Uporaba tehnologij Microsoft Analysis Services (OLAP kocka) nad Yelp podatkovno bazo

Seminarska naloga pri predmetu TEHNOLOGIJE UPRAVLJANJA PODATKOV

Primož Hrovat

Mentor: Luka Šajn

december 2017, januar 2018

Povzetek

Yelp je ameriška multinacionalka, s sedežem v San Franciscu v Kaliforniji. Razvija, gosti in promovira Yelp.com in Yelp mobilno aplikacijo, ki objavlja ocene in recenzije skupnosti o lokalnih podjetjih. Namen naloge je bil spoznati razširitev Microsoftovega SQL Serverja, Analysis Services, ki omogoča dokaj enostavno gradnjo večdimenzionalnih modelov (kock), za procesiranje in pripravo pogledov na različne aspekte podatkov in zakonitosti, ki se v teh zbranih podatkih skrivajo. Yelp kot del tekmovanja »Yelp dataset challenge« del njihovih podatkov objavlja javno in ti podatki so bili tu uporabljeni za gradnjo takšnih modelov.

Kazalo

Tabele

Slike

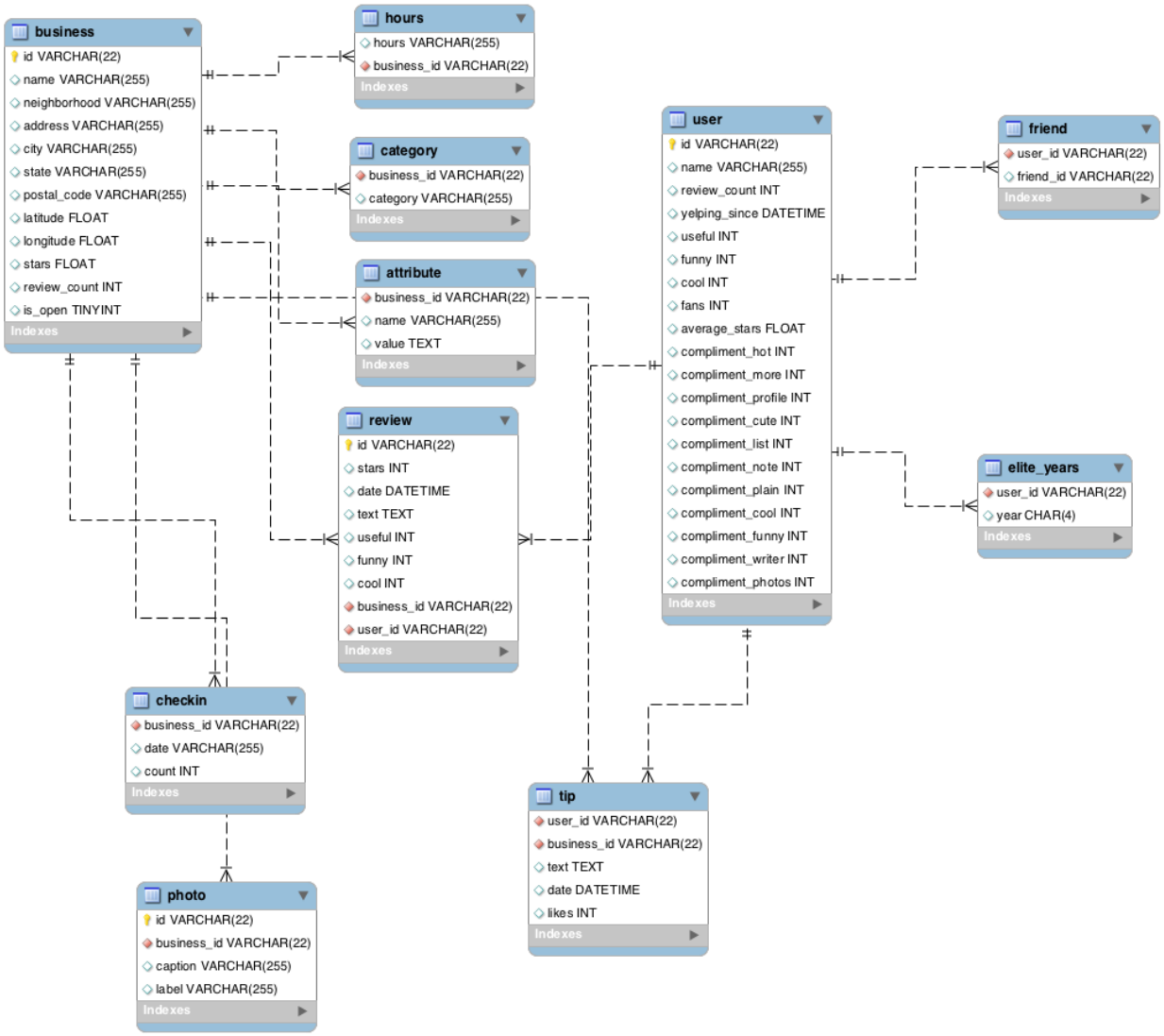
1 Uvod

Pri predmetu Tehnologije upravljanja podatkov sem v okviru seminarske naloge želel analizirati podatke iz podatkovnega nabora Yelp. Prvotno je bila naloga zastavljena v smeri oblačnega gostovanja baze in razširitev na Microsoftovem oblaku Azure. Vendar se je takoj na začetku pojavil problem, ker je v okviru študentske licence, ki jo lahko študent FRI-ja pridobi preko programa IMAGINE, v SQL podatkovno bazo možno brezplačno naložiti le 32MB podatkov. Podatki v tej javni zbirki obsegajo slabih 7GB, kar je privedlo do tega, da se je problema potrebno lotiti nekoliko drugače. Alternativna možnost se je pokazala v uporabi brezplačnega gostovanja virtualnega strežnika na grškem akademskem oblaku Okeanos. Ta deluje v sklopu omrežja Géant, kar članicam omogoča brezplačni dostop.

2 Priprava podatkovne baze

2.1 Podatki

Podatki s katerimi se v tej nalogi ukvarjam, so s spletnega mesta Yelp, dostopni pa so preko programa »Yelp Dataset Challenge«, ki je tekmovanje, namenjeno študentom in sicer z namenom iskanja zakonitosti v podatkih (strojno učenje). Zaradi same velikosti, pa se jo lahko uporabi tudi za namene gradnje OLAP kock. V sami podatkovni