Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.41
DOI: https://doi.org/10.24867/04BE03Kadic

ARHITEKTURA I IMPLEMENTACIJA SISTEMA ZA ANALIZU PODATAKA PAMETNOG GRADA

AN ARCHITECTURE AND IMPLEMENTATION OF A SYSTEM FOR SMART CITY DATA ANALYSES

Simonida Kadić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu prezentovan je sistem skladišta podataka čijom je upotrebom moguće vršiti analize kretanja ljudi na području grada Melburna u Australiji. U prvom delu rada opisani su izvori podataka. Podaci su prikupljeni iz baza podataka vladinih agencija grada Melburna. Dobavljanje podataka zahtevalo je razvoj podsistema za ekstrakciju podataka. Neophodno je bilo izvršiti eksplorativnu analizu prikupljenih podataka, kako bi se razvila OLTP šema baze podataka i programi za punjenje njenih tabela. Za projektovanje dimenzija i činjenica OLAP šeme baze podataka bilo je potrebno uočiti poslovne procese koji su od značaja za analizu kretanja ljudi. Takođe, razvijen je ETL proces čijom primenom se realizuje punjenje OLAP baze podataka. Izveštajna funkcija omogućava parametrizaciju upita.*

Ključne reči: *Sistemi skladišta podataka, OLTP, OLAP, ETL proces, Pametan grad.*

Abstract – *In this paper we present a data warehouse system that provides analyses of the movement of people in the city of Melbourne, Australia. In the first part of the paper we describe sources of data, provided by government agencies of the city of Melbourne. The extraction of data required the development of scraping programs. The development of the OLTP database schema and programs for its loading were based on exploratory data analysis. Next, a set of business processes was established to drive the development of dimensions and facts of the data warehouse. An ETL process was developed to load data into the data warehouse. A user interface was developed to allow customization of reports. Finally, the paper provides reports based on the collected data.*

Ključne reči: *Data warehouse system, OLTP, OLAP, ETL process, Smart city.*

1. UVOD

U vreme rasta populacije već razvijenih gradova, tehnološke inovacije počinju da se primenjuju u korist boljeg svakodnevnog života građana. Podaci sa različitih senzora omogućavaju sprovođenje odluka koje proizilaze iz izveštaja iz prikupljenih podataka.

Uvođenje informacione tehnologije u proces praćenja različitih parametara bitnih za funkcionisanje gradova i njihovih promena, doprinelo je pojavi koncepta pametnog grada (eng. Smart City) [1]. Pametan grad karakteriše primena mnogobrojnih senzora za prikupljanje podataka.

Istraživanje je zasnovano na podacima preuzetim za grad Melburn u Australiji, kao primer gusto naseljenog grada koji je na visokom stepenu kulturnog i ekonomskog razvoja. Australijske državne agencije već duži vremenski period objavljuju skupove podataka kao odgovor na deklaraciju o otvorenoj vladi [2]. Omogućen je i pristup podacima u realnom vremenu sa mnogih javnih senzora.

Decentralizacija podataka prikupljenih od strane različitih agencija onemogućuje jedinstven pristup podacima. Agencije objavljuju podatke u različitim formatima vremena i datoteka ili na odvojenim portalima. Podaci objavljeni na ovaj način moraju se dovesti u korelaciju i ujednačen format pre sprovođenja analiza. Količina prikupljenih podataka dodatno otežava njihovu analizu. Transakcije koje bi pružile izveštaje iz polaznih podataka ne mogu se izvršiti za vreme koje bi bilo adekvatno za praćenje podataka u skoro realnom vremenu. Zbog toga, podatke pojedinačnih izvora najpre je potrebno „očistiti“ od grešaka, a zatim ih dovesti u oblik koji bi omogućio kraće vreme generisanja izveštaja.

Različiti formati podataka, velike količine podataka, potrebe za „čišćenjem“ podataka i njihovo dovođenje u korelaciju, impliciraju potrebu za informacionim sistemom koji bi omogućio parametrizovano generisanje izveštaja. Ovaj rad opisuje postupak izgradnje informa-cionog sistema za analizu podataka pametnog grada. Podaci Melburna iskorišćeni su kako bi iz sistema proizišli realni izveštaji. Informacioni sistem pokriva proces praćenja saobraćaja automobila, pešaka, biciklista na nivou predgrađa. Cilj ovog rada je da se na osnovu prikupljenih podataka o kretanju ljudi dobiju informacije o gustini pešačkog i biciklističkog saobraćaja, kao i informacije o upotrebi parking mesta, u zavisnosti od obuhvaćenog vremenskog intervala, meteoroloških uslova i karakteristika posmatrane lokacije. Rezultati analiza prikupljenih podataka treba da odgovore na sledeća pitanja:

- Kako je kretanje pešaka, biciklista i upotreba parking mesta raspoređeno u toku različitih vremenskih intervala?
- Kako vremenski uslovi utiču na broj zabeleženih pešaka, biciklista i parkiranja?
- Kako je kretanje pešaka, biciklista i upotreba parking mesta raspoređeno na nivou zone grada?

