Haaste

Saada riveittäin allekkain olevat urheiluseurat ja niiden lyhenteet yhdistettyä kuntatietoihin kuntapohjaisen jatkoanalyysin tueksi.

Lähtödata oli alla olevan näköistä:

	A	B
1	Seura	
2	Lyhenne	
3	Alahärmän Kisa ry	
4	AlahK	
5	Alajärven Ankkurit	
6	AA	
7	Alatornion Pirkat ry	
8	ALPi	
9	Alavieskan Viri ry	
10	AV	
11	Alavuden Urheilijat ry	
12	AlavU	
13	Allianssi ry	
14	All	
15	Anjalan Liitto ry	
16	AnjLi	
17	APK-Maratoonarit Ry	
18	APK	
19	Artic Marathon Club	
20	AMC	
21	Asikkalan Raikas ry	
22	AsikRai	
23	Askolan Urheilijat ry	
24	AskU	
25	Endurance Ultrarunning Team Finland	
26	Endurance	
27	Esbo IF	

Lisäksi haettiin vuoden 2016 kuntatiedot **Master Dataksi** eli perustiedoiksi. Urheiluseuran nimen perusteella pitäisi päätellä kunta. Tiedot ladattiin tiedostoista tietokantaan jatkotyöskentelyn helpottamiseksi.

Ensiksi SQL:llä rivinumeroinnin ja modulo-funktion avulla (parilliset ja parittomat rivit) saatiin yhdessä sarakkeessa olevat tiedot jaettua kahteen eri sarakkeeseen.

```
SELECT A.DATA AS SEURA, B.DATA AS LYHENNE
FROM
(SELECT RIVINRO, DATA
  FROM [Test].[dbo].[stg_seurat]
  WHERE RIVINRO % 2 <> 0) AS A
JOIN
(SELECT RIVINRO, DATA
  FROM [Test].[dbo].[stg_seurat]
```

```
WHERE RIVINRO % 2 = 0) B
ON A.RIVINRO + 1 = B.RIVINRO
```

Results | Messages

	SEURA	LYHENNE
1	Alahärmän Kisa ry	AlahK
2	Alajärven Ankkurit	AA
3	Alatornion Pirkat ry	ALPi
4	Alavieskan Viri ry	AV
5	Alavuden Urheilijat ry	AlavU
6	Allianssi ry	All
7	Anjalan Liitto ry	AnjLi
8	APK-Maratoonarit Ry	APK
9	Artic Marathon Club	AMC

Tämän jälkeen hyödynnettiin Microsoft SQL Serverin Enterprise Editionissa olevaa tiedon mätsäys ja puhdistustyökalua ”Fuzzy Lookup”. Sen avulla sisään tulevaa arvoa voidaan verrata perustietoihin ja hakea parhaiten osuva arvo. Mätsäyksen yhteydessä voidaan säätää raja-arvoja kuinka laadukas mätsin pitää olla.

