



데이터와 서비스 활용

<http://www.dbguide.net>

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

데이터 형태별 서비스 활용

<http://www.dbguide.net>

데이터 형태	저장 형태	서비스 활용	활용 API
정형 데이터	DBMS	DBMS 스키마를 읽어 가시적(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

<http://www.dbguide.net>

- 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 수집 방법이 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

수집 방법 정의

<http://www.dbguide.net>

- 데이터의 종류와 형태별로 분류되고 수집 시스템에 저장 · 처리할 수 있는 기능과 정형/반정형 데이터의 형태로 저장할 수 있는 기능 필요
- 수집 데이터의 위치는 수집 방법에 영향을 미치지 않음
- 수집 방법에 영향을 미치는 것은 데이터의 형태와 종류
- 내부/외부를 구분해 나누는 실익은 업무적 제약 사항 때문에 수집 방법별로 존재하는 수집 기술들에 영향을 미침
- 내부에서 수집할 경우에는 DBMS에 저장된 정형 데이터가 주를 이루고, 외부에서 수집할 경우에는 스크립트 형태로 제공하는 반정형 데이터가 주를 이룸
- 실무에서 데이터 수집 계획을 수립할 경우 내부와 외부를 구분하는 것이 데이터 수집의 비용적인 측면과 관리적 측면에서 유리

수집 방법의 종류

<http://www.dbguide.net>

형태	종류	수집방법 정의
정형 데이터	DBMS	<ul style="list-style-type: none"> DBMS 벤더가 제공하는 API를 통해 정형 데이터에 접근 데이터를 수집하고 시스템에 저장
	이진 파일	<ul style="list-style-type: none"> ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드 해당 파일의 API를 통해 데이터 처리
반정형 데이터	스크립트 파일	<ul style="list-style-type: none"> http 프로토콜을 사용해 파일의 텍스트를 스크랩 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
	이진 파일	<ul style="list-style-type: none"> 스트리밍을 사용해 파일의 텍스트를 스크랩 데이터에 저장된 메타정보를 읽어 파일을 파싱해 데이터 처리
비정형 데이터	이진 파일	<ul style="list-style-type: none"> ftp 프로토콜을 사용해 파일을 수집 시스템에 다운로드 해당 파일을 API를 통해 데이터 처리
	스크립트 파일	<ul style="list-style-type: none"> http 프로토콜을 사용해 파일의 텍스트를 스크랩 내부 처리에서 텍스트를 파싱해 데이터 처리

데이터 종류별 연동 방법

<http://www.dbguide.net>

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

수집 기술의 종류

<http://www.dbguide.net>

- 수집 방법에 필요한 기본적인 기능요건들을 만족시키거나 필요한 기능 요건들을 추가해 수집 기술 제작 가능
- 이미 선행되어 명명된 기술이 존재하므로 하나의 수집 방법에 여러 개의 수집기술이 존재
- 기술적으로는 큰 차이가 없지만 서비스 활용 용도에 따라서 수집 기술이 나뉨
- HTTP 수집
 - 웹 스크래핑
 - Open API
- 로그/센서 수집
 - 로그 수집
 - 센서 데이터 수집
- DBMS 수집
- FTP 수집

수집 기술의 종류

<http://www.dbguide.net>

- 수집 방법에 필요한 기본적인 기능요건들을 만족시키거나 필요한 기능 요건들을 추가해 수집 기술 제작 가능
- 이미 선행되어 명명된 기술이 존재하므로 하나의 수집 방법에 여러 개의 수집기술이 존재
- 기술적으로는 큰 차이가 없지만 서비스 활용 용도에 따라서 수집 기술이 나뉘짐
- HTTP 수집
 - 웹 스크래핑
 - Open API
- 로그/센서 수집
 - 로그 수집
 - 센서 데이터 수집
- DBMS 수집
- FTP 수집

웹 스크래핑(Web Scraping)