Rezultati analiza mogu da ukažu na mesta koja zahtevaju više ili manje parking mesta, mesta čestih gužvi u saobraćaju ili mesta gde je potrebno podstaći saobraćaj pešaka i biciklista. Termini izvođenja radova mogu da se planiraju

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Ivan Luković, red. prof.

za periode kada će najmanje ometati saobraćaj. U toku izvođenja radova može se planirati preusmeravanje saobraćaja. Sistem omogućava procenu stanja u saobraćaju u toku radova i nakon završetka radova.

2. IZVORI PODATAKA

Podaci su preuzeti putem Internet stranica agencija vlade Australije. Sledeći skupovi podataka su preuzeti:

- podaci o predgrađima Melburna,
- podaci o poslovnim preduzećima,
- podaci o stambenim prostorima,
- podaci o parking senzorima,
- podaci o senzorima biciklističkih staza,
- podaci o meteorološkim stanicama,
- podaci o senzorima pešaka i
- podaci o senzorima kvaliteta vazduha.

Eksplorativnom analizom utvrđene su nepravilnosti u podacima. Do 33% senzora parking mesta nema zabeleženu lokaciju. Postoje očitavanja parkiranja koja nisu jedinstvena ili imaju negativnu dužinu trajanja parkinga. Postoje parkiranja koja se sukcesivno nastavljaju tako da kraj jednog parkiranja je u trenutku početka sledećeg parkiranja. Utvrđeno je da se takve sekvence protežu kroz više sati, odnosno više dana pa čak i meseci. Sukcesivna parkiranja su u nekim slučajevima vrlo kratka i traju nekoliko sekundi, što doprinosi broju zabeleženih parkiranja u toku jednog sata. Na primer, na jednom parking mestu 19. januara 2012. u 11 sati zabeležena su 202 parkiranja. Očigledno je da takav broj parkiranja u toku jednog sata je malo verovatan, kao i da ovakve nepravilnosti bitno doprinose broju očitavnja sa senzora. Od 6.250 senzora čiji položaj je poznat, 3.125 senzora (50%) ima zabeleženo preko 2.037 parkiranja u toku 2016. Godine. Neki senzori imaju očitavanja samo u toku jednog meseca. Takvi senzori su zabeležili ukupno preko 200 parkiranja, odnosno više parkiranja u intervalu od par meseci, nego senzori bez sekvencijalnih parkiranja u toku cele godine. Utvrđeno je da upravo takvi senzori imaju sukcesivna parkiranja.

U početnim podacima za 43 senzora pešaka postoji 390.423 zapisa koja pokazuju koliko pešaka je zabeleženo u toku predhodnog sata u 2016. godini. Idealno, svaki senzor bi imao 8.784 zapisa za svaki sat u toku 2016. godine, odnosno 24 zapisa po danu. Utvrđeno je da postoje nepravilnosti zbog kojih broj zapisa ni na jednom senzoru ne odgovara ovoj pretpostavci. Ove nepravilnosti nastaju usled dupliranih zapisa, nepostojećih zapisa i lošeg rešenja za letnje računanje vremena.

Duplikate čini do 8% očitavanja pešačkih senzora u 2016. godini. Nakon uklanjanja duplikata, senzori imaju najviše 8.783 očitavanja u toku 2016. godine. 3. Aprila 2016. godine prestalo je letnje računanje vremena i časovnik je pomeren sa 3 (tri) na 2 (dva) sata. U toku ovog datuma nema ni višak ni manjak zapisa. Ovde je problematično to što dolazi do gubljenja podataka jednog sata, s obzirom na to da vreme od 2 (dva) sata nastupa dva puta u toku tog dana: prvi put po letnjem računanju vremena, a zatim i po zimskom računanju vremena. Takođe, ne zna se da li zapisi očitani u 2 (dva) sata ujutru odgovaraju letnjem ili zimskom računanju vremena.

