

Data Science

데이터 분석

데이터 탐색

강사 양석환



데이터의 수집과 변환



• 데이터 수집

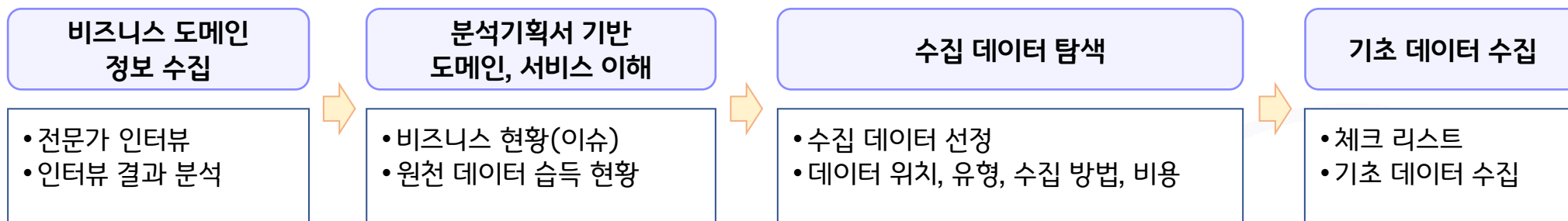
• 데이터 처리 시스템에 들어갈 데이터를 모으는 과정으로 여러 장소에 있는 데이터를 한 곳으로 모으는 작업

• 데이터 수집 수행 자료

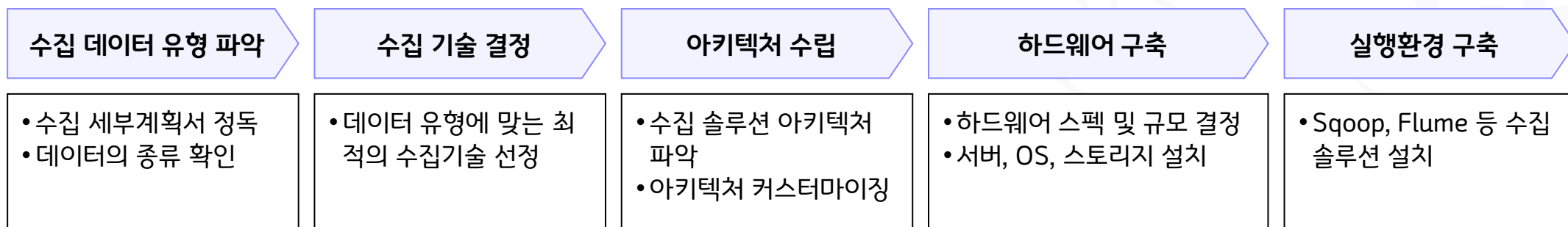
- 용어집
- 원천 데이터 소유기관 정보
- 서비스 흐름도
- 데이터 수집 기술 매뉴얼
- 업무 매뉴얼
- 인프라 구성도
- 데이터 명세서
- 소프트웨어 아키텍처 개념도
- 데이터 수집 계획서
- 수집 솔루션 매뉴얼
- 원천데이터 담당자 정보
- 하둡 오퍼레이션 매뉴얼
- 비즈니스 및 원천 데이터 파악을 위한 비즈니스 모델 등

• 데이터 수집

• 기초 데이터 수집 수행 절차



• 데이터 수집 시스템 구축 절차



• 데이터 수집 기술

• 정형 데이터

- ETL (Extract, Transform, Load)
 - 수집 대상 데이터를 추출 및 가공하여 데이터 웨어하우스에 저장하는 기술
- FTP (File Transfer Protocol)
 - TCP/IP, UDP 프로토콜을 통해 원격지 시스템으로부터 파일을 송수신 하는 기술
- API (Application Programming Interface)
 - 솔루션 제조사 및 서드파티 소프트웨어로 제공되는 도구로, 시스템 간 연동을 통해 실시간으로 데이터를 수신할 수 있도록 기능을 제공하는 인터페이스
- DB to DB
 - 데이터베이스 관리 시스템(DBMS) 간 데이터를 동기화 또는 전송하는 방법
- 스쿱 (Sqoop)
 - 관계형 데이터베이스(RDBMS)와 하둡(Hadoop) 간 데이터를 전송하는 방법