bazi se nahajajo ocene in recenzije lokalnih podjetji, ki jih ocenjujejo uporabniki tega spletnega portala. Shema podatkovne baze, dostopna tudi prek spleta, je na sliki 1. Skupna velikost te baze nanese slabih 7GB, podatki so dostopni v obliki sql, ki ga podpirata sistema MySQL in PostgreSQL, ali v obliki JSON. Med dostopnimi podatki so tudi opisi podjetji, vključno z delovnim časom, kategorizacijo (uporabniško določena), lokacijo, prav tako pa uporabniki lahko prispevajo svoje nasvete, povezane s posameznim podjetjem. Podprto je tudi dodajanje slik.



Slika 1: shema podatkovne baze Yelp

2.2 Namestitev podatkovne baze

Z namenom ohraniti strežniško okolje ločeno od razvojnega, predvsem pa enostavnosti, sem podatkovno bazo in sistem za upravljanje s podatkovno bazo, namestil in gostoval na grškem omrežju Okeanos, ki akademski sferi omogoča brezplačno gostovanje virtualnih instanc (do 80GB diskovnega spomina, do 4GB delovnega spomina, do 4 procesorska jedra), na katere je možno namestiti kopico ponujenih operacijskih sistemov, med njim najdemo nekatere bolj uveljavljene Linux distribucije in tudi Windows Server 2012. Slednjega sem tudi uporabil in nanj dodatno namestil Microsoftov SQL Server 2017 (MS SQL) z nameščeno razširitvijo Analysis Services. V sklopu razvijalske licence so storitve na voljo brezplačno, so pa performanse nekoliko okrnjene. Podatkov neposredno ni moč uvoziti, ker so v neprimerni obliki, zato je bil potreben še dodaten korak, in sicer namestitev sistema MySQL in pa Microsoftovega orodja za migracijo podatkov med sistemoma.

2.3 Priprava podatkov in migracija

Podatke sem najprej s skripto, dostopno na Yelp strani prenesel v sistem MySQL. Z uporabo Microsoft Migration Assistant for MySQL je nato moč podatke tudi migrirati v sistem MS SQL. Postopek prenosa je avtomatiziran in nezahteven, potrebna je namreč samo preslikava sheme, kar zahteva samo par klikov, nato pa se lahko nad bazama požene program za migracijo. Uspešnost prenosa je opisana v spodnji tabeli, je pa skrb vzbujajoča na splošno nizek odstotek uspešnosti prenosov.

