

자기주도 온라인 프로젝트 I

Day 6.

DB 데이터 적재 및 가공

최호근 컨설턴트



Day 6

CONTENTS

Samsung Software Academy For Youth



DB데이터 적재 및 가공

최 호 근 Project consultant

- 와이즈넷 분석설계 수석 컨설턴트
- 서울시 민원 데이터분석 자문위원
- SK플래닛 데이터분석 파트장
- 現 국가대표 인도어사이클 체조선수



00 개요



❖ 실무프로젝트시 데이터 이행 ?

- 프로젝트 오픈시 실제 데이터를 운영DB에 반영하는 작업
- 대량데이터의 경우 오픈 후에는 증분(신규/수정/삭제)데이터만 반영(스케줄 또는 API 연계등)

❖ 데이터 적재 시점

- 서버 Down 타임에 적재(운영서버가 교체 불가인 경우)
- 사전 적재(신규서버 도입인 경우)
- 서비스 오픈 후 적재 (내부오픈 등으로 외부 오픈일이 따로 있는 경우 운영자가 따로 정해진 시간에 적재)

00 개요



❖ 방법 (상황에 따라 여러가지)

- DB백업본 restore(특정시점 이전까지의 데이터 dump)사용하여 데이터 이관
- 텍스트 데이터 이관(다양한 형식의 텍스트파일 CSV,TXT,JSON . . .)
- DB TO DB 마이그레이션(DB 또는 솔루션 사용)
- 정제된 텍스트 데이터를 DB에 직접 인서트
- 정제되지 않은 데이터를 입력이 가능한 형태로 가공 및 인서트
- 데이터 제공자 측에서 제공하는 Restful API를 호출하여 데이터 인서트(외부망으로 가능해야 하며 주로 증분 데이터에만 사용)

00 개요



❖ 정제된 텍스트 데이터 입력시 사용하는 방법을 이용한 실습

- 텍스트 데이터 이관(다양한 형식의 텍스트파일 CSV,TXT,JSON . . .) 및 가공



01 도로명 주소데이터 다운로드

주소DB (구 매칭데이터)
주소 단위의 주소정보를 매월 전체(변동) 자료와 매일 변동 자료로 제공합니다.

주소DB 다운로드

전체자료 < 2020년

1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
----	----	----	----	----	----	----	----	----	-----	-----	-----

월변동 자료 < 2020년

1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
----	----	----	----	----	----	----	----	----	-----	-----	-----

일변동자료 < 2020년 03월

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----










연계신청

* 최근 2년 이내 자료를 제공합니다.
* 19.8월 월변동분부터 관련지번 변동분이 제공됩니다(일변동 자료 현행화 방식 사용).

<http://www.juso.go.kr/addrlink/addressBuildDevNew.do?menu=match>

01 도로명 주소데이터 다운로드



이름	유형
 [가이드]주소DB 활용방법.pdf	Adobe Acrobat Document
 [자료건수]주소DB(2020년 02월 29...	텍스트 문서
 개선_도로명코드_전체분.txt	텍스트 문서
 부가정보_강원도.txt	텍스트 문서
 부가정보_경기도.txt	텍스트 문서
 부가정보_경상남도.txt	텍스트 문서
 부가정보_경상북도.txt	텍스트 문서
 부가정보_광주광역시.txt	텍스트 문서
 부가정보_대구광역시.txt	텍스트 문서

<http://www.juso.go.kr/addrlink/addressBuildDevNew.do?menu=match>

02 데이터 가공 및 변환



❖ 텍스트 데이터 중 입력 대상 데이터 가공

- 구분자 확인(csv변환시 구분기호)

개선_도로명코드_전체분.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말

111102005001|세종대로|Sejong-daero|00|서울특별시|Seoul|종로구|Jongno-gu||2||이|||
111102005001|세종대로|Sejong-daero|01|서울특별시|Seoul|종로구|Jongno-gu|세종로|Sejongno|1|119|0|||
111102005001|세종대로|Sejong-daero|02|서울특별시|Seoul|종로구|Jongno-gu|종로1가|Jongno 1(il)-ga|1|126|0|||



<http://www.juso.go.kr/addrlink/addressBuildDevNew.do?menu=match>

02 데이터 가공 및 변환



❖ 텍스트 속성값 확인

- 필드 확인 후 추후에 필요 없는 컬럼 값 제거
- 입력을 위한 파일 형식 변환(UTF-8)

**개발자센터**

주소DB (구 매칭데이터)

주소 단위의 주소정보를 매월 전체(변동) 자료와 매일 변동 자료로 제공합니다.

