메타구조를 해석해 정형 데이터 형태로 바꿀 수 있는 아키텍처 구조로 수정 필요	중
비정형 데이터	▪ 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움 ▪ 텍스트나 파일을 파싱해 메타구조를 갖는 데이터셋 형태로 바꾸고 정형 데이터 형태의 구조로 만들 수 있도록 아키텍처 구조 수정 필요	상

데이터 형태별 잠재 가치 비교

형태	특징	잠재 가치
정형 데이터	▪ 내부 데이터의 특성상 현실적 가치의 한계상 활용측면에서 잠재적 가치는 상대적으로 낮음	하
반정형 데이터	▪ 데이터의 제공자가 선별해 제공하는 데이터로 잠재적 가치는 정형 데이터 보다 높음	중
비정형 데이터	▪ 수집주체에 의해 데이터에 대한 분석이 선행되었기 때문에 목적론적 데이터 특징이 가장 잘 나타나는 데이터 ▪ 일단 수집이 가능하면 수집주체에게는 가장 높은 잠재적 가치를 제공	상

수집데이터의 위치에 따른 분류

- 수집하려는 데이터를 저장된 위치에 따라 분류하면 동일한 시스템계에 저장되는 내부 데이터와 외부시스템에 저장된 외부 데이터로 나눌 수 있음
- 실시간 처리에서는 저장되는 위치가 아니라 발생하는 위치에 따라 내부 데이터와 외부 데이터로 나눌 수 있음
- 수집시 내부와 외부로 데이터를 분류하는 가장 큰 이유는 원천 시스템과 연계를 위한 인터페이스의 기술적 방법 및 정책적 차이점 때문

내부 데이터Internal Data

- 수집하는 원천 데이터의 데이터 저장소가 내부시스템에 있는 데이터를 의미
- 단순히 물리적 데이터 저장소 외에도 내부데이터와 외부 데이터의 가장 큰 구별점은 데이터 제공자와 상호 협약에 의한 의사소통이 가능하다는 점
- 원천데이터와 수집한 데이터가 동일 시스템계에 저장돼 있으므로 원천데이터가 외부에 있는 경우와 비교했을 때 상대적으로 기술적 제약도 적은 편
- 데이터의 수집주기 및 방법은 데이터 제공자(또는 기관)와의 협약을 통해 제공 받음
- 수집성공 여부에 대한 별도의 인터페이스를 설정해 수집 실패한 데이터에 대해 재수집이 가능하도록 구현할 수 있음

외부 데이터External Data

- 수집하는 원천 데이터의 데이터 저장소가 외부 시스템에 있는 데이터를 의미
- 일반적으로 내부 데이터와 가장 큰 구별점은 데이터 제공자와 협약된 관계가 아니면 상호 의사소통이 불가능하다는 점
- 데이터 수집을 위해 수집주기 및 방법에 관한 분석이 필요
- 외부 데이터의 인터페이스 방법은 수집할 항목을 분석해 수집 시스템을 설계하는 것
- 협약이 되지 않은 시스템의 경우 수집 실패 시의 대안을 마련해야 함
- 가능한 데이터의 전처리 과정 없이 원본 데이터를 수집 후, 수집 시스템에서 처리를 할 수 있도록 인터페이스를 설계하는 것이 바람직

데이터 위치별 수집 및 아키텍처 구성 난이도

위치	특징	난이도
내부	▪ 데이터의 저장소가 내부에 있으므로 해당 소스 데이터 담당자와 의사소통이 원활하기 때문에 수집 난이도가 외부 데이터와 비교해 낮음 ▪ 대부분 정형 데이터이므로 일반적인 CRUD처리 아키텍처와 같은 구성이 가능	하
외부	▪ 외부 소스의 경우 해당 소스 데이터 담당자와 의사소통이 어려워 상대적으로 수집 난이도가 높음 ▪ 대부분 비정형, 반정형 데이터 형태로 일반적인 아키텍처 구성에 반정형, 비정형 데이터를 처리할 수 있는 아키텍처를 추가해야 함	상

데이터 위치별 잠재적 가치 비교

위치	특징	잠재 가치
내부	▪ 내부 데이터의 특성과 현실적 가치의 한계상 활용 측면에서 잠재적 가치는 상대적으로 낮음	보통
외부	▪ 데이터의 제공자가 선별해 제공하는 데이터나 수집주체에 대한 분석이 이루어진 후 수집을 하는 데이터이기 때문에 데이터의 목적론적 특징이 가장 잘 나타나는 데이터 ▪ 내부 데이터와 비교할 경우 상대적으로 잠재적 가치가 높음	높음