Tabela 1: migracija podatkov iz MySQL v MS SQL

Originalna tabela	Ciljna tabela	Uspešnost %
yelp_db.attribute	yelp_db.attribute	72,73
yelp_db.bussiness	yelp_db.bussiness	100
yelp_db.category	yelp_db.category	100
yelp_db.checkin	yelp_db.checkin	29,42
yelp_db.elite_years	yelp_db.elite_years	100
yelp_db.friend	yelp_db.friend	NA
yelp_db.hours	yelp_db.hours	100
yelp_db.photo	yelp_db.photo	100
yelp_db.review	yelp_db.review	5,21
yelp_db.tip	yelp_db.tip	42,52
yelp_db.user	yelp_db.user	30,95

Slabemu odstotku migracije navkljub sem za občutek poskusil ustvariti začetno bazo za OLAP analizo, vendar se izkaže, da za dostop do Analysis Services potrebujemo uporabniško ime in geslo lokalne domene Windows Active Directory. Precej nepraktično, glede na to, da MS SQL poleg avtentikacije z AD omogoča tudi interno, SQL avtentikacijo. Stanje je bilo potemtakem tako: do podatkov je bilo moč priti, ustvarjati/urejati analitične podatkovne baze pa praktično nemogoče brez ustvarjanja lokalnih računov v AD in dodatne konfiguracije odjemalcev. Enostavnejša rešitev se je izkazala v lokalni namestitvi MS SQL, obenem pa tudi priložnost za ponovni poskus vnosa podatkov v bazo.

Praktično neuporabno malo podatkov se je preneslo prav v shemah, kjer se skriva največ »relavantnih« informacij – »review« in »user«. Ker bi delo nad tako okrnjenim naborom podatkov predstavljalo veliko napako v interpretaciji samih rezultatov, sem ubral drugačen pristop. Uporabil sem zapis podatkov v obliki JSON in s pomočjo dodatne SQL skripte »ročno« prenesel podatke v MS SQL. Osredotočil sem se predvsem

na tabele, ki mi predstavljajo potencialno večjo vrednost in tako recimo tabele s fotografijami sploh nisem prenašal. Uspešnost prenosa je bila tu 100%, je pa zaradi narave JSON hrambe podatkov dobljena shema denormalizirana. Delna normalizacija je bila izvedena takoj po prenosu, z ustvarjanjem novih tabel in razbijanjem obstoječih.

2.4 Kreiranje projekta

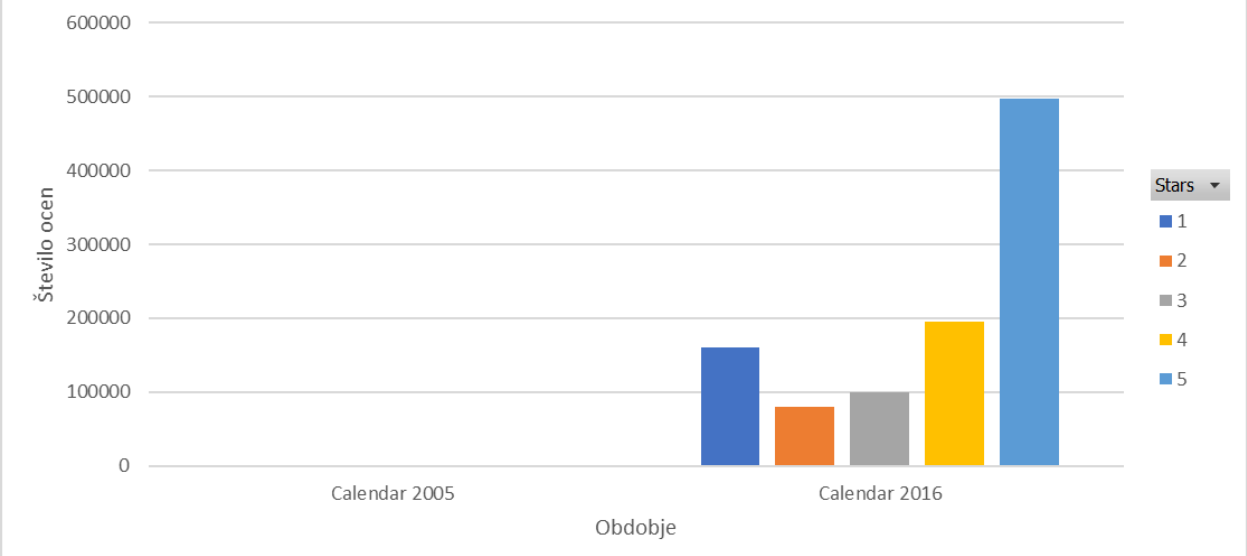
Za povezavo, urejanje in ustvarjanje podatkovnih modelov se v okolju Windows lahko uporablja programski paket SSDT (SQL Server Data Tools), ki je v osnovi razširitev Visual Studia, omogoča pa načrtovanje, oblikovanje in izvedbo modelov pretežno z uporabo grafičnih elementov. Nismo zavezani uporabi tega orodja, na voljo so tudi druge metode, recimo preko Microsoft SQL Server Management Studia (SSMS), vendar je tu vse stvari potrebno ustvariti in popravljati ročno.

