Izrada skladišta podataka o zločinima u Chicagu uz primjenu CDC tehnike za inkrementalno punjenje

Mauko, Sara

Sveučilište Jurja Dobrile u Puli

Fakultet informatike u Puli

Sara Mauko

IZRADA SKLADIŠTA PODATAKA O ZLOČINIMA U CHICAGU UZ PRIMJENU CDC TEHNIKE ZA INKREMENTALNO PUNJENJE

Završni rad

Pula, 2024.

Sveučilište Jurja Dobrile u Puli

Fakultet informatike u Puli

Sara Mauko

IZRADA SKLADIŠTA PODATAKA O ZLOČINIMA U CHICAGU UZ PRIMJENU CDC TEHNIKE ZA INKREMENTALNO PUNJENJE

JMBAG: 0303103398, redovni student

Studijski smjer: Informatika

Kolegij: Skladišta i rudarenje podataka

Mentor: izv. prof. dr. sc. Goran Oreški

Pula, 2024.

Sažetak

Predmet ovog rada je izrada skladišta podataka o zločinima u Chicagu uz primjenu Change Data Capture tehnike za inkrementalno punjenje. U radu se koristi timestamp CDC metoda radi jednostavne implementacije, minimalne promjene na izvornoj bazi i efikasnog inkrementalnog punjenja skladišta podataka. U teorijskom dijelu rada opisana je definicija skladišta podataka, razlike između OLTP i OLAP sustava, pristupi skladištenju podataka (Inmonov i Kimballov model), dimenzionalno modeliranje, ETL proces i OLAP analiza. Implementacijski dio rada obuhvaća izradu baze podataka, dizajn star sheme, primjenu Pentaho Data Integration alata za ETL proces, te vizualizaciju podataka u Power BI-u. Kako bi se provjerila ispravnost i učinkovitost inkrementalnog punjenja skladišta podataka, provedeno je jedno inicijalno punjenje te deset simulacija inkrementalnog punjenja podataka.

Poveznica na Github repozitorij:https://github.com/smauko/Crimes-in-Chicago

Ključne riječi: skladište podataka, Change data capture, ETL proces, OLAP analiza, dimenzijsko modeliranje, star shema, kriminal, zločin, OLTP sustavi.

Summary

The subject of this thesis is the development of a data warehouse on crimes in Chicago using the Change Data Capture (CDC) technique for incremental data loading. The timestamp CDC method is used in this study due to its simple implementation, minimal changes to the source database, and efficient incremental data loading into the data warehouse.

The theoretical part of the thesis describes the definition of data warehouses, differences between OLTP and OLAP systems, approaches to data warehousing (Inmon and Kimball models), dimensional modeling, the ETL process, and OLAP analysis. The implementation part includes database design, star schema modeling, the use of Pentaho Data Integration for the ETL process, and data visualization in Power BI.

To verify the accuracy and efficiency of incremental data loading, one initial load and ten simulations of incremental loading were conducted.

Github repository link : https://github.com/smauko/Crimes-in-Chicago

Keywords: data warehouse, Change Data Capture, ETL process, OLAP analysis, dimensional modeling, star schema, crime, OLTP systems.

Sadržaj

1. Uvod

U današnje vrijeme, organizacije se oslanjaju na podatke kako bi donosile informirane odluke. Posebno u području kriminalistike i sigurnosti, analiza podataka o kriminalu može pomoći u prepoznavanju obrazaca, predviđanju budućih incidenata i optimizaciji strategija prevencije. No, s obzirom na velike količine podataka koji se generiraju svakodnevno, potrebno je osigurati efikasno pohranjivanje, pristup i analizu tih podataka.

Skladišta podataka predstavljaju rješenje za sustavno organiziranje i analizu velikih skupova podataka. Njihova struktura omogućava učinkovito izvođenje analitičkih upita, što je ključno za donošenje strateških odluka. U ovom radu, fokus će biti na izradi skladišta podataka o zločinima u Chicagu, koristeći dimenzionalno modeliranje kako bi se osigurala brža i jednostavnija analiza podataka.

Jedan od izazova u radu sa skladištima podataka je osiguravanje ažurnosti informacija. Budući da se podaci o zločinima neprestano ažuriraju, ključno je implementirati strategiju inkrementalnog punjenja podataka kako bi se smanjila redundancija i poboljšala učinkovitost. U ovom radu koristit će se Change Data Capture tehnika za inkrementalno punjenje podataka, omogućujući učinkovito praćenje promjena i njihovo ažuriranje u skladištu podataka.