3. 데이터 수집 절차

데이터 수집 프로젝트의 특성

- 일반 업무 프로젝트는 데이터의 확보가 프로젝트의 품질과 성패를 좌우하지는 않으며, 데이터의 관리적인 측면에 더 집중
- 데이터를 수집하는 절차가 포함된 프로젝트일 경우 데이터의 확보가 프로젝트의 품질과 성패를 결정하는 경우가 대부분
- 데이터를 수집하기 위해서는 수집절차를 설계하고 충분한 테스트를 걸쳐 수집을 진행해야 함
- 데이터 수집은 서비스의 품질을 좌우하기 때문에 데이터 수집 절차 중 심각한 문제가 발생한다면, 프로젝트 전체를 다시 설계해야 할 경우도 발생

프로젝트의 프로세스 정의

프로세스 분류	프로젝트 분류	특징
데이터 입력	일반 프로젝트	▪ 사용자가 데이터 발생의 주체 ▪ 데이터 입력자가 데이터에 대한 오너십을 가짐 ▪ 데이터는 트랜잭션 하나에 한 건씩 발생
	수집 관련 프로젝트	▪ 수집 프로세스 자체가 데이터 발생의 주체 ▪ 수집 정책에 관련된 기술에 대한 데이터 오너십이 존재 ▪ 데이터는 프로세스 하나에서 여러건 발생 가능
데이터 처리	일반 프로젝트	▪ 입력된 데이터의 사전처리 및 사후처리 과정이 거의 존재하지 않음 ▪ 원본 데이터의 변경이 이뤄지지 않음
	수집 관련 프로젝트	▪ 대부분 데이터의 전처리 과정이 필요 ▪ 수집한 데이터에 대해 후처리(다른 데이터와 병합)가 이루어짐
서비스 제공	일반 프로젝트	▪ 처리된 데이터를 변형 없이 제공 ▪ 서비스 제공은 데이터의 변형이 없으므로 데이터 생산과 데이터 출력이 일치 ▪ OLTP 기반의 프로젝트가 주를 이룸
	수집 관련 프로젝트	▪ 처리된 데이터를 분리・병합해 새로운 인사이트를 제공하는 서비스가 주를 이룸 ▪ 재생산의 과정을 거쳐 생성된 데이터가 출력 ▪ 데이터 분석 관련 프로젝트가 주를 이룸

프로세스별 비교

프로세스 분류	프로젝트 분류	특징
데이터 입력 ↓ 데이터 처리	일반 프로젝트	▪ 입력 받은 데로 데이터 처리를 진행 ▪ 데이터 처리시 정규화가 중요한 요소
	수집 관련 프로젝트	▪ 데이터 수집과 데이터 처리는 누락 되는 데이터에 대해 사전처리 작업이 필요 ▪ 데이터 처리시 수집 데이터의 가공이 발생
데이터 처리 ↓ 서비스 제공	일반 프로젝트	▪ 데이터 처리는 처리된 데이터와 원본 데이터가 정규성을 가짐 ▪ 서비스 제공에서도 처리된 데이터와 제공하는 서비스항목이 순수하게 데이터의 조회 형태로만 제공
	수집 관련 프로젝트	▪ 서비스 제공의 대부분이 분석과 관련이 있음 ▪ 데이터 처리는 원본 데이터의 변환이 일어나고 서비스 제공 형태에 따라 처리 방법의 수정이 발생할 수 있음
데이터 입력 ↓ 서비스 제공	일반 프로젝트	▪ 입력한 데이터와 서비스에 제공되는 데이터가 가역적 선형 관계 ▪ 서비스 제공 데이터로부터 입력 데이터로의 역추적이 가능 ▪ 데이터 입력이 중단돼도 서비스 제공은 가능
	수집 관련 프로젝트	▪ 서비스에 제공되는 데이터는 수집한 데이터를 내부 처리과정을 거쳐 서비스 목적에 맞춰 변형(재생산)된 새로운 데이터 ▪ 서비스 제공과 데이터 수집은 하나의 업무처리 선상에 있으므로 서비스의 제공은 데이터 수집에 영향을 받음

데이터 수집절차 설계

데이터 선정

- 수집 가능성
- 보안 문제
- 데이터의 정확성
- 수집 난이도
- 수집 비용

수집 세부계획 수립

- 수집 데이터의 위치 파악
- 데이터 유형 파악 및 수집 방법 적용
- 수집 계획서 작성

테스트 수집 실행

- 수집 테스트 기술적 검토
- 수집 테스트 업무적 검토

데이터 선정

- 수집을 포함하는 프로젝트 진행시 데이터의 선정은 프로젝트의 품질을 결정할 뿐 아니라 성공 여부, 진행 여부에 영향을 미치는 핵심업무
- 수집 대상 데이터를 분석할 수 있는 업무 지식을 보유하거나 수집 노하우를 보유한 전문가가 수행하는 것이 바람직
- 데이터의 목적론적 특성에 대해 이해했다면 전문 기술을 보유하지 않은 사람들도 경험을 통해 진행할 수 있을 것
- 먼저 수집 대상 데이터를 선정하기 위해 살펴 볼 것은 수집 가능성과 활용 데이터의 보안 문제, 그리고 데이터의 정확성