개발+
제안하기
문의답하기
Tech & Tips

주소검색솔루션
주소검색솔루션
PC용 주소검색기

오픈API
API활용체험
도로명주소API

주소DB는 여러 개의 건물이 하나의 도로명주소를 갖는 집합 건물(예: 아파트)의 경우 한 건의 주소정보를 제공하도록 구성된 주 600만여 건의 주소와 800만여 건의 지번정보를 바탕으로 사용자의 필요에 따라 선택적으로 활용할 수 있도록 도로명코드 / 주소 / 지번 / 부가정보로 분리-구성하였습니다.

주소DB 레이아웃

[주소DB 관계도 보기](#)  [도로명코드 보기](#) [도로명주소 보기](#) [지번\(대표지번+관련지번\) 보기](#)

[주소 변동이력조회](#) [조회](#)

03 보건복지부 선별진료소 데이터 다운로드

DATA 공공데이터포털
 . GO . KR

마이페이지 로그인 사이트맵 ENGLISH
 



데이터셋
 제공신청
 활용사례
 정보공유
 이용약관

/ 데이터셋 / 파일데이터

파일데이터

·제공기관
 공공데이터활용지원센터

·관리부서명
 공공데이터활용지원센터

·관리부서 전화번호
 02-1566-0025

·등록일
 2020-03-04

보건복지부_선별진료소

보건복지부로부터 데이터를 받아 활용지원센터에서 등록하였습니다.
 선별진료소 현황 및 주소 데이터입니다.

매체유형 : 텍스트
 파일, 링크 건수 : 7
 전체 행 수 : N/A
 확장자 : CSV
 다운로드 횟수(바로그기 횟수) : 907

☐ 전체

※ 서비스 오류가 있을시 오류신고 버튼을 이용해주세요.

☐ CSV
 보건복지부_선별진료소_현황_주소_2...
☐ CSV
 보건복지부_선별진료소_현황_주소_2...

 멀티다운로드
 닫기
 오류신고


 멀티다운로드
 상세정보
 오류신고


보건복지부_선별진료소_현황_주소_20200311			
업데이트 주기	주간	차기등록예정일	2020-03-12
비용부과유무	무료	비용부과기준 및 단위	없음
다운로드 횟수	11		

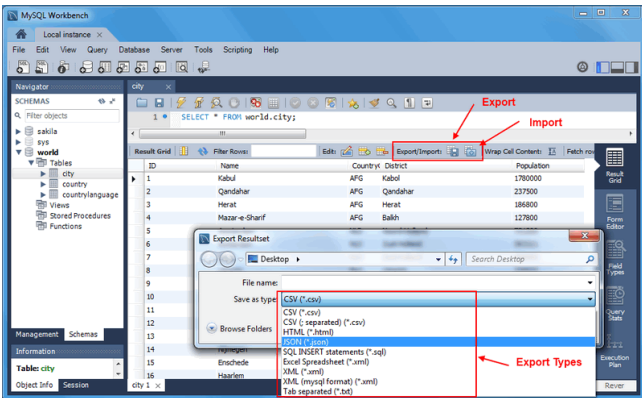
<https://www.data.go.kr/dataset/15043008/fileData.do>

04 데이터 import



❖ 다양한 방법을 통해 데이터 적재

- Mysql workbench, mysql load data 쿼리등 사용
- 도로명주소 → addr테이블 테이블로 import
- 보건복지부 선별진료소 → medical 테이블로 import



```
mysql> use bharathi;
Database changed
mysql> show tables;
+-----+
| Tables_in_bharathi |
+-----+
| cash_order          |
+-----+
1 row in set (0.00 sec)

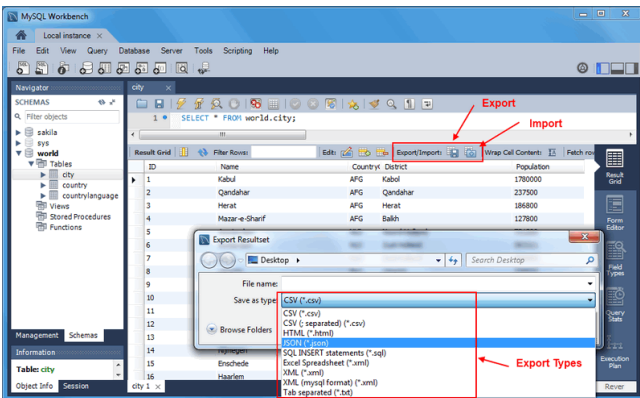
mysql> load data local infile 'C:\Python27\cash_order.csv'
-> into table cash_order
-> fields terminated by ','
-> lines terminated by '\n'
-> ;
ERROR 2 (HY000): File 'C:\Python27\cash_order.csv' not found (Errcode: 2 - No
file or directory)
mysql> load data local infile 'C:/Python27/cash_order.csv'
-> into table cash_order
-> fields terminated by ','
-> lines terminated by '\n';
Query OK, 9868890 rows affected, 65535 warnings (23 min 19.78 sec)
Records: 142316285 Deleted: 0 Skipped: 132447395 Warnings: 142316285
```