Glavni cilj ovog rada je dizajnirati i implementirati skladište podataka koje omogućava OLAP analizu podataka o zločinima u Chicagu, koristeći prikladne tehnike modeliranja i optimizacije podataka. Kroz rad će se detaljno objasniti teorijske osnove skladišta podataka, proces izgradnje i modeliranja baze podatka i skladišta podataka, primjena CDC tehnike i analiza podataka pomoću OLAP pristupa.

2. Teorijska osnova

2.1. Uvod u skladišta podataka

Transakcijski sustavi, poznati i kao OLTP sustavi (eng. *Online Transaction Processing*), ključni su za svakodnevno poslovanje organizacija. Ovi sustavi optimizirani su za obradu velikog broja transakcija u stvarnom vremenu, omogućujući brzo i točno upravljanje podacima poput narudžbi, plaćanja, rezervacija i evidencija. OLTP sustavi oslanjaju se na visoko normalizirane baze podataka koje osiguravaju konzistentnost podataka i podržavaju sve operacije unosa, ažuriranja, brisanja i dohvaćanja podataka. Međutim, kako organizacije rastu, a količina podataka eksponencijalno se povećava, javlja se potreba za dugoročnim očuvanjem i analizom povijesnih podataka.

Transakcijski sustavi, iako izvrsni za svakodnevne operacije, nisu dizajnirani za analitičku obradu podataka. Povijesni podaci često se uklanjaju iz OLTP sustava kako bi se održale performanse i smanjila složenost, no njihova vrijednost za organizaciju ostaje značajna. Očuvanje tih podataka postalo je ključno za podršku dugoročnim strategijama i donošenju informiranih odluka, što je dovelo do razvoja skladišta podataka. [3] [4]

2.2. Definicija skladišta podataka

Skladišta podataka su skupovi trenutnih i povijesnih transakcijskih podataka specijalno strukturirani za analizu poslovanja, generiranje izvješća i donošenje poslovnih odluka.

William H. Inmon, poznat kao otac skladišta podataka, definira skladište podataka kao subjektno orijentiran, integriran, vremenski obilježen, te postojan skup podataka koji služi za donošenje odluka o poslovanju [2].

Ova definicija naglašava ključne karakteristike skladišta podataka:

- **Subjektno orijentiran –** Za razliku od OLTP sustava, gdje se podaci strukturiraju prema aplikacijama ili procesima, podaci u skladištima podataka se organiziraju prema određenim poslovnim područjima koji pomažu u procesu odlučivanja. Ako uzmemo za primjer ovoj završni rad možemo utvrditi da su podaci podijeljeni prema ključnim tematskim područjima kao što su vrsta zločina, lokacija, vrijeme, počinjeni zločini i počinitelj. Ovakav način raspodijele podataka nam omogućuje analize poput učestalost zločina po strani grada, uhićenja po vrsti zločina, broj zločina prema spolu i prisutnosti obiteljskog nasilja.

- **Integriran –** Da bi se povećala efikasnost skladišta podataka i prikupili svi potrebni podaci za uspješniju analizu, podaci se prikupljaju iz više izvora. Tim procesom osiguravamo detaljniji pregled na podatke. Podaci da na kraju posluže kao jedinstven i pouzdan izvor informacija prolaze kroz postupak standardizacije kako bi se osigurala dosljednost, kao i proces čišćenja kojim se uklanjaju duplicirani i nepotrebni podaci. Tipičan primjer standardizacije je usklađivanje formata datuma i vremena, brojeva telefona, mjera, spola….

- **Vremenski obilježen –** Nakon dodavanja nekog novog aktualnog podatka u skladište podataka, dodaje mu se vremenska oznaka koja pokazuje od kada do kada je taj podatak aktivan. Na ovaj način skladišta podataka ne pohranjuju samo trenutne podatke, već i sve izmijenjene, obrisane i povijesne informacije. Takav pristup omogućuje praćenje promjena kroz vrijeme i analizu podataka u njihovom vremenskom okviru, što pridonosi u donošenju boljih odluka.

- **Postojan –** Podaci koji su se jednom unijeli u skladište podataka više se ne mijenjaju niti brišu, ako se neki podatak promijeni u izvornom sustavu, skladište ne zamjenjuje postojeći zapis, već pohranjuje novu verziju podatka uz pripadajuću vremensku oznaku. Ovakav način pohranjivanja nam dozvoljava čuvanje povijesnih podataka kojih u izvorima više nema.