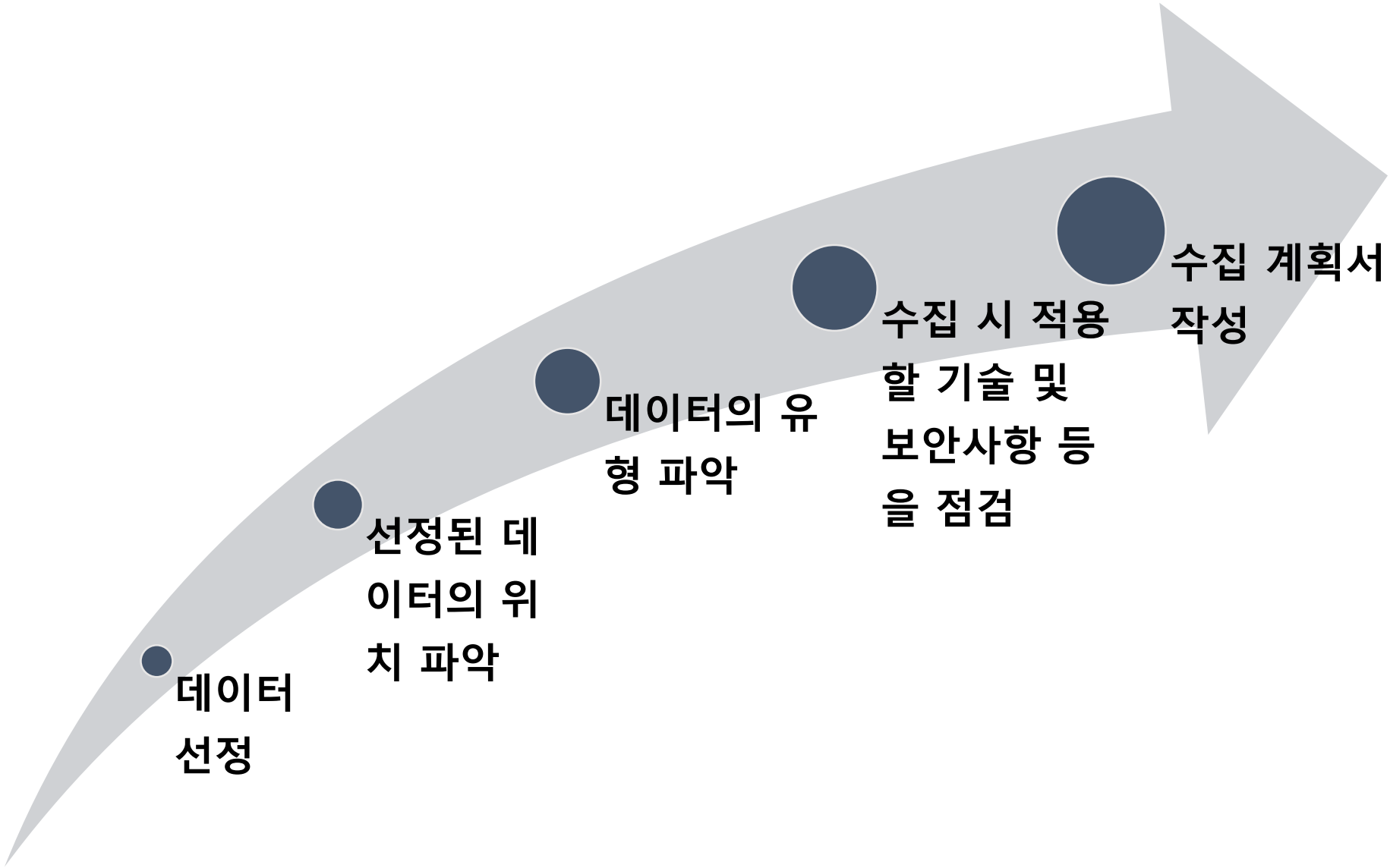
데이터 수집의 고려사항

- 수집 가능성
  - 데이터 선정시 가장 우선적으로 고려해야 할 사항
  - 서비스 활용에 아무리 좋은 데이터가 있다고 가정하더라도 수집이 불가능하거나 통제 불가능한 주기를 가지고 있다면 서비스 활용을 원천 데이터의 정책에 의존하게 되므로 바람직하지 않음
  - 수집이 아무리 용이하더라도 서비스 활용측면에서 데이터를 활용하기 위한 전처리・후처리에 비용이 많이 들어가게 되면 좋은 데이터 선정이라 할 수 없음
- 개인정보보호 및 저작권 문제
  - 수집된 데이터에 대해 개인정보보호 문제나 저작권에 대한 문제 발생시 서비스 활용에 대해 심각한 문제가 발생하므로 반드시 살펴보아야 함