Senzori pešaka koja imaju očitavanja do poslednjeg dana u godini, a prvo očitavanje pre oktobra, nemaju očitavanje 2. oktobra 2016. godine u 2 sata. Ovo je sat koji nedostaje senzorima koji imaju 8.783 umesto 8.784 očitavanja nakon uklanjanja duplikata. Upravo 2. oktobra 2016. godine u 2 sata počelo je letnje računanje vremena i časovnik je pomeren za jedan sat u napred. Za taj datum i vreme nema zapisa ni za jedan senzor. Međutim, 1 (jedan) sat po zimskom računanju vremena i 3 (tri) sata po letnjem računanju vremena preslikavaju se na susedne sate po griničkom vremenu, zbog čega takvi zapisi ne treba ni da postoje.

Navedene nepravilnosti utiču na rezultate izveštaja, zbog čega se moraju očistiti putem *ETL* procesa. Detaljniji opis izvora podataka prikazan je u master radu autora.

3. ŠEMA OLTP BAZE PODATAKA

OLTP baza podataka (eng. *On-Line Transaction Processing,* transakciono procesiranje podataka) predstavlja izvor podataka za OLAP (eng. *On-Line Analytical Processing,* analitičko procesiranje podataka) bazu podataka. Sastoji se od sledećih tabela: tabela predgrađa - sadrži nazive i granice predgrađa, tabela stambenih prostora - sadrži podatke sa popisa stambenih prostora, tabela poslovnih preduzeća - sadrži podatke sa popisa poslovnih preduzeća, tabela pešačkih senzora - sadrži podatke o pešačkim senzorima, uključujući lokaciju i identifikator senzora, tabela očitavanja pešačkih senzora - sadrži očitavanja pešačkih senzora uključujući identifikator senzora, tabela lokacija biciklističkih senzora - sadrži podatke i identifikatore lokacija na biciklističkim stazama gde su postavljena najčešće dva senzora koja su usmerena u suprotnim smerovima, tabela biciklističkih senzora - sadrži podatke i identifikatore biciklističih senzora sa identifikatorom lokacije na kojoj se nalaze, tabela očitavanja biciklističkih senzora - sadrži podatke o očitavanjima biciklističkih senzora, uključujući identifikator lokacije senzora i identifikator senzora, tabela lokacija parking senzora - sadrži identifikatore i lokacije na kojima se nalazi više parking senzora, tabela očitavanja parking senzora - sadrži podatke očitavanja parking senzora, uključujući identifikator lokacije senzora i identifikator parking senzora, tabela meteoroloških stanica - sadrži podatke o mestima očitavanja meteoroloških uslova, uključujući identifikatore meteoroloških stanica, tabela očitavanja meteoroloških stanica - sadrži podatke o očitavanjima meteoroloških uslova, uključujući opis očitanog parametra i identifikator meteorološke stanice, tabela mesta očitavanja kvaliteta vazduha - sadrži identifikatore i lokacije na kojima se nalazi više senzora kvaliteta vazduha, tabela senzora kvaliteta vazduha – sadrži podatke i identifikatore senzora kvaliteta vazduha kao i identifikator lokacije merenja na kojoj se nalazi senzor, tabela vremenskih osnova očitavanja kvaliteta vazduha sadrži podatke o intervalima očitavanja parametara kvaliteta vazduha, uključujući identifikator senzora i identifikator lokacije očitavanja, tabela očitavanja kvaliteta vazduha - sadrži podatke o očitavanjima parametara kvaliteta vazduha, uključujući identifikator vremenske osnove očitavanja, identifikator senzora i identifikator lokacije očitavanja.

Detaljna specifikacija šeme OLTP baze podataka prikazana je u master radu autora.

4. ŠEMA OLAP BAZE PODATAKA

Šema *OLTP* baze podataka predstavlja izvor podataka za izveštajnu funkciju. Njen sadržaj obezbeđuje *ETL* proces. Sastoji se od tabela dimenzija i tabela činjenica. Dimenzije čine vremenska dimenzija, dimenzija lokacije, dimenzija meteoroloških uslova i dimenzija kvaliteta vazduha. Činjenične tabele opisuju pešački saobraćaj, biciklistički saobraćaj, automobilski saobraćaj i upotrebu prevoznih sredstava. Tabela 1 prikazuje matricu Poslovni procesi / Dimenzije (tzv. *Bus Matrix*), u okviru koje su izdvojeni ključni procesi poslovanja. Prikazani procesi poslovanja su predmet projektovanja skladišta podataka. Tabela prikazuje i dimenzije po kojima se posmatraju poslovni procesi.
Detaljna specifikacija šeme *OLAP* baze podataka data je u master radu autora.

