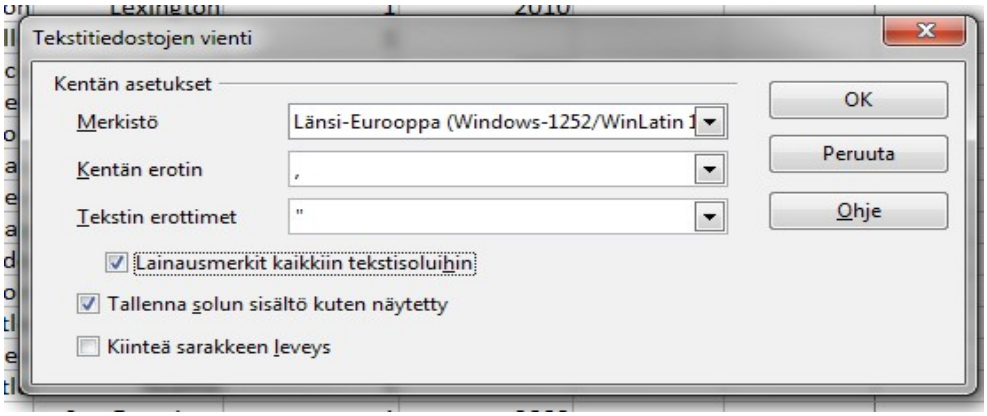
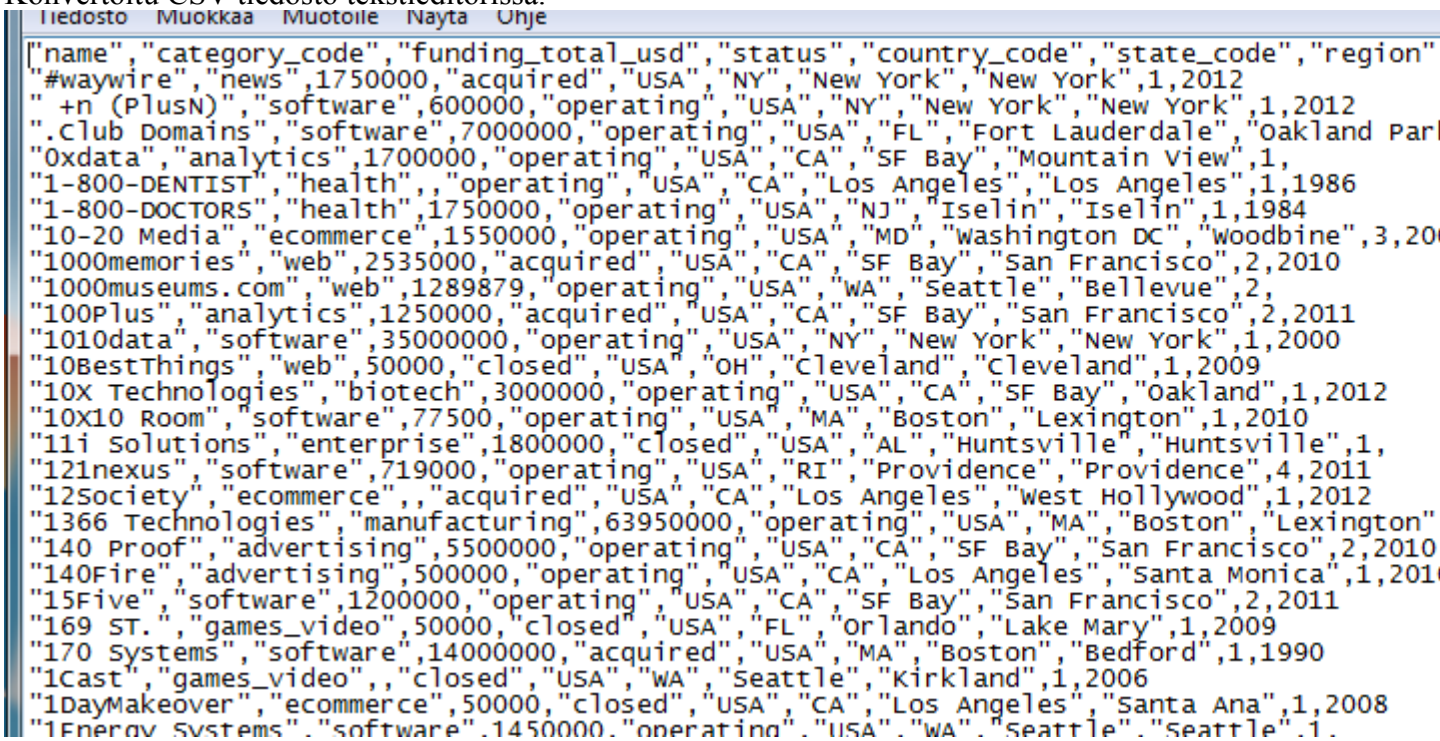
BDPa harjoitustehtävä 1 Ilmari Anturaniemi & Felix Bjugg