데이터 수집의 고려사항

- 데이터의 정확성
  - 수집한 데이터의 정확성은 서비스의 활용목적에 세부항목이 정확히 존재하는가에 대해 검토 필요
  - 수집목적에 맞는 데이터를 수집하기 위해서는 사전처리 과정이 필요하고 수집한 데이터의 사후처리 방안 필요
- 수집 난이도
  - 수집 난이도는 데이터 수집 및 처리에 들어가는 구축비용Ongoing Cost이 많이 들어갈 경우와 데이터 수집의 분석・설계와 필요한 데이터를 얻기 위해 많은 정제 과정이 필요할 경우로 구분하여 대안을 고려
  - 정성적 기준으로 주로 비용으로 직접 산출이 어려운 경우 수집난이도 측면에서 트래픽량과 저장처리 장치의 용량 등이 고려대상이며, 수집 대상의 대안을 찾아야 함
- 수집비용
  - 수집비용은 데이터를 획득하기 위해 직접적으로 들어가는 획득비용
  - 정량적 기준으로 적용된 수집기술에 들어가는 비용이 발생할 경우에는 수집기술에 대한 검토가 필요

수집 세부계획 수립



수집 데이터의 위치 파악

- 수집할 데이터가 내부 혹은 외부에 있느냐에 따라 소유기관과의 필요한 협의를 진행

내부	▪ 내부 시스템과 데이터 연계 가능여부를 파악, 데이터 종류 및 수집주기 인터페이스 정의서 작성
외부	▪ Open API는 개방하는 데이터의 종류 및 형태를 파악해 데이터의 양과 트래픽의 정도 확인 ▪ 연계방식 및 절차 수집기술 적용방안 검토, 소스데이터 시스템의 데이터 개방 정책 파악 ▪ 크롤링을 통해 데이터를 가져올 경우 외부 시스템의 수명주기 및 저작권 문제 등 수집가능 ▪ 여부 체크, 수집기술 적용방안 검토, 서비스 종료시 다른 소스에 대한 대안 검토

데이터 유형 파악 및 수집방법 적용

- 데이터의 유형 파악은 데이터 수집기술을 결정하는 중요한 과정

형태	특징	존재 형태
정형 데이터	데이터스키마 지원	RDB, 파일
반정형 데이터	데이터 내에 메타 속성이 존재	파일
비정형 데이터	분석이 가능한 텍스형 파일과 데이터 형태가 아닌 이미지나 동영상 파일로 존재	RDB, 파일

형태	데이터유형	수집방법
RDB	정형 데이터, 비정형 데이터	DBtoDB, ETL, RDB 벤더제공 드라이버
파일	반정형 데이터	크롤링, Open API, FTP, HTTP

수집 계획서 작성

- 데이터 소스 구성요소

항목	작성 내용	고려사항
소스 위치	▪ 내부 시스템일 경우 특정 RDB의 IP, PORT 등이 포함될 수 있으며, 외부 시스템일 경우 URI 등이 기술	여러 소스가 있을 경우 중요한 소스별로 모두 기술돼야 함
데이터 유형	▪ 물리적으로 존재하는 데이터의 유형을 파일 종류, RDB일 경우 DBMS 종류 등을 기술	데이터 유형이 혼합된 경우 의존관계가 기술돼야 함
인터페이스	▪ 수집하는 항목의 세부내용에 대해 인터페이스 요소에 대해 기술	
데이터 담당자	▪ 소스 데이터의 데이터 담당자와 연락처를 기술 ▪ 담당자를 알지 못하는 경우 기관명 혹은 대표 URL만 기술 가능	
협약 내용	▪ 데이터 원천 담당자와 협약한 내용을 기록 ▪ 협약사항은 별도의 첨부 문서로 관리될 수 있음	법적 검토가 필요

