Harjoitus 1 - Datatyypit

Big Data -ympäristöt

Tapani Alastalo

Jyväskylän ammattikorkeakoulu
JAMK University of Applied Sciences

Sisältö

1 Vaihe 1

1.1 Mitä tauluja SQL-tiedoston rakenne sisältää ja mikä on taulujen rakenne?

Tiedosto sisältää taulut DEPT, EMPS ja SDEPT.

DEPT koostuu merkkijonoista (varchar) DEPT, MANAGER ja DUTY, sekä decimaaliarvosta (decimal) BUDGET.

EMPS koostuu useammista merkkijonoista (char ja varchar) ja decimaaliarvoista (decimal).

SDEPT koostuu merkkijonoista (varchar) DEPT ja MANAGER, sekä decimaaliarvosta (decimal) CODE.

1.2 Sisältääkö tiedosto myös tauluihin tallennettavaa dataa?

Kyllä. DEPT sisältää kymmeniä esimerkki syötteitä. EMPS sisältää 500 esimerkki syötettä ja SDEPT kolme esimerkki syötettä.

1.3 Jos sisältää, niin anna pari esimerkkiriviä niistä tauluista, joihin dataa tallennetaan.

DEPT:

('A','SMITH A','ACCOUNTING',100000.00), ('B','JONES B','INF SYSTEMS',120000.00)

EMPT:

(1,'A','KOO','SARA','234 WEST ST','BANFF','AB','T0R0R0','421-6923',6.00,0,100.00),(2,'B','MARSH','JOHN','456 EAST AVE','BANFF','AB','T0R0R0','963-2176',NULL,1,75.00)

SDEPT:

('A','SMITH A',1),('B','JONES B',2)

2 Vaihe 2

2.1 Vertaile avaamaasi tiedostoa rakenteen osalta sekä tekstieditorissa että Excel/CSV Viewer -ohjelmistossa. Miten tiedoston rakenne näyttäytyy tekstieditorissa ja kuinka Excel/CSV Viewer -ohjelmistot sen esittävät?

Tekstieditorissa tiedoston rakenne on rivitettyä pilkulla eroteltua tekstiä, minkä esimerkiksi Excel osaa jäsennellä taulukoksi, missä pilkku toimii erottimena rivin tekstin jakamisena sarakkeisiin.

2.2 Millaista dataa tiedosto sisältää ja mikä on mahdollisesti ollut sen alkuperäinen käyttötarkoitus?

Tiedosto sisältää myynnissä olleiden asuntojen tietoja, kuten osoite, kaupunki, postinumero, osavaltio, makuuhuoneiden lukumäärä, hinta, myyntipäivä ja koordinaatit.

Datan käyttötarkoitus on mahdollisesti ollut myytyjen asuntojen hinnoittelu esimerkiksi kaupungeittain. Lisäksi on voitu vertailla makuuhuoneiden lukumäärän vaikutusta hintakehitykseen eri alueilla.

3 Vaihe 3

3.1 Vertaa tätä tiedostoa edellisen tehtävän CSV-tiedostoon. Miten rakenne poikkeaa?

CSV tiedostossa data oli , eroteltua. XML:ssä erottelu tapahtuu tageilla, jotka toimivat samalla tavalla kuin HTML kielessä. Täten XML tiedoston lukeminen on ihmiselle helpompaa, mutta vastaavasti se vie enemmän tilaa.

3.2 Millaista dataa tiedostoon on tallennettu?

Tiedostoon on tallennettu levykokoelma. Levy (CD) sisältää otsikon, artistin, maan, levy-yhtiön tekstimuodossa, hinnan desimaalina ja vuoden numeroina.

3.3 Anna muutaman rivin esimerkki siitä millaiselta aiemmassa tehtävässä esitetty CSV-data näyttäisi XML-muodossa.

```
<APARTMENTS>
...
 <APARTMENT>
  <STREET>3526 HIGH ST</STREET>
  <CITY>SACRAMENTO</CITY>
  <ZIP>95838</ZIP>
  <STATE>CA</STATE>
  <BEDS>2</BEDS>
  <BATHS>1</BATHS>
  <SQ_FT>836</SQ_FT>
  <TYPE>Residential</TYPE>
  <SALE_DATE>Wed May 21 00:00:00 EDT 2008</SALE_DATE>
  <PRICE>59222</PRICE>
  <LATITUDE>38.631913</LATITUDE>
  <LONGITUDE>121.434879</LONGITUDE>
 </APARTMENT>
...
</APARTMENTS>
```

4 Vaihe 4

4.1 Millainen datarakenne on kahteen aiempaan semirakenteelliseen datatyyppiin nähden?

JSON koostuu objekteista, mitkä on sullottu aaltosulkeiden sisälle. Objektit sisältävät elementtejä, jotka koostuvat avain-arvo parista. Taulut koostuvat useasta objektista.

4.2 Millaista dataa tiedostoon on tallennettu?

Tiedostoon on tallennettu leivonnaisia, joille on kuorrutteet ja taikinat. Kuorrutteet ja taikinat voivat koostua useasta vaihtoehdosta.

4.3 Luo esimerkki JSON-tiedosto, johon on tallennettu kaksi riviä toisen vaiheen CSV-tiedostosta ja kolme riviä kolmannen vaiheen XML-tiedostosta.


```
{
"APARTMENTS": [
    {
     "STREET": "3526 HIGH ST",
     "CITY": "SACRAMENTO",
     "ZIP": "95838",
     "STATE": "CA",
     "BEDS": "2",
     "BATHS": "1",
     "SQ_FT": "836",
     "TYPE": "Residential",
     "SALE_DATE": "Wed May 21 00:00:00 EDT 2008",
     "PRICE": "59222",
     "LATITUDE": "38.631913",
     "LONGITUDE": "121.434879"
    },
    {
     "STREET": "51 OMAHA CT",
     "CITY": "SACRAMENTO",
     "ZIP": "95823",
     "STATE": "CA",
     "BEDS": "3",
```

```
      "BATHS": "1",
      "SQ_FT": "1167",
      "TYPE": "Residential",
      "SALE_DATE": "Wed May 21 00:00:00 EDT 2008",
      "PRICE": "68212",
      "LATITUDE": "38.4789",
      "LONGITUDE": "121.431233"
     }
   ],
  "CATALOG": {
   "CD": [
    {
     "TITLE": "Empire Burlesque",
     "ARTIST": "Bob Dylan",
     "COUNTRY": "USA",
     "COMPANY": "Columbia",
     "PRICE": "10.90",
     "YEAR": "1985"
    },
    {
     "TITLE": "Hide your heart",
     "ARTIST": "Bonnie Tylor",
     "COUNTRY": "UK",
     "COMPANY": "CBS Records",
     "PRICE": "9.90",
     "YEAR": "1988"
    },
    {
     "TITLE": "Greatest Hits",
     "ARTIST": "Dolly Parton",
     "COUNTRY": "USA",
     "COMPANY": "RCA",
     "PRICE": "9.90",
     "YEAR": "1982"
    }
   ]
  }
}
```


5 Vaihe 5

5.1 Millaista dataa kuva sisältää?

Kuva sisältää hexa dataa.

5.2 Onko siitä havaittavissa minkäänlaista rakenteellista piirrettä?

Tiedoston alun hexadata kertoo miten kuva on pakattu.