Postopek gradnje modelov poteka nekako takole: znotraj orodja SSDT se ustvari povezavo/e na podatkovno bazo, ki nam služi kot vir podatkov. Nad zbranimi podatki lahko oblikujemo poglede, tako da kreiranje, urejanje projekta ne zahteva stalne povezave na vire. Okolje nas nato preko pozivnih oken vodi pri ustvarjanju osnovne kocke, kateri dodajamo dimenzije, po kateri želimo podatke filtrirati. Najpomembnejša stvar so seveda mere (Measures), v katerih definiramo operacije (večinoma so to agregacije) nad podatki. Mere in dimenzije je potrebno med seboj povezati, model pa se nato lokalno zgradi (build), v primeru uspeha pa se ga lahko tudi prenese na strežnik, kjer se dejansko izvaja Analysis Services baza. Ko se model uspešno namesti na strežnik, se podatki sprocesirajo in pripravijo osnovne dimenzije, definirajo mere in pravila, ki smo jih navedli v projektu. Do ustvarjenih modelov, v tem primeru kocke, imamo potem dostop preko množice različnih orodij, kot so Reporting Services, Excel, Visual Studio…

Začetni model kocke sem iterativno izboljševal, vsaka iteracija in testiranje pa s seboj potegne nekaj časa, ko se stvari na strežniku dejansko procesirajo. Do podatkov sem dostopal iz programa Excel, in ustvarjal vrtljive tabele in grafikone, s katerimi sem lahko na podatke pogledal z različnih zornih kotov (uporaba dimenzij). Naknadno sem modelu dodal tudi časovno dimenzijo, ki sem jo potreboval.

3 Analiza podatkov

V nadaljevanju so predstavljeni rezultati v obliki grafikonov.

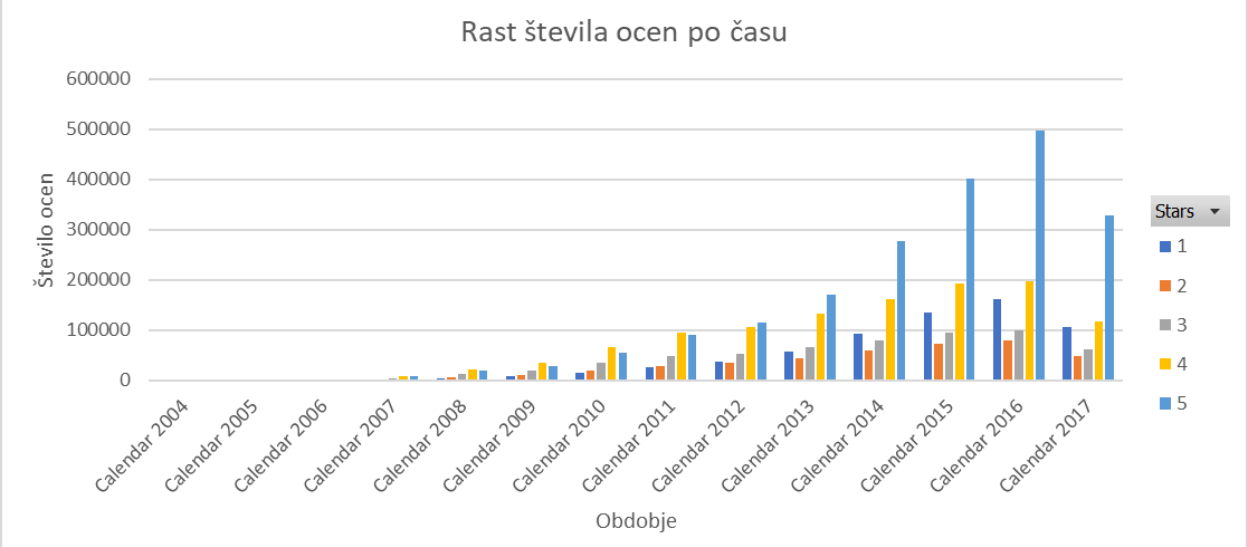


Slika 2: primerjava števila ocen za leto 2005 in 2016

Tabela 2: Število komentarjev leta 2005 in 2016

Review Count	Oznake stolpcev					
Oznake vrstic	**1**	**2**	**3**	**4**	**5**	**Skupna vsota**
Calendar 2005	26	34	152	341	313	866
Calendar 2016	160519	79751	99087	195925	497842	1033124
Skupna vsota	**160545**	**79785**	**99239**	**196266**	**498155**	**1033990**

Dober občutek o tem, kako je popularnost spletne storitve Yelp rasla, je pregled števila komentarjev, ki so jih uporabniki ustvarili skozi leta. Zgornji graf in tabela prikazujeta razliko med letom 2005, in 11 let kasneje. Spodnji graf in pripadajoča tabela pa prikazujeta, kako število recenzij narašča skozi leta, ko se je Yelp čedalje bolj uveljavljal. Podatki za leto 2017 obsegajo samo 3 četrtletja, kar pojasnjuje manjše skupno število recenzij.