수집 계획서 작성

- 데이터 수집주기의 구성요소

항목	작성 내용	고려사항
주기설정	주기는 주기성을 갖지 않는 불규칙 수집일 경우와 일정 시간을 정해 수집하는 배치일 경우, 이벤트가 발생할 경우 등으로 배치처리와 실시간 처리를 구분해 기술	수집실패 시 재수집 정책을 수립할 필요가 있음
데이터량	1회 수집 시 발생하는 데이터의 양을 기술하고, 여기에 수집주기를 곱해 서비스 시작일과 서비스의 예상 종료일까지의 총 데이터량을 기술	수집 시스템의 저장소의 용량을 예측
트래픽량	수집 시 발생하는 네트워크 트래픽량을 계산해 작성	과도한 트래픽발생 시 해결책이 필요

수집 계획서 작성

- 데이터 수집방법의 구성요소

항목	작성 내용	고려사항
적용기술	수집에 필요한 적용기술을 기술 일반적으로 하나의 기술만 선택되는 것이 아니므로 수집 프로세스별로 나누어 사용되는 기술의 이름, 버전을 기술	처리과정에 대해 과정별로 나누어 기술해야 함
데이터 사전처리	사전 데이터 처리가 필요한 경우 사전 처리 작업에 대해 기술	사전처리 규칙을 명시해야 함
데이터 사후처리	데이터 수집 후 사후처리가 필요한 경우 사후처리 방법에 대해 명시해 기술	

테스트 수집 진행

- 수집 계획서가 완성되었으면 테스트 수집 진행
- 테스트 수집은 수집 관련 프로젝트에서 단순히 수집 기능의 테스트만 진행하는 것이 아님
- 데이터 선정 시 고려했던 수집 가능성, 보안문제, 데이터의 정확성을 만족시키는지 검증
- 수집된 데이터의 서비스 활용 측면까지 검토

수집 테스트 기술적 검토

- 먼저 원하는 데이터가 제대로 수집됐는지 기술적 방법의 적용은 최적의 방법으로 적용했는지 등에 대해 검토
- 기술적 검토는 수집한 데이터 세트의 누락여부, 소스 데이터와의 비교, 데이터 활용 측면에서의 데이터의 정확성에 대해 검토해야 함

검토사항	검토방법
데이터세트 누락	▪ 원본 데이터 요청 후 확인 ▪ 재수집을 통해 누락데이터 세트 확인
소스 데이터와 비교	▪ 파일일 경우 사이즈 비교 ▪ 수집한 데이터와 개수 비교
데이터의 정확성	▪ 서비스 활용에 수집한 데이터의 사후처리가 필요한지 확인

수집 테스트 업무적 검토

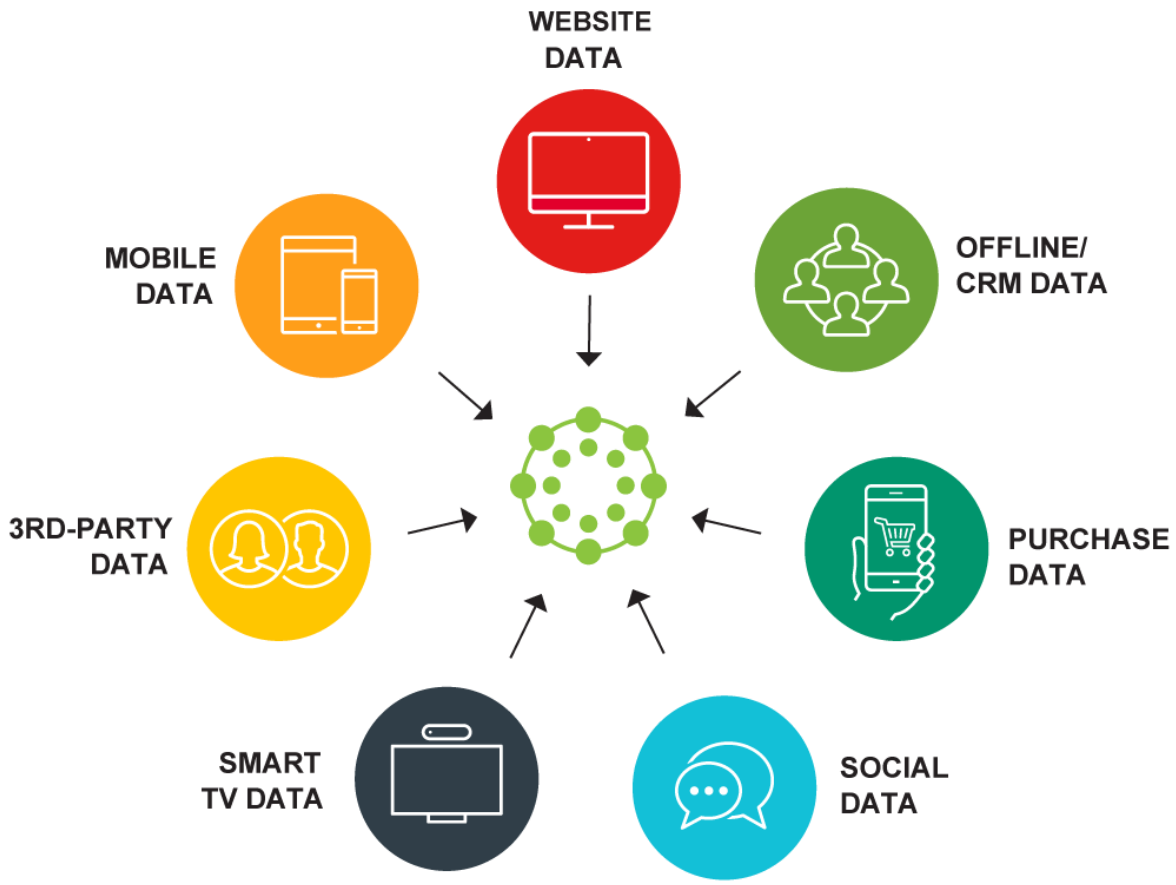
- 수집한 데이터에 대한 개인정보 보안 , 수집 데이터의 저작권 관련 사항 등 업무적 검토
- 협약기관에 많은 트래픽을 발생시킬 경우, 제약 사항들을 다시 한번 검토