05 데이터 export



❖ 최종데이터 가공 및 내보내기

- 본인 소속 지역데이터만 남김(대전의 경우 대전의 도로명주소, 진료소 정보만 테이블에 하도록 함)
- 해당 테이블 데이터 내보내기 (CSV파일, 파일명: addr_ssafy아이디.csv,
- 각 테이블 select 결과 화면 스크린샷 저장



```
mysql> use bharathi;
Database changed
mysql> show tables;
+-----+
| Tables_in_bharathi |
+-----+
| cash_order          |
+-----+
1 row in set (0.00 sec)

mysql> load data local infile 'C:\Python27\cash_order.csv'
-> into table cash_order
-> fields terminated by ','
-> lines terminated by '\n'
->
ERROR 2 (HY000): File 'C:\Python27\cash_order.csv' not found (Errcode: 2 - No
file or directory)
mysql> load data local infile 'C:/Python27/cash_order.csv'
-> into table cash_order
-> fields terminated by ','
-> lines terminated by '\n';
Query OK, 9868890 rows affected, 65535 warnings (23 min 19.78 sec)
Records: 142316285 Deleted: 0 Skipped: 132447395 Warnings: 142316285
```

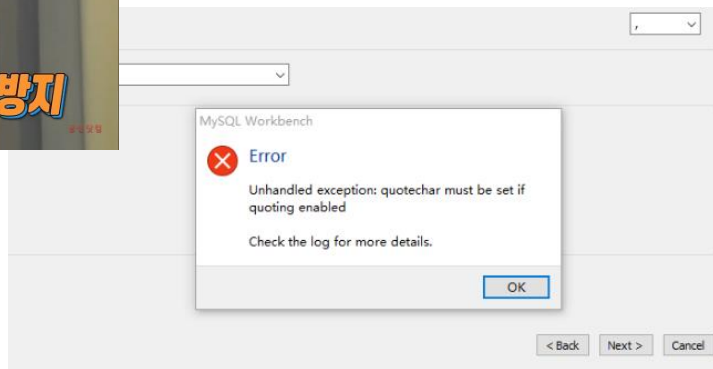
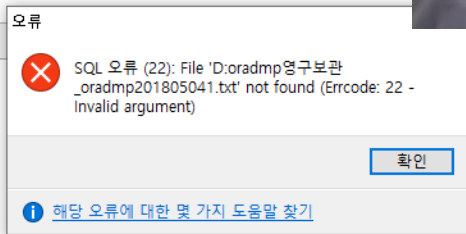
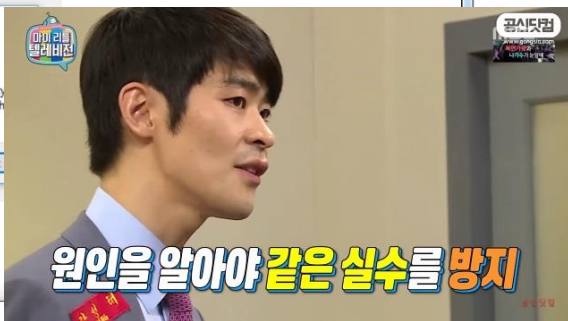
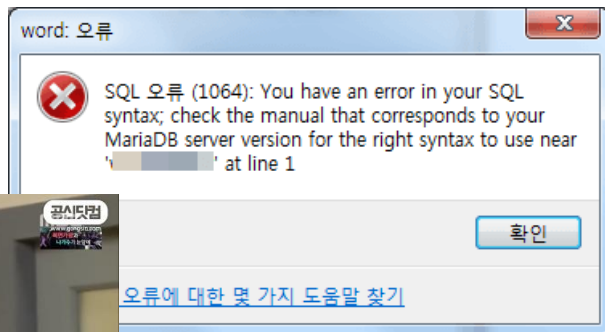
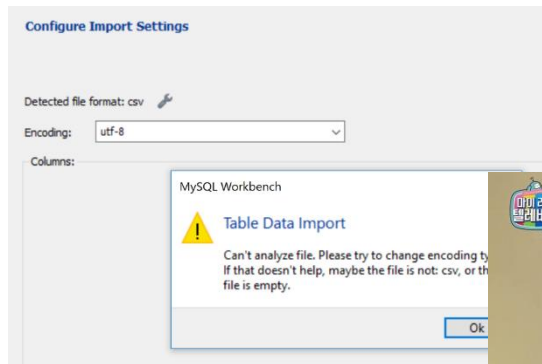
06 최종산출물



❖ 최종산출물

- 해당데이터 내보내기 (CSV파일, 파일명: addr_ssafy아이디.csv,
- 각 데이터 셀렉트 화면 샘플 스크린샷 저장 (select * from 테이블명 쿼리결과 스크린샷, ssafy아이디, 입력날짜 가 나오도록 함)
- Readme(MD파일)

07 error 발생시 원인 확인 및 해결





응원합니다