Tabela 1. BUS Matrix

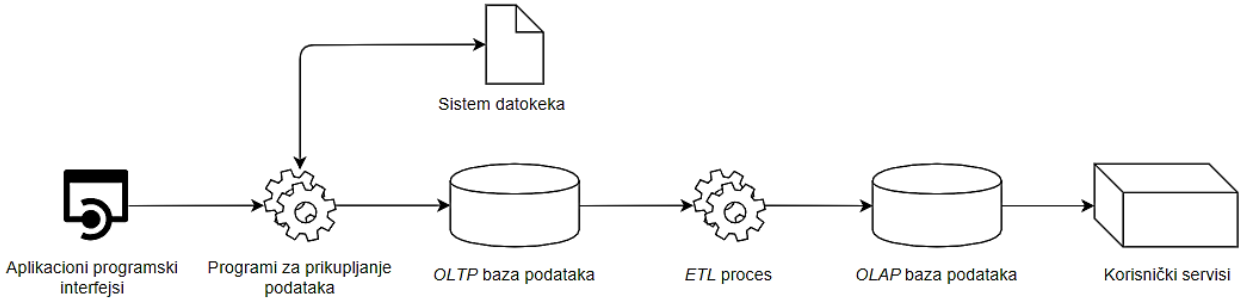
Dimenzija / **Poslovni proces**	**Vreme**	**Meteorološki uslovi**	**Predgrađe**	**Senzor**	**Kvalitet vazduha**
Praćenje upotrebe prevoznih	X	X	X		
Praćenje biciklističkog saobraćaja	X	X	X	X	
Praćenje kretanja pešaka	X	X	X	X	
Praćenje kvaliteta vazduha	X	X	X	X	
Praćenje upotrebe parking mesta	X	X	X	X	X

5. ETL PROCES

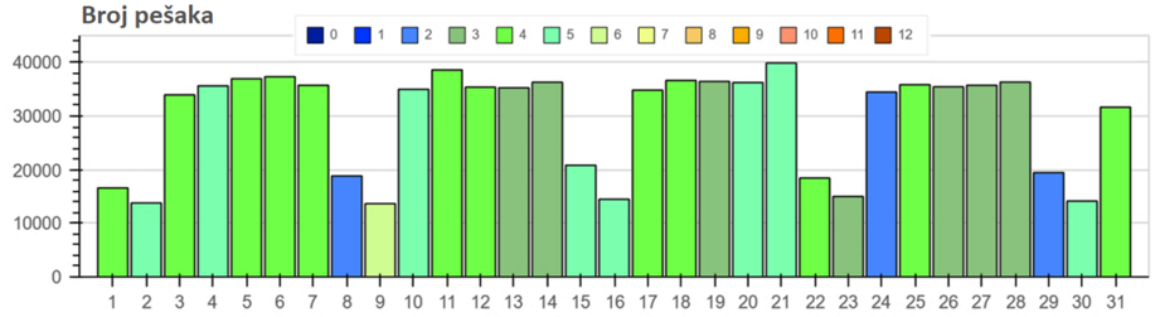
Na slici 2. prikazan je tok podataka kroz sistem, od sistema agencija do korisničkih servisa sistema. Programi za prikupljanje podatka unose podatke u *OLTP* bazu podataka. *ETL* proces obrađuje podatke iz *OLTP* baze podataka i beleži ih u *OLAP* bazu podataka. Korisnički servisi prikazuju rezultate analiza nad podacima u *OLAP* bazi podataka na osnovu zadatatih kriterijuma.
Putem programa za prikupljanje podataka ekstrahuju se podaci sa servisa agencija. Kako bi se smanjila količina podataka koja se preuzima preko Interneta, podaci se snimaju u sistemu datoteka. Programi zatim prenose podatke u *OLTP* bazu podataka. Prenos može da bude ostvaren putem *Python* ili *SQL* skripti.
Upite nad *OLTP* bazom podataka karakteriše upotreba geoprostornih metoda. U takve upite spada poređenje položaja dve tačke na karti, koje mogu predstavljati građevinske objekte, senzore za prikupljanje podataka o saobraćaju ili poslovna preduzeća. Međutim, pokazalo se da su upiti vremenski zahtevni kada podrazumevaju veliki broj provera preseka dva mnogougla, ili provera prisustva tačke unutar mnogougla.
Razvoj *ETL* procesa zahtevao je vremena za istraživanje mogućnosti korišćenja adekvatnog programskog alata. Microsoft integracioni servisi i servisi za analizu ne podržavaju geoprostorni tip podataka. Dalje unapređenje ETL procesa podrazumeva pronalaženje programskog paketa koji podržava geoprostorni tip podataka.
Željeni nivo kvaliteta je postignut usaglašavanjem formata različitih senzora. Projektovanjem ograničenja u *OLTP* bazi podataka sprečen je unos nekonzistentnih podataka u sistem (na primer upisivanje očitavanja nepostojećeg senzora). Radi ostvarenja polaznog cilja rada, analitička funkcija sistema zahteva da se upiti nad *OLAP* bazom podataka izvršavaju brže nego nad *OLTP* bazom podataka. *ETL* proces i *OLAP* baza podataka smanjuju broj spajanja tabela i poređenja kroz agregaciju podataka i kategorizaciju. U zahtevne procedure spadaju prevođenje atributa u geoprostorni tip i poređenje lokacija stanica i granica predgrađa. *ETL* proces omogućava da se ovakve procedure izvše jednom, i to pre potrebe za izvršenjem upita.
Detaljna specifikacija *ETL* procesa prikazana je u master radu autora ovog rada.