검토사항	검토방법
보안사항	수집 데이터에 개인정보 수집 등 보안 사항이 없는지 검토
저작권	수집한 데이터가 저작권 등 법적문제가 없는지 검토
대용량 트래픽	수집 대상 시스템에 트래픽을 많이 발생시키는 요소가 없는지 검토

4. 데이터 수집 방법 및 기술

수집방법의 분류

- 일반적 수집 데이터의 형태와 종류
  - 크롤링
  - ETL
  - 로그수집,
  - ftp
  - http
  - RDB
  - …



데이터와 서비스 활용

- 서비스 활용에 필요한 모든 데이터는 디지털화한 결과물
- 디지털 데이터는 형태가 어떤 것이든 각각의 해석기가 존재
- 해석기를 통해 보통은 텍스트 형태의 가시적 데이터로 혹은 음성이나 동영상 같이 서비스 이용자가 인지할 수 있는 형태로 서비스를 제공
- 파일 형태의 스프레드시트라면 스프레드시트 지원 프로그램을 통해 데이터를 해석
- 데이터베이스에 저장된 정형 데이터라면 데이터베이스 벤더가 제공하는 API를 통해 해석

데이터 형태별 서비스 활용

데이터 형태	저장 형태	서비스 활용	활용 API
정형 데이터	DBMS	DBMS 스키마를 읽어 가지적(텍스트) 데이터 세트 형태로 제공	DBMS
	이진 파일	이진코드화한 데이터를 해석해 가시적 데이터 세트 형태로 제공	응용소프트웨어
반정형데이터	스크립트 파일	스크립트 형태의 파일을 브라우저나 응용소프트웨어가 해석해 가시적(텍스트) 데이터 형태로 제공	브라우저, 응용소프트웨어 (JAVA, DBMS)
비정형데이터	이진 파일 스크립트 파일	이진화한 형태의 데이터를 해석해 인지적(음성, 영상, 텍스트) 데이터 형태로 제공	응용소프트웨어 (동영상 플레이어)

데이터와 서비스 활용

- 여러 형태의 데이터가 각자의 서비스 활용 목적에 맞게 API를 통해 서비스를 제공
- 데이터 수집과 관련된 프로젝트에서 서비스 활용에 필요한 기능을 살펴보면 데이터를 수집하고 수집된 데이터를 가공해 재생산된 데이터를 가시적 형태로 제공하는 것
- **데이터 수집**: 데이터의 형태나 종류에 관계 없이 수집 가능 (단, 데이터의 형태나 종류에 따라 수집방법이 결정)
- **데이터의 재생산**: 수집한 데이터를 가공 재생산할 수 있어야 하며, 수집방법에는 데이터를 가공・재생산하는 과정을 포함
- **가시적 형태로 서비스 제공**: 가시적 형태로 서비스를 제공하고 데이터를 효율적으로 활용할 수 있도록 저장해야 하며, 수집방법에는 정형/반정형 형태로 데이터를 저장하는 기능을 제공

