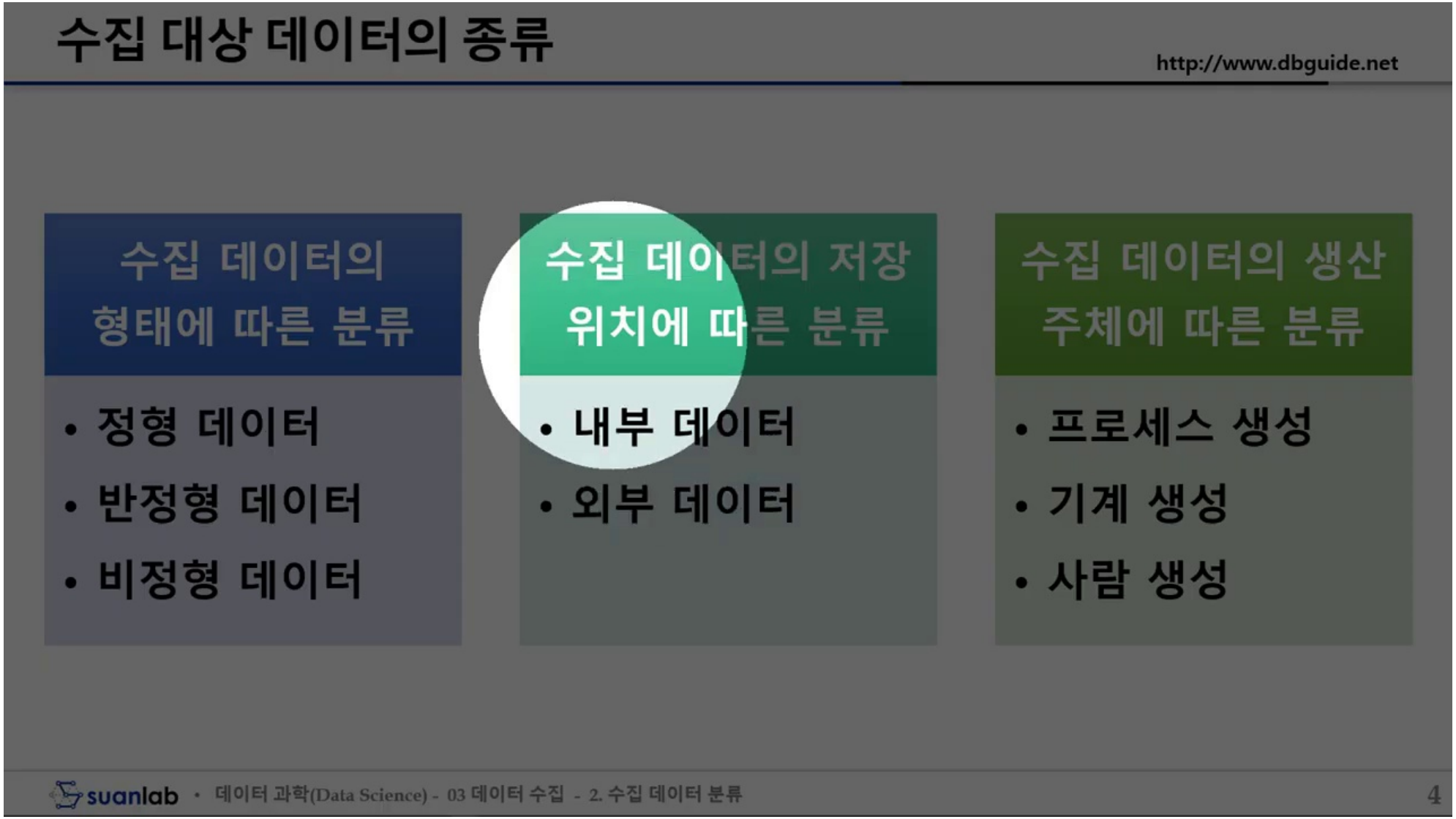
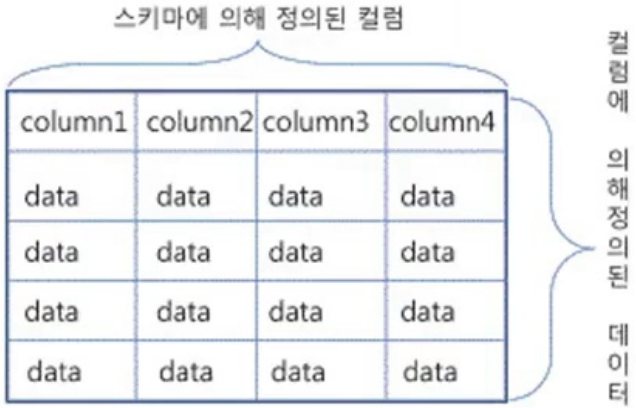