• 데이터 수집 기술

• 비정형 데이터

- 크롤링 (Crawling)

- 인터넷상에서 제공되는 다양한 웹 사이트, 소셜 네트워크 정보, 뉴스, 게시판 등으로부터 웹 문서 및 정보를 수집하는 기술

- RSS (Rich Site Summary)

- 블로그, 뉴스, 쇼핑몰 등의 웹 사이트에 게시된 새로운 글을 공유하기 위해 XML 기반으로 정보를 배포하는 프로토콜

- Open API

- 응용 프로그램을 통해 실시간으로 데이터를 수신할 수 있도록 공개된 API

- 척와 (Chukwa)

- 분산 시스템으로부터 데이터를 수집, 하둡 파일 시스템에 저장, 실시간으로 분석할 수 있는 기능 제공

- 카프카 (Kafka)

- 대용량 실시간 로그처리를 위한 분산 스트리밍 플랫폼 기술

• 데이터 수집 기술

• 반정형 데이터

- 플럼 (Flume)
 - 분산 환경에서 대량의 로그 데이터를 수집, 전송하고 분석하는 기능 제공
- 스크라이브 (Scribe)
 - 다수의 수집 대상 서버로부터 실시간으로 데이터를 수집, 분산 시스템에 데이터를 저장하는 기능 제공
- 센싱 (Sensing)
 - 센서로부터 수집 및 생성된 데이터를 네트워크를 통해 수집하는 기능 제공
- 스트리밍 (Streaming): TCP, UDP, Bluetooth, RFID
 - 네트워크를 통해 센서 데이터 및 오디오, 비디오 등의 미디어 데이터를 실시간으로 수집하는 기술

- **데이터 유형 및 속성 파악**

- **데이터 수집 세부 계획 작성**

- 데이터 유형, 위치, 크기, 보관방식, 수집주기, 확보비용, 데이터 이관절차 등을 조사하여 세부계획 작성

- **데이터 위치 및 비용**

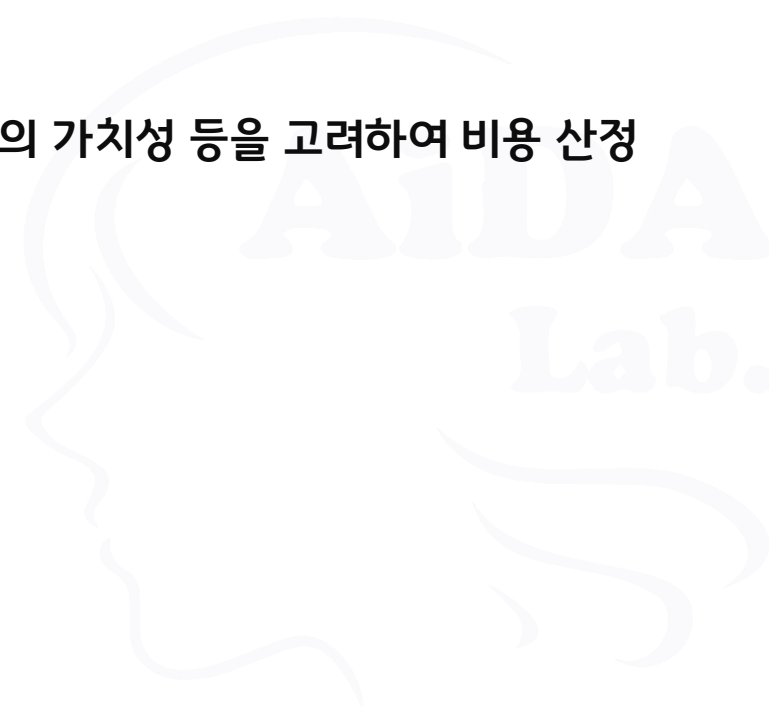
- 데이터의 종류, 크기 및 보관주기, 수집주기, 수집방식, 수집기술, 데이터의 가치성 등을 고려하여 비용 산정