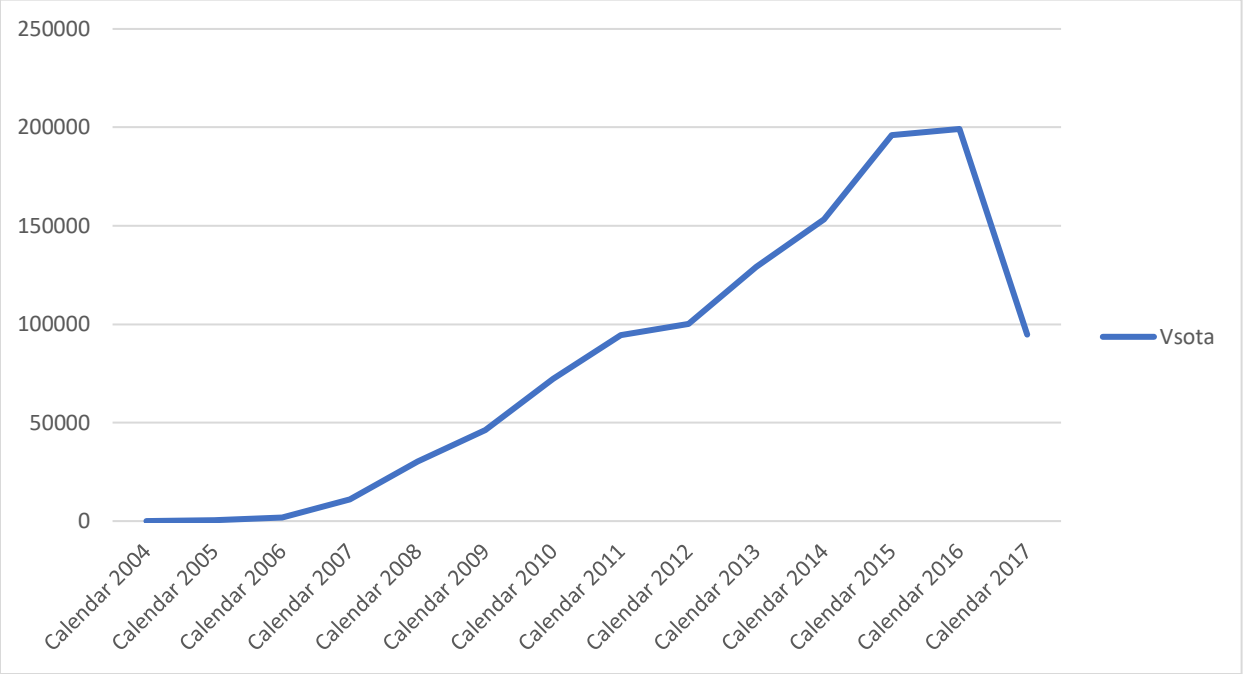


Slika 3: Pregled števila komentarjev od ustanovitve naprej

Tabela 3: Pregled števila komentarjev od ustanovitve naprej

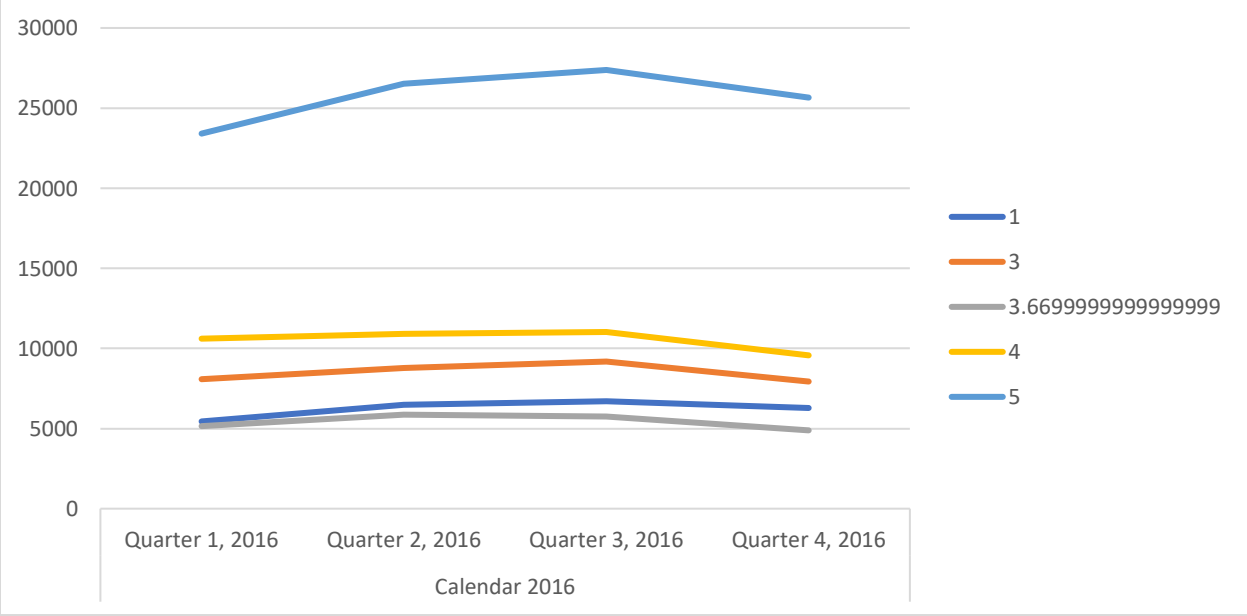
Oznake vrstic	Review Count
Calendar 2004	14
Calendar 2005	866
Calendar 2006	5642
Calendar 2007	22895
Calendar 2008	61093
Calendar 2009	97834
Calendar 2010	185643
Calendar 2011	287814
Calendar 2012	345637
Calendar 2013	468608
Calendar 2014	670440
Calendar 2015	897835
Calendar 2016	1033124
Calendar 2017	659452
Skupna vsota	**4736897**