수집방법 정의

- 수집방법은 데이터의 종류와 형태별로 분류되고 수집 시스템에 저장・처리할 수 있는 기능과 정형/반정형 데이터의 형태로 저장할 수 있는 기능이 있어야 함
- 수집 데이터의 위치는 수집방법에 영향을 미치지 않음
- 수집방법에 영향을 미치는 것은 데이터의 형태와 종류
- 내부/외부를 구분해 나누는 실익은 업무적 제약 사항 때문에 수집방법별로 존재하는 수집기술들에 영향을 미침
- 내부에서 수집을 할 경우에는 DBMS에 저장된 정형 데이터가 주를 이루고, 외부에서 수집할 경우에는 스크립트 형태로 제공하는 반정형 데이터가 주를 이루기 때문에 실무에서 데이터 수집계획을 수립할 경우 내부와 외부를 구분하는 것이 데이터 수집의 비용적인 측면과 관리적 측면에서 유리

수집방법의 종류 (외부)

위치	형태	종류	수집방법 정의
외부	정형 데이터	DBMS	DBMS 벤더가 제공하는 API를 통해 정형 데이터에 접근해 데이터를 수집하고 시스템에 저장
		이진 파일	ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드하고 해당 파일의 API를 통해 데이터 처리
	반정형 데이터	스크립트 파일	http 프로토콜을 사용해 파일의 텍스트를 스크랩하고 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
		이진 파일	스트리밍을 사용해 파일의 텍스트를 스크랩하고 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
	비정형 데이터	이진 파일	ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드하고 해당 파일을 API를 통해 데이터 처리
		스크립트 파일	http 프로토콜을 사용해 파일의 텍스트를 스크랩하고 내부 처리에서 텍스트를 파싱해 데이터 처리

수집방법의 종류 (내부)

위치	형태	종류	수집방법 정의
내부	정형 데이터	DBMS	DBMS 벤더가 제공하는 API를 통해 정형 데이터에 접근해 데이터를 수집하고 시스템에 저장
		이진 파일	ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드 하고 해당 파일의 API를 통해 데이터 처리
	반정형 데이터	스크립트 파일	http 프로토콜을 사용해 파일의 텍스트를 스크랩하고 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
		이진 파일	스트리밍을 사용해 파일의 텍스트를 스크랩하고 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
	비정형 데이터	파일	ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드 하고 해당 파일을 API를 통해 데이터 처리

데이터 종류별 연동방법

데이터의 종류	수집 대상 데이터의 구현형태	연동방법	내부처리
DBMS	DBMS에 저장된 데이터	소켓	저장
반정형 이진 파일	로그 형태로 센서, 서버 등 머신이 발생하는 데이터	스트리밍	저장
비정형 이진 파일	텍스트 형태의 파일, 동영상 파일, 이미지 파일	ftp	저장・파싱
스크립트 파일	웹상에 html, xml, JSON 형태로 존재	http	저장・파싱

연동 방법에 따른 수집 방법

- 원본 데이터 요청 후 확인
- 소켓Socket 통신으로 연동하는 DBMS 수집방법
  - 주로 DBMS에서 벤더가 제공하는 드라이버를 통해 데이터를 연동
- 스트리밍 방식으로 연동하는 로그 데이터, 센서 데이터 수집방법
  - 주로 시스템에서 발생하는 데이터로 스트리밍 방식의 연동방법은 tcp, 블루투스, RFID 등 여러 가지 통신 프로토콜이 존재
- ftp 프로토콜을 사용하는 이진 파일 수집방법
  - 비정형 데이터를 수집할 때 필요한 방법으로 데이터 수집 후 수집한 데이터를 그대로 사용할 경우도 있지만, 서비스 활용을 위해 데이터의 파싱이 필요
- http 프로토콜 사용하는 스크립트 파일 수집방법
  - 웹상에 존재하는 데이터는 반정형 데이터로 존재하므로 수집기 내에서 활용 가능한 데이터의 형태로 파싱처리 후 시스템에 저장

수집방법의 기본적 기능요건

- 데이터 수집 정의 요건에 맞도록 수집방법의 기능적인 측면을 주기성, 데이터의 위치 설정을 위한 수집방법에 공통적으로 필요한 환경설정 기능, 서비스 활용을 위해 데이터 처리・전처리・후처리의 요건을 만족시키도록 하는 기능들을 정의

수집기술의 환경 설정기능

수집방법	기능항목	기능항목 요건
DBMS 수집	주기성	수집주기 설정 기능
	위치	수집 데이터의 위치설정 기능
로그/센서수집	주기성	실시간/근실시간 환경설정 기능
	위치	머신으로부터 받는 정보설정 기능
FTP 수집	주기성	수집주기 설정 기능
	위치	수집 데이터의 위치설정 기능
HTTP 수집	주기성	수집주기 설정 기능
	위치	수집 데이터의 위치설정 기능