데이터 과학
Data Science
Information
Product
Brand
Success
Network
Information
Marketing
Social Media
Analysis
Development


수집 대상 데이터의 종류
http://www.dbguide.net
수집 데이터의 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
• 프로세스 생성
• 기계 생성
• 사람 생성
suanlab · 데이터 과학(Data Science) - 03 데이터 수집 - 2. 수집 데이터 분류
4

정형 데이터(Structured Data)

http://www.dbguide.net

- 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 데이터와 비교할 때 가장 큰 차이점은 데이터의 스키마를 지원하는 것
- 데이터의 스키마 정보를 관리하는 DBMS와 데이터 내용이 저장되는 데이터 저장소로 구분

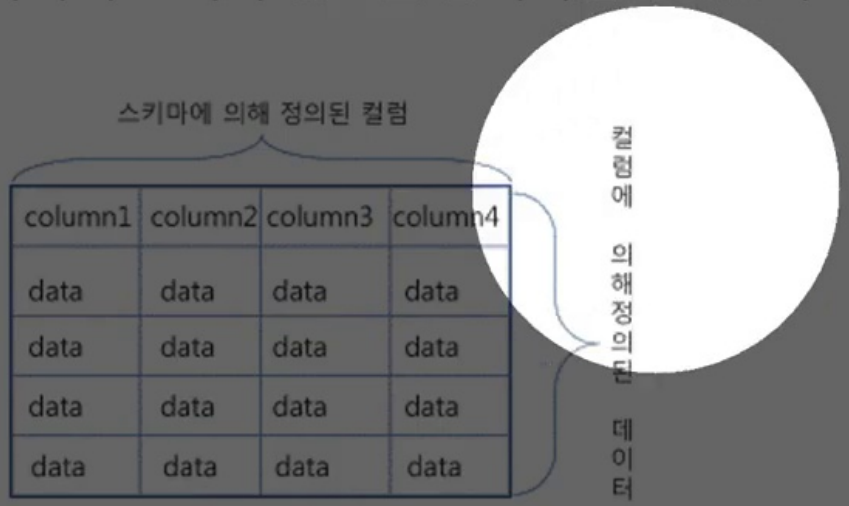


	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1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

정형 데이터(Structured Data)

http://www.dbguide.net

- 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 데이터와 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 데이터(Semi-Structured Data)

http://www.dbguide.net

- 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" }]
```

반정형 데이터(Semi-Structured Data)

http://www.dbguide.net

- 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" }]
```

반정형 데이터(Semi-Structured Data)