Data muutettiin OpenOfficen tallenna nimellä-toimintoa käyttämällä XLSX-muodosta CSV-muotoon. Alla olevasta kuvasta käyvät ilmi valitut asetukset, joista ulostulon kannalta merkittävimmät ovat pilkku erottimena ja lainausmerkkien käyttö soluissa. Lainausmerkit olivat käytössä pääosin siksi että dataa olisi ollut helpompi siivota tarpeen vaatiessa joukon keskeltä. Esimerkiksi hae ja korvaa-toiminnolla oltaisiin voitua etsiä tyhjät kohdat hakemalla ,, ja korvata ne. Tämä ei kuitenkaan olisi toiminut tyhjien founding_year kohtien kanssa, koska ne olivat rivien lopussa.



Konvertoitu CSV-tiedosto tekstieditorissa.



```
Tiedosto Muokkaa Muotoile Näytä Ohje
"name","category_code","funding_total_usd","status","country_code","state_code","region"
"#waywire","news",1750000,"acquired","USA","NY","New York","New York",1,2012
" +n (PlusN)","software",600000,"operating","USA","NY","New York","New York",1,2012
".Club Domains","software",7000000,"operating","USA","FL","Fort Lauderdale","Oakland Par
"0xdata","analytics",1700000,"operating","USA","CA","SF Bay","Mountain View",1,
"1-800-DENTIST","health",,"operating","USA","CA","Los Angeles","Los Angeles",1,1986
"1-800-DOCTORS","health",1750000,"operating","USA","NJ","Iselin","Iselin",1,1984
"10-20 Media","ecommerce",1550000,"operating","USA","MD","Washington DC","Woodbine",3,20
"1000memories","web",2535000,"acquired","USA","CA","SF Bay","San Francisco",2,2010
"1000museums.com","web",1289879,"operating","USA","WA","Seattle","Bellevue",2,
"100Plus","analytics",1250000,"acquired","USA","CA","SF Bay","San Francisco",2,2011
"1010data","software",35000000,"operating","USA","NY","New York","New York",1,2000
"10BestThings","web",50000,"closed","USA","OH","Cleveland","Cleveland",1,2009
"10X Technologies","biotech",3000000,"operating","USA","CA","SF Bay","Oakland",1,2012
"10X10 Room","software",77500,"operating","USA","MA","Boston","Lexington",1,2010
"11i Solutions","enterprise",1800000,"closed","USA","AL","Huntsville","Huntsville",1,
"121nexus","software",719000,"operating","USA","RI","Providence","Providence",4,2011
"12Society","ecommerce",,"acquired","USA","CA","Los Angeles","West Hollywood",1,2012
"1366 Technologies","manufacturing",63950000,"operating","USA","MA","Boston","Lexington"
"140 Proof","advertising",5500000,"operating","USA","CA","SF Bay","San Francisco",2,2010
"140Fire","advertising",500000,"operating","USA","CA","Los Angeles","Santa Monica",1,201
"15Five","software",1200000,"operating","USA","CA","SF Bay","San Francisco",2,2011
"169 ST.","games_video",50000,"closed","USA","FL","Orlando","Lake Mary",1,2009
"170 Systems","software",14000000,"acquired","USA","MA","Boston","Bedford",1,1990
"1Cast","games_video",,"closed","USA","WA","Seattle","Kirkland",1,2006
"1DayMakeover","ecommerce",50000,"closed","USA","CA","Los Angeles","Santa Ana",1,2008
"1Energy Systems","software",1450000,"operating","USA","WA","Seattle","Seattle",1,
```

Seuraavaksi tämä edellä mainittu CSV-tiedosto vietiin BigQuery-palveluun ja siitä tehtiin taulukko projektiin. Taulukon kanssa käytetty skeema oli seuraavanlainen.