Uporabniki imajo možnost poslovne subjekte označiti kot »useful«, »funny«… Spodnji diagram prikazuje, število recenzij z oznako »cool«.



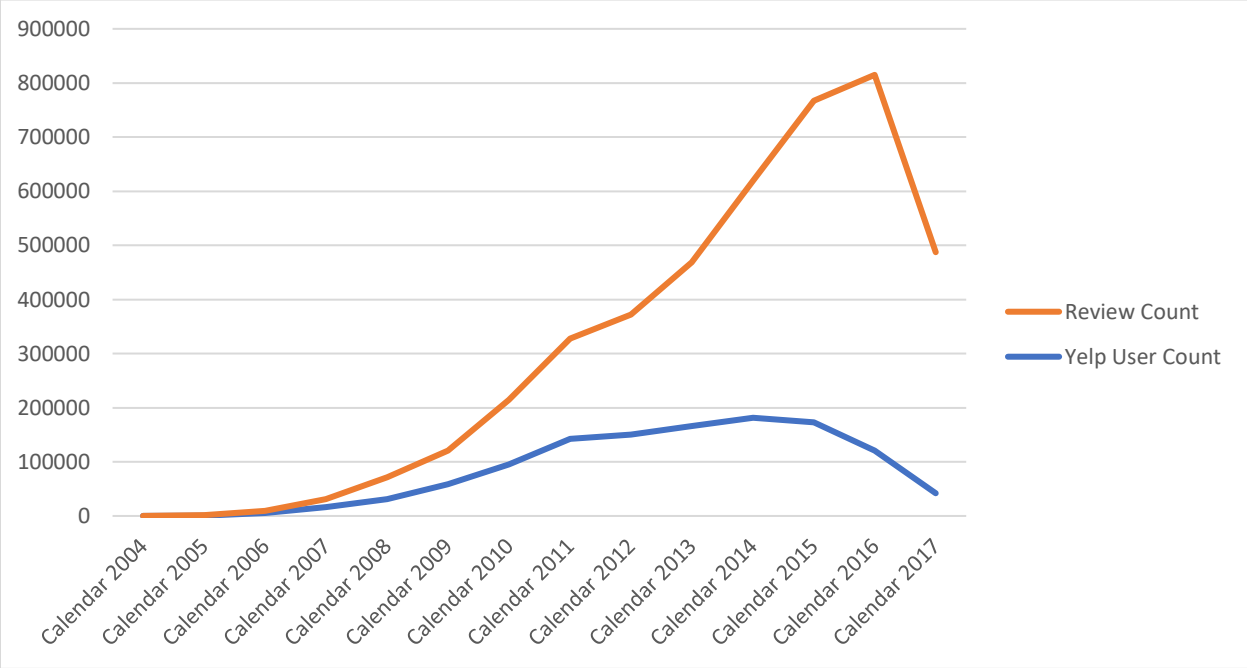
Slika 4: Število novih "useful" komentarjev skozi leta

Eden zanimivejših pogledov je pregled najpogostejših povprečnih ocen podjetji, uporabnikov, ki so označeni kot »elite«. Prikazano je število recenzij za »elitne« uporabnike z najpogostejšo povprečno oceno za leto 2016 po četrtletjih. Kar je zanimivo je to, da so uporabniki nagnjeni k podeljevanju predvsem višjih ocen (3, 4, 5).