http://www.dbguide.net

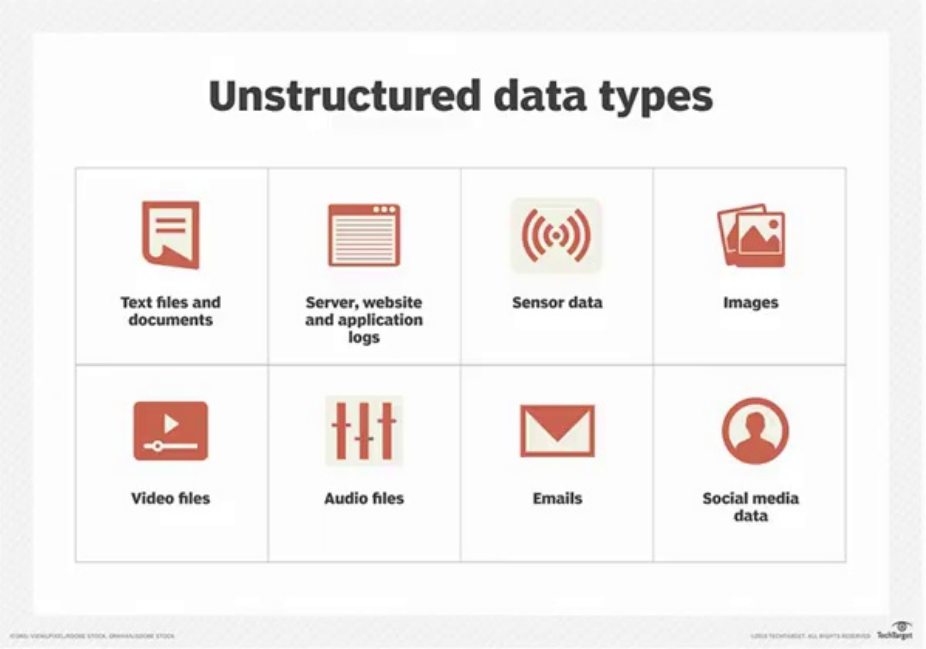
- 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 갖고 있음
- 일반적으로 파일 형태로 저장
- 데이터 내부에 데이터 구조에 대한 메터 정보를 갖고 있기 때문에 어떤 형태를 가진 데이터인지를 파악하는 것이 필요
- 데이터 내부에 있는 규칙성을 파악해 데이터를 파싱할 수 있는 파싱 규칙을 적용
- 반정형 데이터 예
  - URL 형태로 존재 – HTML
  - 오픈 API 형태로 제공 - XML, JSON
  - 로그형태 - 웹로그, IOT에서 제공하는 센서 데이터

```
[{ "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 데이터(Unstructured Data)

http://www.dbguide.net

- 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

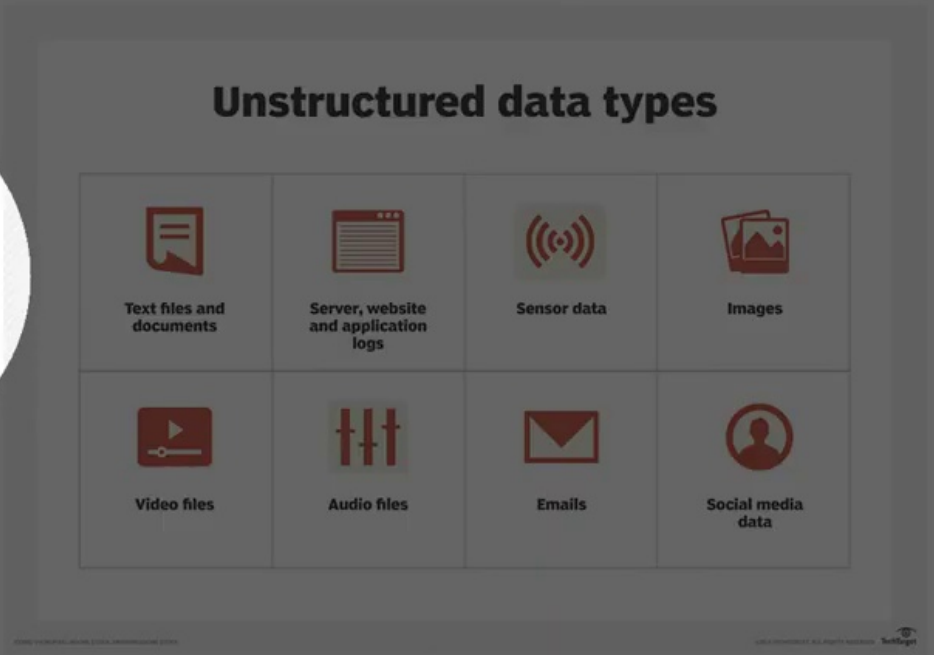


https://searchenterpriseai.techtarget.com/feature/Convert-unstructured-data-to-structured-data-with-machine-learning

비정형 데이터(Unstructured Data)

http://www.dbguide.net

- 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

http://www.dbguide.net

형태	특징	난이도
정형 데이터	▪ 내부 시스템인 경우가 대부분이라 수집이 쉬움 ▪ 파일 형태의 스프레드시트라도 내부 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

http://www.dbguide.net

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

http://www.dbguide.net

- 수집하려는 데이터를 저장된 위치에 따라 분류하면 내부 시스템에 저장되는 내부 데이터와 외부 시스템에 저장된 외부 데이터로 구분
- 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 데이터(Internal Data)

http://www.dbguide.net

- 수집하는 원천 데이터의 데이터 저장소가 내부 시스템에 있는 데이터를 의미
- 단순히 물리적 데이터 저장소 외에도 내부데이터와 외부 데이터의 가장 큰 구별점은 데이터 제공자와 상호 협약에 의한 의사소통이 가능하다는 점
- 원천 데이터와 수집한 데이터가 동일 시스템에 저장되어 있으므로 원천 데이터가 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 데이터(External Data)

http://www.dbguide.net

- 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
- 데이터의 전처리 과정 없이 원본 데이터를 수집 후, 수집 시스템에서 처리를 할 수 있도록 인터페이스를 설계하는 것이 좋음

데이터 위치별 수집 및 아키텍처 구성 난이도

http://www.dbguide.net

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 CRUD 처리 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

http://www.dbguide.net

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