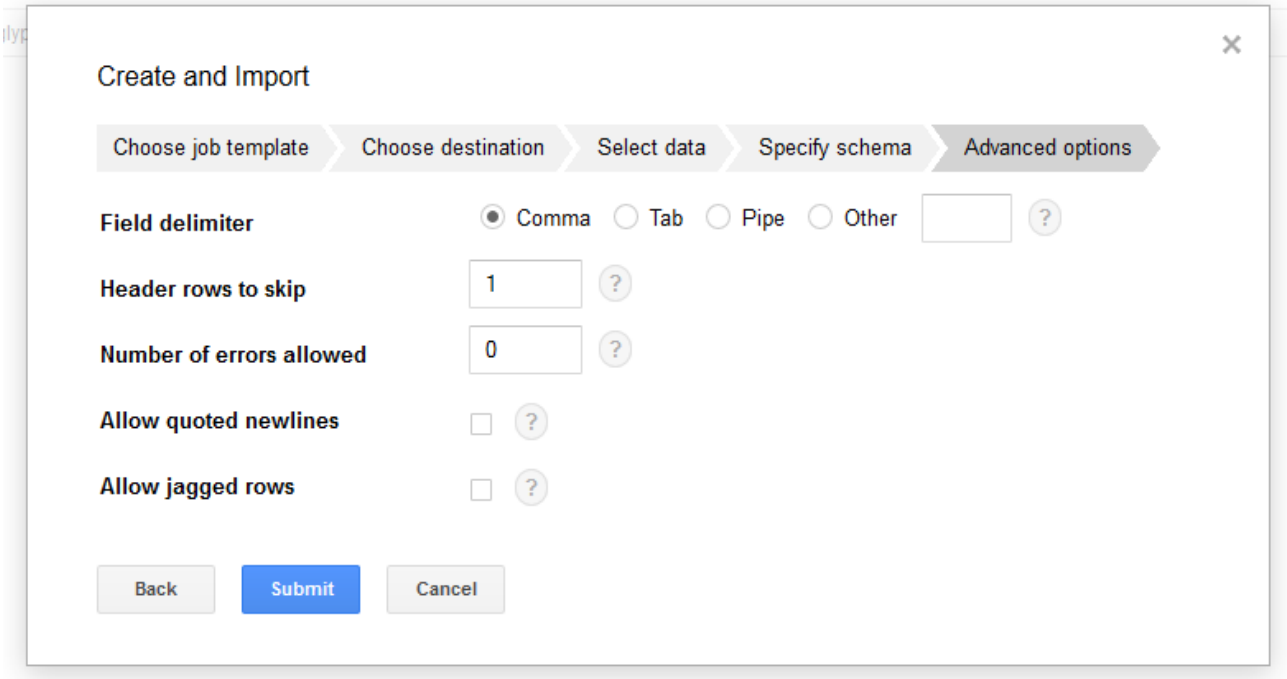
Table Details: data

Schema

name	STRING	NULLABLE
category_code	STRING	NULLABLE
funding	INTEGER	NULLABLE
status	STRING	NULLABLE
country_code	STRING	NULLABLE
state_code	STRING	NULLABLE
region	STRING	NULLABLE
city	STRING	NULLABLE
rounds	INTEGER	NULLABLE
year	INTEGER	NULLABLE

Muuttujia funding, rounds, ja year oli mahdollista käyttää integereinä. Muut sisälsivät tekstimerkkejä, joka vaati string-muuttujan käytön. Alkuperäisessä taulukossa (ensimmäinen versio) vuosi oli myös string, mutta se muutettiin integeriksi, koska se mahdollisti paremman käsittelyn funktioiden kanssa (ei pitkiä OR-ketjuja kohdassa 2008-2013). Syy stringin käyttöön alunperin oli parempi käsittely tyhjien kenttien suhteen. Niitä oli myös hankalampi paikkailla itse taulukosta, koska ne olivat viimeiset muuttujat.