- **수집되는 데이터의 형태**

- HTML, XML, JSON 등

- **데이터 저장 방식**

- 파일 시스템, 관계형 데이터베이스, 분산처리 데이터베이스 등



• 데이터 유형 및 속성 파악

• 데이터 적절성 검증

- 데이터 누락 점검: 수집 데이터 셋의 누락, 결측 여부를 판단하여 누락 발생 시 재 수집함
- 소스 데이터와 비교: 수집 데이터와 소스 데이터의 크기 및 개수를 비교, 검증함
- 데이터의 정확성 점검: 유효하지 않은 데이터의 존재여부를 점검함
- 보안 사항 점검: 수집 데이터의 개인정보 유무 등 보안 사항의 점검이 필요함
- 저작권 점검: 데이터의 저작권 등 법률적 검토를 수행함
- 대량 트래픽 발생 여부: 네트워크 및 시스템에 트래픽을 발생시키는 데이터의 존재여부를 검증함

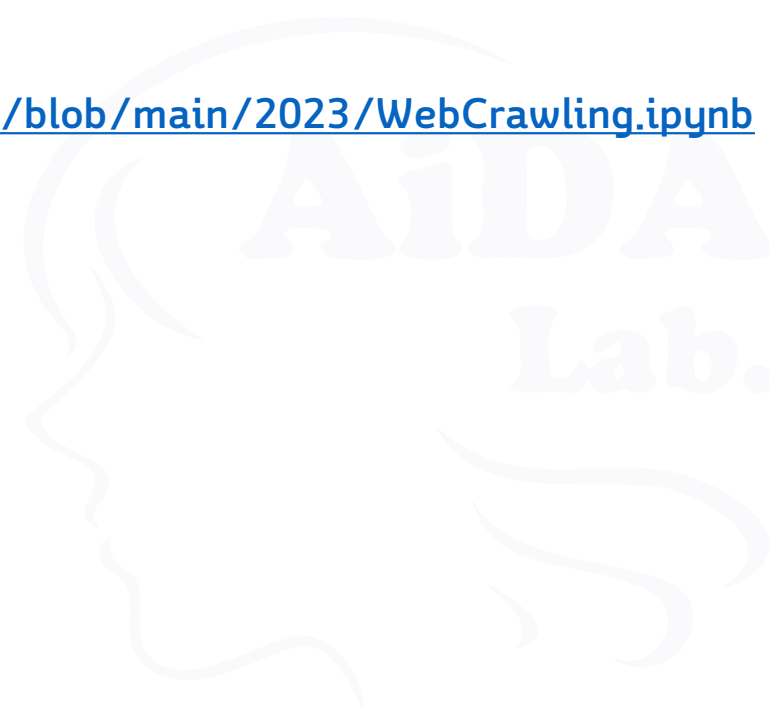
- 데이터 수집 실습

- 크롤링

- Google Colab 사용

- 네이버 뉴스 크롤링하기

- <https://colab.research.google.com/github/aidalabs/Lectures/blob/main/2023/WebCrawling.ipynb>



• 데이터 변환 (Data Transformation)

- 데이터를 하나의 표현 형식에서 다른 형식으로 변형하는 과정
- 데이터 변환 방식의 종류
 - 비정형 데이터를 정형 데이터 형태로 저장하는 방식(관계형 데이터베이스)
 - 수집 데이터를 분산 파일 시스템으로 저장하는 방식(HDFS 등)
 - 주제별, 시계열적으로 저장하는 방식(데이터 웨어하우스)
 - Key-Value 형태로 저장하는 방식(NoSQL)