Prema drugoj definiciji Kimball i Caserta [5] skladište podataka definiraju kao sustav koji iz izvora dohvaća, čisti, usklađuje i pohranjuje podatke u multidimenzionalno spremište podataka, a zatim pruža podršku i implementira mogućnost postavljanja upita i analize u svrhu donošenja poslovnih odluka. Objašnjavaju kako je skladištenje podataka proces uzimanja podataka iz transakcijskih baza i ostalih izvora te njihovo pretvaranje u organizirane informacije u korisnički prilagođenom formatu.

2.3. Razlika između OLTP i skladišta podataka

Parteek Bhatia [1] objašnjava razlike između dvaju vrlo važnih alata za upravljanje podacima unutar organizacije, no koji imaju značajno različite ciljeve. OLTP su tradicionalni sustavi baza podataka dizajnirani za maksimizaciju kapaciteta obrade transakcija organizacijskih informacija, dok skladišta podataka sadrže detaljne i povijesne podatke koji se mogu sažeti na različite načine i rijetko se mijenjanju.

Kroz tablicu predstavljam ključne razlike OLTP i skladišta podataka. [3][4]

OLTP sustav	Skladište podataka
Sadrži podatke u stvarnom vremenu	Sadrži povijesne podatke
Koristi normalizirane podatke radi smanjenja redundantnosti	Koristi de-normalizirane podatke za multidimenzionalne analize
Podaci su promjenjivi	Podaci su postojani
Orijentiran prema svakodnevnim transakcijama	Orijentiran prema analizi podataka
Obrađuju jednostavne upite	Koriste se složenim upitima za analizu podataka
Brza obrada podataka (upiti, ažuriranja, dodavanja, brisanja) oko nekoliko milisekundi	Brzina obrade ovisi o količini podataka, ali varira od nekoliko sekundi pa do nekoliko sati
Izvor podataka su svakodnevne transakcije	Različiti izvori podataka
Malen prostor pohranjivanja (100 MB-10 GB)	Velik prostor pohranjivanja (1 TB – 100 PB)
Koriste ga krajnji korisnici poslovnih aplikacija i operativno osoblje	Koriste ga menadžeri i analitičari
Visoka frekvencija upita	Niska frekvencija upita
Česta i kontinuirana ažuriranja podataka u stvarnom vremenu.	Ažuriranja podataka se provode periodično (npr. dnevno, tjedno ili mjesečno)

Tablica 1:Razlika između OLTP i skladišta podataka

2.4. Ciljevi skladišta podataka

Skladišta podataka predstavljaju ključnu tehnologiju za organizacije koje žele iskoristiti svoje podatke za donošenje strateških odluka. Osim pohrane povijesnih podataka, njihova svrha je osigurati jednostavan pristup informacijama, konzistentnost podataka i podršku za naprednu analizu. Kimball i Ross [6] su po iskustvu i frustracijama menadžera odlično opisali ciljeve skladišta podataka:

- **Skladište podataka mora omogućiti jednostavan pristup informacijama organizacije –** Sadržaj skladišta podataka mora bit razumljiv, podaci moraju biti intuitivni, očiti i čitljivi za poslovne korisnike. Želja menadžera i analitičara je da mogu razdvajati i kombinirati podatke u skladištu na beskonačno mnogo načina. Alati za pristup moraju biti laki i jednostavni za korištenje, a rezultati upita moraju se brzo vratiti korisniku.

- **Skladište podataka mora dosljedno predstavljati informacije organizacije –** Da bi podaci bili vjerodostojni, potrebno ih je pažljivo prikupiti iz različitih izvora, očistiti, provjeriti njihovu kvalitetu i objaviti tek kada su spremni za upotrebu. Informacije iz različitih poslovnih procesa trebaju biti usklađene, ne smije se dogoditi da dva različita podatka imaju isti naziv. Na kraju, svi podaci trebaju biti obuhvaćeni i potpuni kako bi informacije bile kvalitetne i korisne za donošenje odluka.

- **Skladište podataka mora biti prilagodljivo i otporno na promjene.** Potrebe korisnika, poslovni uvjeti, podaci i tehnologija podložni su promjenama, što znači da su promjene u skladištima podataka neizbježne. Skladište podataka treba biti dizajnirano da podnese te promjene. Promjene moraju biti minimalne, što znači da ne smiju mijenjati i poremetiti postojeće podatke.