Varmaankin eniten ongelmia projektissa aiheutti juurikin tämä taulukon tuonti ja se oli seurausta pääasiassa siitä, että ”Header rows to skip”-asetus oli oletuksena nolla, mutta ongelma saatiin paikannettua pikaisesti virheilmoitusten avulla. Tärkeää kyseissä kohtaa oli myös valita oikea erotin soluille, joka oli tässä tapauksessa pilkku. Ohessa kuva käytetyistä asetuksista.



Kun taulukko saatiin lopulta sisälle järjestelmään, ei SQL-käskyjen kehittely ollut kovinkaan hankalaa. Ongelmia oli pääosin ORDER BY-funktion käytössä. Alla on kysytty kysymys sen vastauksen selvitykseen käytetty koodi ja vastaus tummennetulla tai taulukossa.

1. Kuinka monta yritystä taulussasi on?
SELECT COUNT(name) FROM cb_companies_1.data;
19373

2. Kuinka monta vuonna 2013 perustettua yritystä on saanut rahoitusta?
SELECT COUNT(name) FROM cb_companies_1.data
WHERE year ="2013" AND funding > 0;
334

3. Mitä (toimialoja) category_code:ja oli käytössä vuoden 2000 jälkeen perustetuissa yrityksissä?
SELECT DISTINCT (category_code) FROM cb_companies_1.data
WHERE YEAR > 2000;
news, software, ecommerce, web, analytics, biotech, manufacturing, advertising, games_video, social, enterprise, cleantech, education, mobile, hardware, other, search, security, messaging, real_estate, network_hosting, consulting, health, music, design, nanotech, medical, semiconductor, finance, public_relations, travel, fashion, hospitality, sports, photo_video, nonprofit, legal, local, automotive, transportation, pets, government

4. Mitkä ovat suosituimmat 10 toimialaa vuonna 2002 perustetuissa rahoitusta saaneissa yrityksissä?
SELECT category_code, COUNT(name) AS maara FROM cb_companies_1.data
WHERE YEAR = 2002 AND funding > 0 AND category_code != ”null”
GROUP BY (category_code)
ORDER BY maara DESC LIMIT 10;

1	software	96
2	biotech	65
3	enterprise	23
4	hardware	19
5	cleantech	18
6	security	18
7	web	17
8	medical	17
9	semiconductor	14
10	games_video	13

5. Mitkä ovat suosituimmat 10 toimialaa vuonna 2012 perustetuissa rahoitusta saaneissa yrityksissä?

```
SELECT category_code, COUNT(name) AS maara FROM cb_companies_1.data
WHERE YEAR = 2012 AND funding > 0 AND category_code != ”null”
GROUP BY (category_code)
ORDER BY maara DESC LIMIT 10;
```

1	software	183
2	mobile	145
3	web	134
4	ecommerce	83
5	biotech	75
6	enterprise	66
7	analytics	56
8	education	55
9	social	54
10	advertising	39

6. Mitkä 10 toimialaa ovat keränneet eniten rahoitusta ja paljonko?

```
SELECT category_code, SUM(funding) AS rah FROM cb_companies_1.data
GROUP BY category_code
ORDER BY rah DESC LIMIT 10;
```

1	biotech	51624512243
2	software	31869304347
3	cleantech	29154677255
4	mobile	23274563282
5	enterprise	17474748001
6	web	13558825697
7	medical	11449549970
8	network_hosting	10414891725
9	advertising	10336147064
10	hardware	10034830717

7. Selvitä mitkä viisi toimialaa (top5) saivat eniten rahoitusta vuosina 2008-2013?

```
SELECT category_code, SUM(funding) AS rah FROM cb_companies_1.data
WHERE YEAR >= 2008 AND YEAR <=2013
GROUP BY category_code
ORDER BY rah DESC LIMIT 5;
```

1	biotech	7487828772
2	enterprise	5842189571
3	software	5043210071
4	web	3630598489
5	mobile	3599086284

Ylipäätään projekti oli melko yksinkertainen. Ongelmia oli pääosin BigQueryn kanssa, koska se ei ollut niin tuttu, myös SQL oli hieman ruosteessa, mutta se on hyvin helppoa. Dataa ei puhdistettu lainkaan vaan mahdolliset virheelliset tulokset pyrittiin kiertämään ehtojen kautta, kuten WHERE funding > 0 ja WHERE category_code != ”null”.