수집 데이터의 저장형태에 따른 데이터 변환 솔루션

수집 데이터 저장 형태	저장 솔루션	라이선스
관계형 데이터베이스	MySQL, Oracle, DB2, PostgreSQL 등	상용 라이선스, 오픈소스
분산 데이터 저장	HDFS (Hadoop Distributed File System)	오픈소스
데이터 웨어하우스	네티자, 테라데이타, 그린폴럼의 DW 솔루션 등	상용 라이선스
NoSQL	HBase, Cassandra, MongoDB	오픈소스

- 데이터 변환 (Data Transformation)

- 데이터 형태에 따른 데이터 변환 작업

- 관계형 데이터베이스

- 데이터베이스 구조 설계를 기반으로 함

- DBMS 구축여부 결정 → 저장 데이터베이스 결정 → DBMS 설치 → 테이블 구조 설계

- 비정형/반정형 데이터의 변환

- 데이터 전처리나 후처리가 수행되기 전에 비정형/반정형 데이터를 구조적 형태로 전환하여 저장하는 과정

- 수집 데이터의 속성 구조 파악 → 데이터 수집 절차에 대한 수행 코드 정의 → 데이터 저장 프로그램 작성 → DB에 저장
(추출하려는 정보의 위치, 구조 파악 후 데이터 추출)

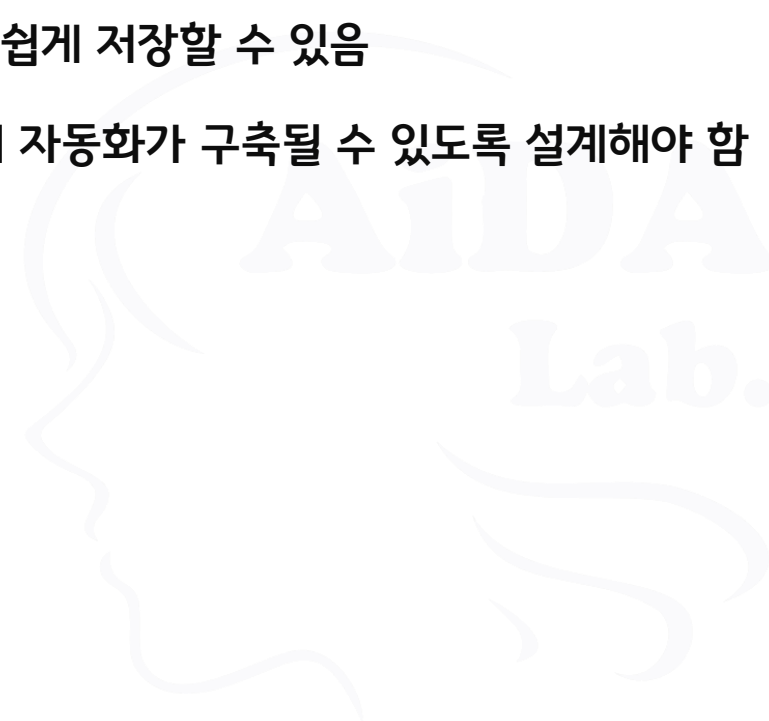
- 융합 데이터베이스 설계

- 데이터의 유형과 의미를 파악하여 활용 목적별 융합 DB 설계

- 데이터 변환 (Data Transformation)

- 데이터 변환 시 고려사항

- 비정형, 반정형 데이터를 데이터 분석의 용이성을 위해 정형화된 데이터베이스로 변환하는 작업에 집중할 것
 - 수집 데이터의 속성 구조를 정확히 파악하여야 틀을 이용하여 데이터를 쉽게 저장할 수 있음
 - 융합 DB 구성은 활용 업무 목적을 정확히 판단하는 것이 중요하며, 쉽게 자동화가 구축될 수 있도록 설계해야 함



**THANK
YOU**