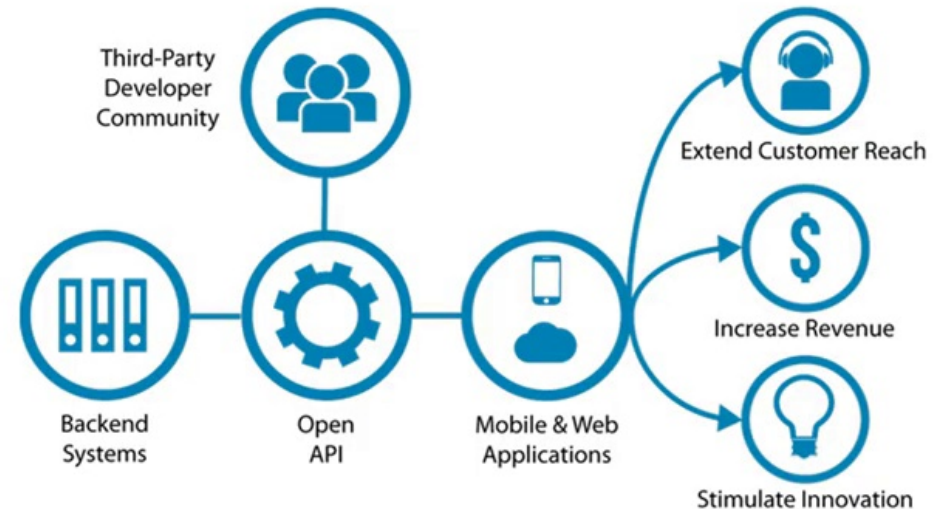
<http://www.websitescraper.com/>

- 뉴스, 블로그, SNS, 웹 페이지 등 웹에서 공개되어 있는 데이터를 자동으로 수집하여 데이터를 추출하고 저장
- 일반적으로 HTTP GET 요청을 보낸 다음 웹 서버가 전송하는 모든 정보를 복사하여 저장
- 수백 또는 수천 개의 웹 페이지가 있는 대규모 사이트의 경우에도 몇 초 안에 웹사이트의 모든 콘텐츠를 다운로드 가능

Open API

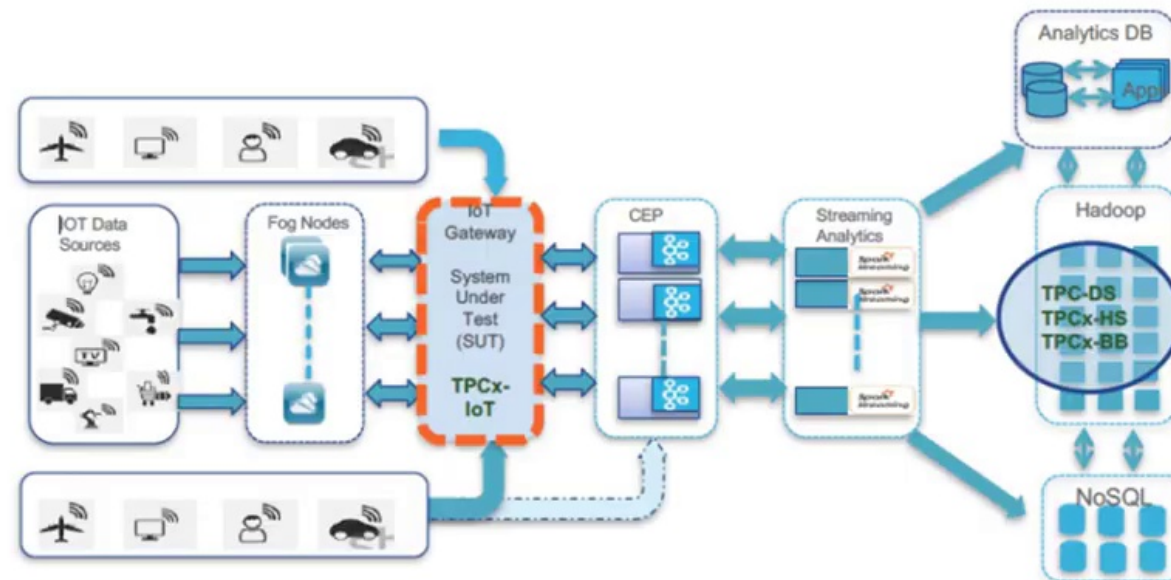
https://ko.wikipedia.org/wiki/오픈_API

- Open API(Open Application Programming Interface)는 누구나 사용할 수 있도록 공개된 API
- 개발자에게 사유 응용 소프트웨어나 웹 서비스에 프로그래밍적인 권한 제공
- 네이버 지도, 구글맵, 오픈스트리트맵 등이 대표적인 예
- 대한민국 정부에서는 공공데이터포털을 통해 도로명주소조회서비스, 동네예보정보조회서비스 등 오픈API 운영 및 제공