6. IZVEŠTAJI

Dinamički izveštaji imaju prednost u odnosu na statičke izveštaje, s obzirom da omogućavaju prilagođavanje upita izveštaja potrebama korisnika. Na ovaj način, korisnici bez informatičkog znanja i iskustva ne moraju da se upoznavaju sa tehnologijom generatora izveštaja, kako bi prilagodili upite svojim potrebama. Sa druge strane, programeri nemaju potrebu za ručnim implementiranjem pojedinih tipova izveštaja.
Slika 3 prikazuje broj zabeleženih pešaka po danu u toku oktobra 2016. godine na senzoru u centru Melburna. Boje stubova grafikona odgovaraju maksimalnoj katego-riji vetra. Legenda povezuje boje sa kategorijama vetra. Sa slike može se utvditi da maksimalne kategorije vetra su se kretale u intervalu od drugog do šestog stepena. 26. januar beleži manje pešaka, pošto je u pitanju državni praznik „Dan Australije“. Većina zaposlenog stanovništva tada koristi slobodan dan, državne ustanove i prodavnice su zatvorene ili imaju skraćeno radno vreme, a pojedini gradski servisi javnog prevoza ne rade. Pored ovog dana u godini, i prvi dan godine (1. januar) beleži manje pešaka. Na osnovu izveštaja o kretanju pešaka, mogu da se planiraju lokacije promocija, anketiranja ili poslovnih preduzeća. Preduzeća mogu da procene vrednost lokacija poslovanja, potreban broj radnika ili potrebu za obezbeđenjem.



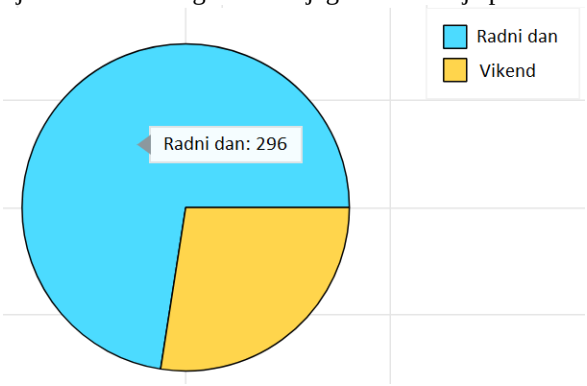
Slika 2. *Tok podataka kroz sistem*



Slika 3. *Izveštaj o broju pešaka po danu u toku oktobra 2016. godine*

Marketing može da se prilagodi izvešta-jima kako bi se privuklo što više kupaca.

Slika 4. prikazuje udeo parkiranja u prekršaju koji su zabeleženi u toku 2016. godine. Grafikon prikazuje da u toku 2016. godine, više od četvrtine prekršaja pri parkiranju zabeleženo je u toku vikenda. Može se zaključiti da je potrebno podstaći plaćanje parkinga vikendom, na primer kroz češće kontrole. Izgradnjom infrastrukture za bicikliste u blizini mesta gde se dešavaju prekršaji može da se podstakne upotreba bicikla u mesto automobila. Ovo je primer izveštaja koji može da podstakne anketiranje korisnika usluge radi boljeg razumevanja problema.