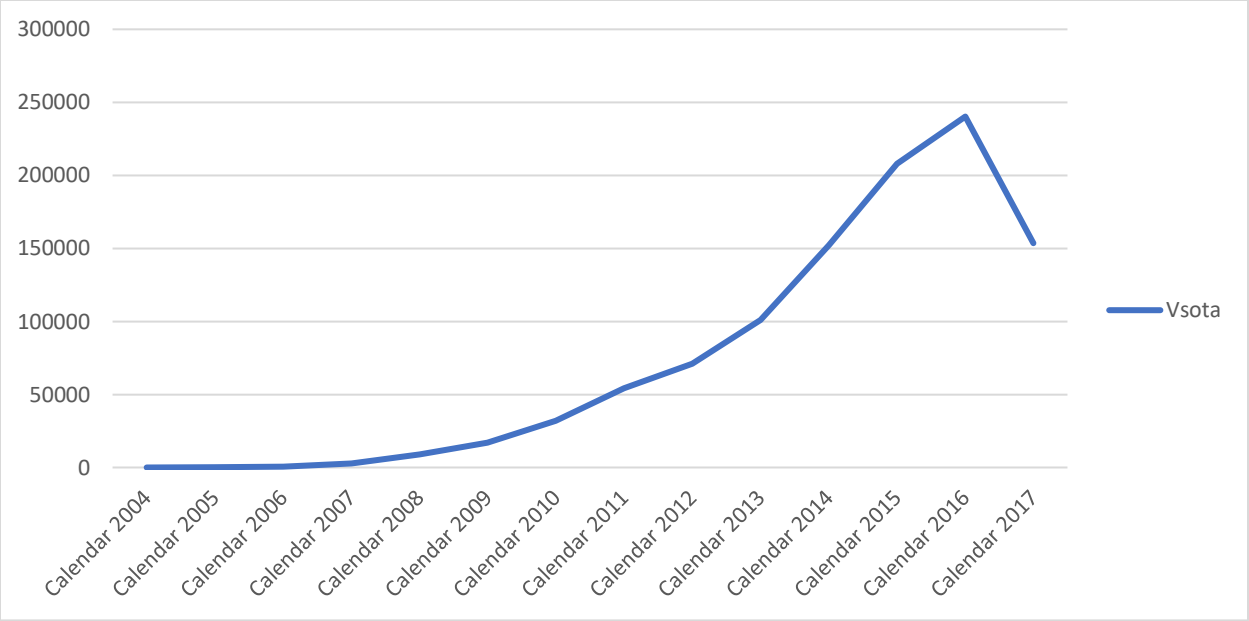
Slika 5: Najpogostejše povprečne ocene "elite" uporabnikov

Če primerjamo rast števila novih registriranih uporabnikov glede na število novo dodanih komentarjev se nam izriše zanimiva slika, ki kaže, da se število komentarjev povečuje nesorazmerno hitreje, kot raste število novih uporabnikov. To pomeni, da so uporabniki (ali vsaj delež njih) zelo aktivni. Natančneje, na spodnjem grafu je prikazano število recenzij z oceno 4 ali 5, z modro pa število novo registriranih uporabnikov.



Slika 6: Recenzije (4, 5) in število novo registriranih uporabnikov

Za primerjavo: na spodnjem diagramu je prikazano število podeljenih recenzij z ocenami 1 ali 2. Razlike so bolj opazne v podanih tabelah.