센서 데이터 수집

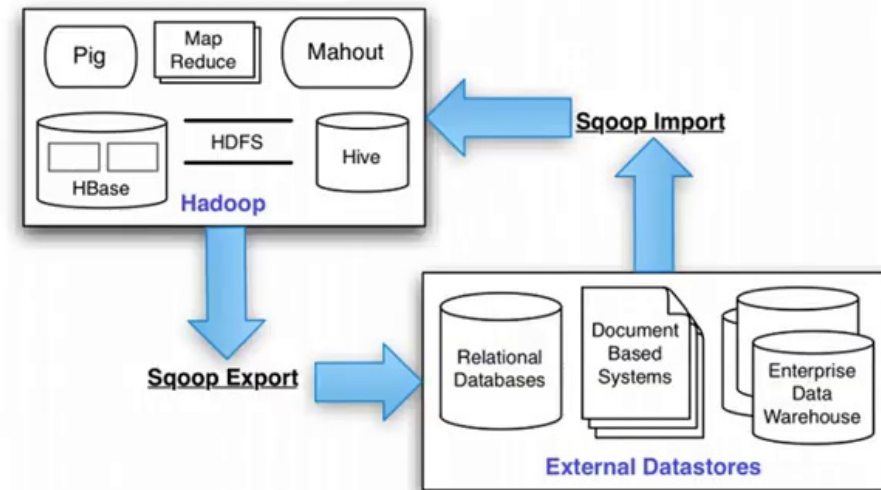
- IoT 기술의 발달로 수많은 기기로부터 각종 센서 데이터 수집 가능
- 온도, 습도, 자이로, 압력, 속도/가속도, 가스, 초음파, 자기 센서 등 수 많은 센서 존재
- 도시, 차량, 철도, 비행기, CCTV, 스마트 기기 등에서 센서 데이터 수집



<https://www.machbase.com/kr/blog/?mod=document&uid=87>

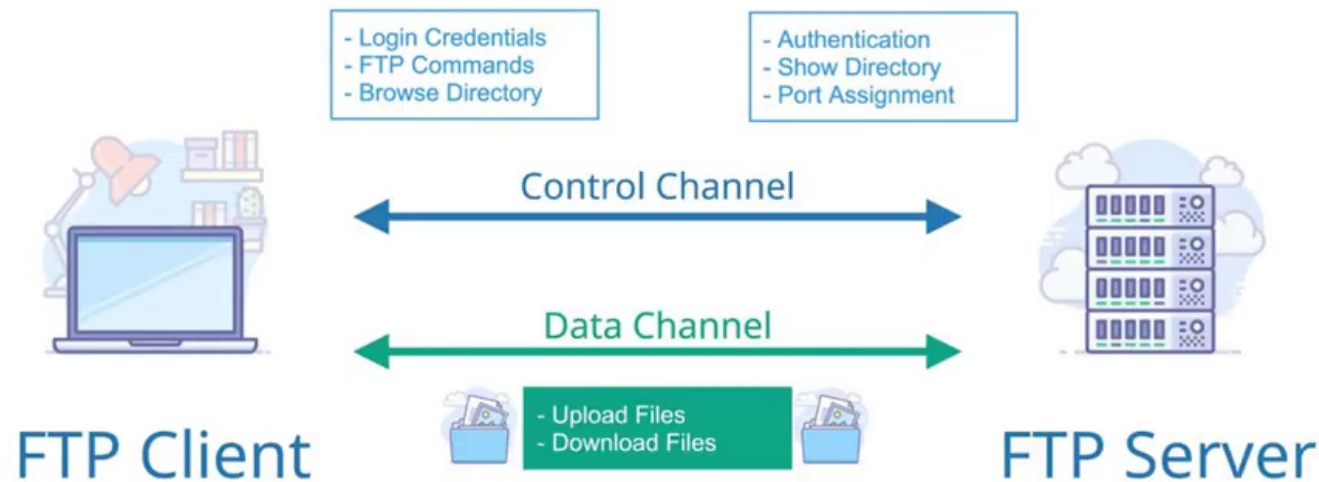
DBMS 수집

- 데이터베이스 시스템으로부터 특정 테이블과 컬럼을 선택하여 정형 데이터 수집 가능
- 레코드 단위로 수집해 분산파일시스템으로 전송 가능
- 기존에 운영되는 DBMS의 연결만으로 쉽게 데이터 수집 가능



FTP 수집

- TCP/IP 프로토콜을 활용하는 인터넷 서버로부터 각종 파일들을 송수신
- 파일 전송에 최적화되어 있어서 대용량 데이터 수집에 용이



<https://www.exavault.com/blog/what-is-ftp-tutorial-video-blog/>