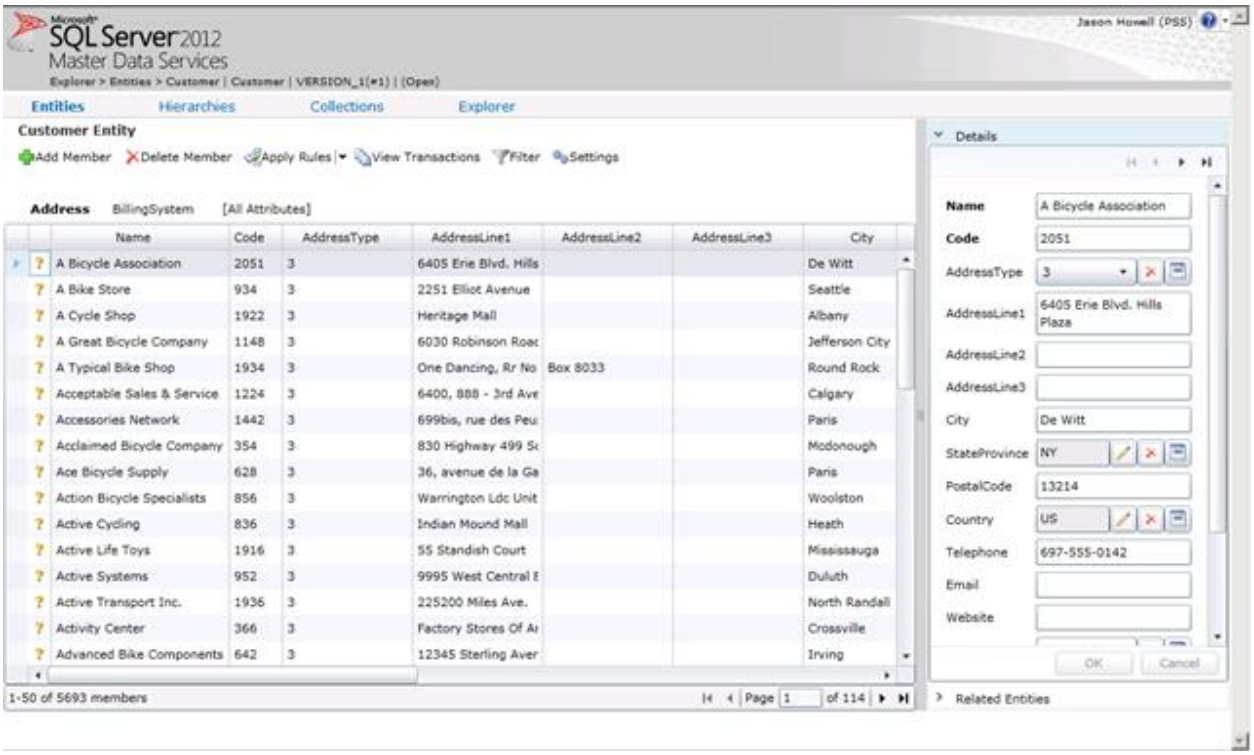
Lopputulos

Lopputuloksesta saatiin varsin kelvollinen ottaen huomioon, että osasta nimistä ei voi ollenkaan päätellä paikkakuntaa. Lopputuloksesta olisi saatu vielä tarkempi, jos kuntatiedoista olisi ladattu myös ruotsinkielinen nimi. Joka tapauksessa tällä säästettiin jo paljon aikaa näinkin pientä massaa käsiteltäessä. Hyödyt ovat paljon merkittävämmät, kun vastaava prosessi saadaan automatisoitua käsittelemään jatkuvasti hyödynnettäviä suurempia tietomassoja!

Results | Messages

	Seura	Lyhenne	Kunta	_Similarity	_Confidence	_Similarity_Nimi
233	Oulunsuun Heitto ry	OlsHe	NULL	0	0	0
234	Padasjoen Yritys ry	PadY	Padasjoki	0,2592593	0,2863302	0,2592593
235	Pakilan Visa ry	PakVi	NULL	0	0	0
236	Palosaaren Urheiluseura ry.	PUS	NULL	0	0	0
237	Pargas Idrottsförening r.f	PIF	NULL	0	0	0
238	Parikkalan Urheilijat ry	ParikU	Parikkala	0,2999995	0,8035603	0,2999995
239	Perniön Urheilijat ry	PerU	NULL	0	0	0
240	Petäjäveden Petäjäiset ry	PetPet	Petäjävesi	0,2424242	0,2815485	0,2424242
241	Pieksämäen Veikot ry	PiVe-98	Pieksämäki	0,2666667	0,2875984	0,2666667
242	Pihtiputaan Tuisku ry	PIHTU	Pihtipudas	0,2364246	0,2833724	0,2364246
243	Polvijärven Urheilijat ry	PoU	Polvijärvi	0,2727272	0,2861978	0,2727272
244	Polvijärven Vauhti ry	POLVA	Polvijärvi	0,2727272	0,2877661	0,2727272
245	Porin Marathonyhdistys ry (Maratonk...	MkP	Pori	0,2	0,303242	0,2
246	Porin Tarmo ry	PorTa	Pori	0,2665788	0,2912095	0,2665788
247	Porin Urheiluveteraanit ry	PUV	Pori	0,2665788	0,2888695	0,2665788
248	Porin yleisurheilu	PorY	Pori	0,3998682	0,2902603	0,3998682
249	Pornainen Heavy Team ry	porHT	Pornainen	0,25	0,3083218	0,25
250	Porvoon Akilles ry	PorvA	Porvoo	0,2857058	0,2926582	0,2857058
251	Porvoon Urheilijat ry	PorvU	Porvoo	0,2857058	0,2912923	0,2857058

Loppu on sitten puhdasta manuaalityötä sekä lopputuloksen varmistuksen, että puuttuvien tietojen lisäyksien osalta. Esimerkiksi SQL Server ympäristössä nämä voidaan toteuttaa MDS:llä (Master Data Services) joko selainpohjaisen tai Excel-käyttöliittymän avulla. Nörtti tekee asiat tietenkin suoraan tietokantaan, mutta business-ihmiselle kyseiset käyttöliittymät ovat varsin mukavia etenkin, jos ylläpitäjiä on useampia.



Saatua tietoa hyödynnettiin Kalevan-kisojen tulosten analysointiin kuntapohjaisesti. Urheilijan osalta oli mainittu vain seuralyhenne, jonka takia yllä kuvattu prosessi toteutettiin kuntatiedon saamiseksi. Mutta sen analyysin tulos onkin sitten toinen tarina.