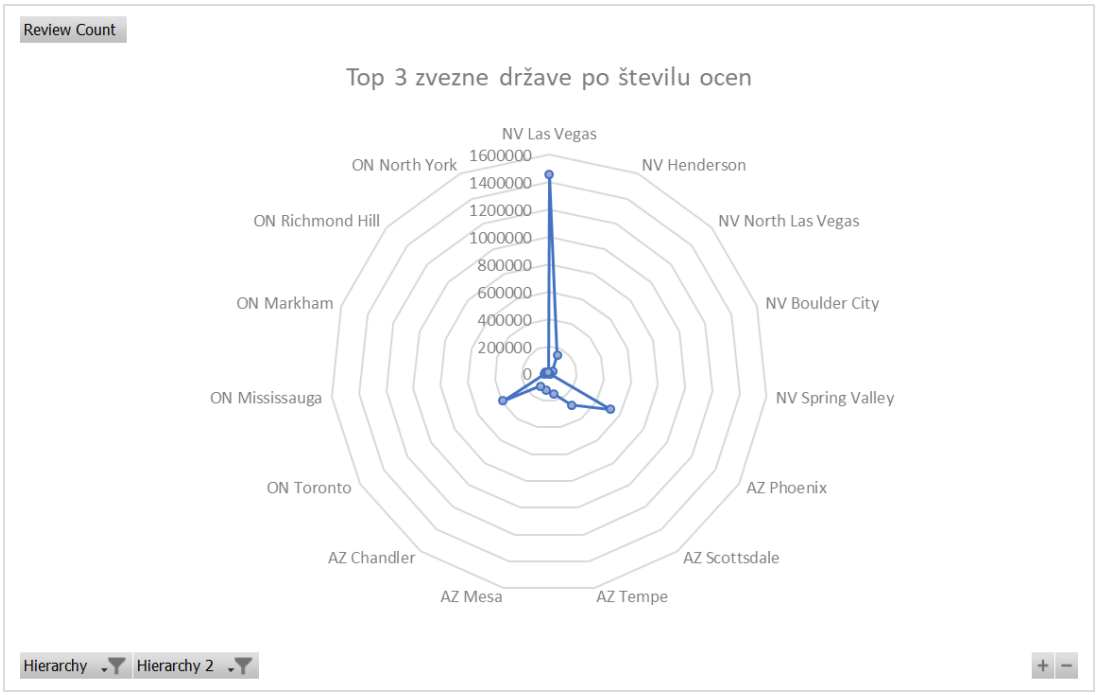


Slika 7: Recenzije z ocenami 1 ali 2

Tabela 4: Število recenzij

Oznake vrstic	Ocena 4 ali 5	Ocena 1 ali 2
Calendar 2004	12	1
Calendar 2005	654	60
Calendar 2006	3887	684
Calendar 2007	15625	2892
Calendar 2008	40116	9075
Calendar 2009	62795	17084
Calendar 2010	119717	31993
Calendar 2011	184970	54403
Calendar 2012	221463	71025
Calendar 2013	302711	100938
Calendar 2014	438204	152183
Calendar 2015	594294	207924
Calendar 2016	693767	240270
Calendar 2017	445618	153713
Skupna vsota	**3123833**	**1042245**

Še en zanimiv pogled na podatke je, kje se »zgosti« število recenzij. Podatek je kar presenetljiv, in sicer mar 69,98% vseh recenzij je pripisanih 15 mestom v 3 zveznih državah.



Slika 8: 15 mest z največ recenzijami (po zveznih državah)

Tabela 5: 15 mest z največ recenzijami

Oznake vrstic	Review Count
NV	
Las Vegas	1456806
Henderson	147562
North Las Vegas	33059
Boulder City	7523
Spring Valley	1063
AZ	
Phoenix	519246
Scottsdale	279716
Tempe	148074
Mesa	117960
Chandler	109246
ON	
Toronto	391792
Mississauga	37877
Markham	33923
Richmond Hill	16153
North York	14731
Skupna vsota	**3314731**

4 Zaključek

Za zaključek najprej nekaj opomb na rezultate, ki so predstavljeni v tem poročilu. Potrebno je izpostaviti, da je set podatkov močno okrnjen. Glede na podatke s spletne strani Yelp[1] je bilo do konca tretjega četrtletja 2017 skupaj napisanih kar 142 milijonov recenzij – v dostopnem setu jih je »le« slabih 5 (natančneje 4736897). Prav tako je pričakovano, da je posledično manj poslovnih subjektov, le podmnožica vseh uporabnikov... Skratka, rezultati, pridobljeni na tej množici ne nujno odražajo dejanskega stanja. Glede na vse »ad-hoc« poizvedbe, ki sem jih izvedel tekom priprave naloge, sem precej trdno prepričan, da so podatki iz zelo omejenega seta podatkov, ki se nanaša samo na podjetja iz Združenih držav Amerike. Verjetno je temu tako, ker podatki primarno niso namenjeni obdelavi z takšnimi OLAP modeli, ampak se spodbuja iskanje zakonitosti v njih s pomočjo metod strojnega učenja, kar pa ni bila tema te naloge. Prav z metodami strojnega učenja, bi lahko iz podatkov »izvlekli« še kakšen drobec ali pa presenetljivo veliko informacije, ki je skrita globoko v njih.

Pri izdelavi naloge sem predvsem pridobil občutek, da delo z velikimi količinami podatkov zahteva bistveno več časa, kot s šolskimi primeri, zato je bistvenega pomena, da so podatkovne baze in skladišča, ki hranijo te podatke, dobro in premišljeno načrtovana in vzdrževana. Prav tako pri pregledovanju teh ogromnih količin podatkov, običajna orodja, ki so v vsakdanjem življenju uporabna, tukaj odpovejo na celi črti. Kot nadgradnja tega raziskovanja po podatkih bi omenil možnost uporabe katere od metod strojnega učenja, uporaba Analysis Services pa je bolj primerna za agregiranje podatkov iz več podatkovnih virov in modeliranje ter projeciranja teh agregatov.

[1] Yelp, https://www.yelp.com/about

5 Bibliografija

1. **Yelp.** ***Yelp.*** **[Elektronski] https://www.yelp.com/about.**

2. Yelp dataset. ***Yelp dataset challenge.*** **[Elektronski] https://www.yelp.com/dataset/challenge.**

3. Microsoft. ***Microsoft Analysis Services.*** **[Elektronski] https://docs.microsoft.com/en-us/sql/analysis-services/analysis-services.**