Slika 4. *Izveštaj o broju parkiranja u prekršaju u toku 2016. godine, podeljeno na radne dane i vikend*

S obzirom da implementirani izveštaji o kretanju pešaka, biciklista i upotrebi parking mesta proizilaze iz realnih podataka vladinih agencija Melburna, moguća je njihova upotreba pri planiranju radova na saobraćajnoj infrastrukturi i praćenje njihovog uticaja na saograćaj. Izveštaji omogućavaju procenu isplativosti radova na saobraćajnoj infrastrukturi. Takođe, sistem omogućava praćenje uticaja manifestacija na saobraćaj i organizaciju bezbednosti u gradu.

7. ZAKLJUČAK

U ovom radu predstavljen je sistem skladišta podataka sa izveštajnom funkcijom koji omogućava uvid u kretanje građana grada Melburna. Podaci uključuju očitavanja senzora pešaka, biciklista, parkinga, meteoroloških uslova i kvaliteta vazduha, zatim podatke o granicama predgrađa, kao i podatke o poslovnim preduzećima i stambenim prostorima. Izvori podatka su portali vladinih agencija i API servisi. Podaci koji potiču sa servera zahtevaju programe kako bi se snimili i preneli u *OLTP* bazu podataka. Količina zabeleženih podataka nije prikladna za analizu nad datotekama ili nad *OLTP* sistemom.

Eksplorativna analiza prikupljenih podataka otkrila je nepravilnosti u prikupljenim podacima. Manifestuju se u obliku duplikata ili neispravnih zapisa, što doprinosi još većoj količini zabeleženih podatka sa senzora. Kod pojedinih senzora postoje „rupe" u podacima usled prekida rada senzora. Uočena je i nekonzistentnost u formatiranju vremena i lokacije kod različitih senzora. Poseban problem čine senzori koji beleže očitavanja u lokalnom vremenu, iz razloga što propuštaju sate pri prelasku iz letnjeg u zimsko računanje vremena.

Formirana je *OLTP* šema baze podataka koja je namenjena čuvanju prikupljenih podatka. Uočeni su poslovni procesi, kao i dimeznije po kojima se prate. Formirana je šema *OLAP* baze podataka, koja kasnije ispunjava ulogu izvora podataka za izveštaje.

Prenos podataka iz *OLTP* baze podataka u *OLAP* bazu podataka zahteva implementaciju ETL procesa koji uzima u obzir navedene nedostatke u podacima. Izveštajna funkcija implementirana je na *Web* platformi, što omogućava dinamičko generisanje izveštaja i dostupnost izveštaja velikom broju uređaja. Predlaže se beleženje vremena koje je nezavisno od vremenske zone (griničkom vremenu), ili u *Unix Epoch* vremenu.

Dalji razvoj projekta podrazumeva implementaciju sistema skladišta podataka putem *Azure SQL Data Warehouse* servisa. Punjenje skladišta podataka obezbedio bi *Azure Databricks* servis, na osnovu podataka iz *Azure Data Lake Store* servisa. Prenos sistema na *Azure* servise omogućava bolje performanse i veću dostupnost sistema, kao i jednostavniji razvoj i održavanje. Sistem nema pristup bazama podataka agencija, već ima ograničen pristup Internet servisima agencija. Istorijski podaci senzora pešaka agregirani su na jedan sat, a sat pri prelasku iz letnjeg u zimsko računanje vremena je preskočen. Zbog ovoga, podaci nekih senzora u *OLTP* bazi podataka već su agregirani. U cilju postizanja boljih rezultata, potrebno je razviti proces prenošenja podataka iz baza podataka agencija.

8. LITERATURA

[1] Cocchia A. (2014) „Smart and Digital City: A Systematic Literature Review", Dameri R. Rosenthal-Sabroux C. „Smart City" , pp 13-43 ,2014

[2] *Melbourne Data*, URL https://data.melbourne.vic.gov.au

Kratka biografija:

Simonida Kadić rođena je u Novom Sadu 1994. god. Fakultet tehničkih nauka upisala je 2013. god. Bečelor rad iz oblasti računarske grafike odbranila je 2017. god