수집기술의 데이터 처리기능

수집방법	데이터 처리기능	기능항목 설명
DBMS 수집	에이전트 기능	데이터 수집을 처리·관리하는 수집 에이전트 기능
로그/센서수집	에이전트 기능	데이터 수집을 처리·관리하는 수집 에이전트 기능
	콜렉터 기능	대용량 처리시 노드나 전송을 관리하는 콜렉터 기능
FTP 수집	파일전송 기능	ASCII 혹은 BINARY 파일을 전송하는 기능
HTTP 수집	에이전트 기능	데이터 수집을 처리·관리하는 수집 에이전트 기능
	RDB 매핑 기능	처리된 데이터를 수집 시스템의 RDB와 매핑하는 기능

수집기술의 종류

- 수집방법에 필요한 기본적인 기능요건들을 만족시키거나 필요한 기능요건들을 추가해 수집기술을 제작 가능
- 이미 선행돼 명명된 기술이 존재하므로 하나의 수집방법에 여러 개의 수집기술이 존재
- 기술적으로는 큰 차이가 없지만 서비스 활용 용도에 따라서 수집기술이 나뉘기도 함
- 데이터와 서비스 활용
  - HTTP에서 데이터를 수집하는 방법에 RDB 매핑 기능을 추가해 오픈 API 수집기술이 파생
  - 외부 데이터의 HTTP 수집방법은 크롤링 기술이라 미리 명명
  - 로그/센서 데이터를 수집하는 방법이 각각 데이터 처리 에이전트의 구별을 통해 로그 수집기술, 머신2머신 수집기술로 분리

수집방법별 수집기술의 종류

- HTTP 수집
  - 크롤링 기술
  - Open API 수집 기술
- 로그/센서 수집
  - 로그 수집 기술
  - 센서 데이터 수집 기술
- DBMS 수집
- FTP 수집

크롤링 기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 웹문서, 콘텐츠 수집기능 ▪ 수집한 웹문서에서 URL 추출기능 ▪ 머신런닝 기능 ▪ URL 리스트 탐색 기능 ▪ 스크랩한 문서를 DB에 업로드하는 기능 ▪ 병렬 크롤링 기능 ▪ 불가역 데이터를 수집하므로 불필요 데이터를 전처리하는 기능
접근제어기능	▪ 에이전트가 웹 사이트에 접근하는 것을 방지하기 위한 로봇 배제 표준규약 권고안을 준수하도록 robots.txt를 탐지하는 기능

Open API 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 에이전트가 mash-up이 가능하도록 RESTFUL 방식의 API 제공기능 ▪ 콘텐츠 자원에 유일한 URI를 부여하는 기능 ▪ CRUD를 에이전트의 메소드로 제공하는 기능 ▪ 반정형 데이터(XML, JSON, RSS)의 정보 제공방식을 지원 ▪ 수집 데이터의 성공/실패에 대한 응답기능
테이블 매핑기능	▪ 수집한 반정형 데이터의 요소와 수집 시스템의 컬럼에 대해 매핑하는 기능

로그 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 수집한 파일을 Chunk 단위로 전송하는 기능 ▪ 주기/블록 단위로 잘라서 전송하는 기능 ▪ 압축지원 기능 ▪ 수집대상 파일 체크기능
콜렉터기능	▪ 다수의 에이전트로 받은 데이터를 직렬화된 chunk에 분산 파일시스템의 시퀀스 파일로 전송하는 기능 ▪ 트래픽 밸런싱을 자동 조정하거나 수동으로 관리할 수 있는 기능 ▪ 파일 전송 모니터링 기능

로그 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 수집 시스템의 콘텐츠 자원에 접속가능한 표준 API기반의 서비스 제공 ▪ 다양한 외부 애플리케이션을 지원하는 기능 ▪ 센서 데이터 검색 메소드를 지원하는 기능 ▪ 최적 라우팅 경로를 탐색하는 기능
콜렉터기능	▪ 에이전트 모듈에서 수집한 하위 머신의 정보와 센서 데이터를 실시간으로 서버로 전송하는 기능

DBMS 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ DBMS 메타 정보에서 테이블을 선택하는 기능 ▪ DBMS 메타 정보에서 컬럼을 선택하는 기능 ▪ 레코드 단위로 수집해 분산파일시스템으로 전송하는 기능 ▪ 수집의 성공/실패시 응답 기능
클린징기능	▪ 수집한 데이터를 타겟시스템에 맞게 정제하여 로드하는 기능

FTP 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ ACTIVE, PASSIVE 연결에 필요한 통신포트 설정기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 클라이언트 방화벽 운영시 액티브, 패시브 모드 변환기능 ▪ 액티브 연결기능 ▪ 패시브 연결기능 ▪ 파일전송 연결기능
파일전송기능	▪ ASCII, BINARY 파일 전송기능 ▪ Get, put, mget, mput 명령기능 ▪ 수집파일에 대한 무결성 확인기능

Q & A