- **Skladište podataka mora biti čvrst i siguran sustav za zaštitu informacija.** Najvrjedniji podaci organizacije nalaze se u skladištu podataka, uključujući ključne informacije poput prodaje, podataka o klijentima i cijenama. Ako takvi podaci dospiju u pogrešne ruke, mogu izazvati značajnu štetu. Zbog toga skladište podataka mora osigurati strogu i učinkovitu kontrolu pristupa povjerljivim informacijama organizacije.

- **Skladište podataka mora služiti kao temelj za bolje donošenje odluka.** U skladištu podataka moraju biti pohranjeni točni i relevantni podaci koji podržavaju proces odlučivanja. Glavni rezultat skladišta podataka nisu samo analize, već odluke donesene na temelju prikupljenih i obrađenih informacija. Takve odluke izravno pridonose stvaranju poslovne vrijednosti i uspjehu organizacije.

- **Poslovna zajednica mora prihvatiti skladište podataka da bi ono bilo uspješno.** Bez obzira na to koliko je skladište podataka tehnički napredno, ako korisnici ne vide njegovu vrijednost i ne koriste ga, sve to postaje nevažno. Ključ uspjeha nije samo početno korištenje sustava, već i njegova dugoročna upotreba. Ako korisnici prestanu koristiti skladište podataka unutar šest mjeseci nakon implementacije i obuke, sustav nije prošao "test prihvaćanja" jer očito ne zadovoljava njihove potrebe ili je prekompliciran za upotrebu. Najvažniji faktor za prihvaćanje skladišta podataka od strane korisnika je jednostavnost. Ako je sustav previše kompliciran, čak i najnapredniji korisnici će ga izbjegavati, što znači da skladište neće ispuniti svoju svrhu.

Kimball i Ross [6] da zaključe ovu listu objašnjavaju kako uspješno upravljanje skladištima podataka zahtjeva puno više od vrhunskog administratora baze podataka. Uspješno upravljanje skladištima podataka podrazumijeva spojiti dva svijeta: tehnološki i poslovni. S jedne strane, postoji poznati teren informacijskih tehnologija, gdje se radi s podacima, sustavima i tehničkim procesima. S druge strane, nalazi se nepoznat svijet poslovnih korisnika, njihovih zahtjeva, izazova i očekivanja. Moramo pronaći ravnotežu između ta dva svijeta, prilagođavajući i usavršavajući svoje postojeće vještine kako bismo ispunili specifične zahtjeve skladištenja podataka.

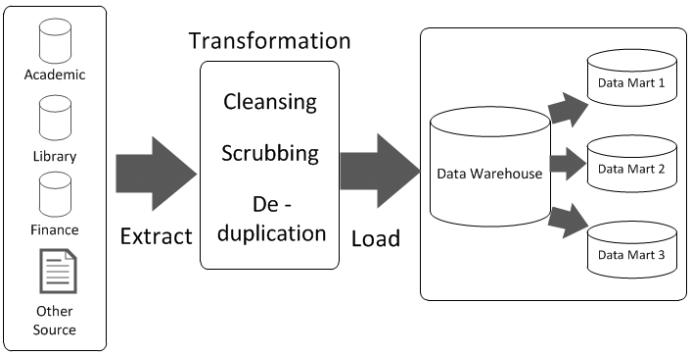
2.5. 2 pristupa skladištenju podataka

U dizajnu skladišta podataka ključnu ulogu igraju različiti pristupi implementacije, posebno top-down (od vrha prema dolje) i bottom-up (od dna prema vrhu) strukture. Ovi pristupi određuju kako se skladište podataka planira, razvija i izrađuje te imaju značajan utjecaj na cijeli njegov životni ciklus.

2.5.1. Top Down (Inmon model)

Bill Inmon razvio je koncept skladišta podataka koji se temelji na izgradnji centraliziranog skladišta podataka (EDW, eng. *Enterprise Data Warehouse*), iz kojeg se kasnije izdvajaju područna skladišta podataka prilagođena pojedinim poslovnim funkcijama. Područna skladišta podataka (eng. *Data Mart*) predstavljaju podskup EDW-a koji se fokusira na određeno poslovno područje, poput financija, marketinga ili prodaje.

Struktura skladišta se temelji na normaliziranom modelu, što smanjuje redundantnost podataka i olakšava održavanje, ali istovremeno može otežati složene analitičke upite zbog većeg broja povezanih tablica. [7] [8]



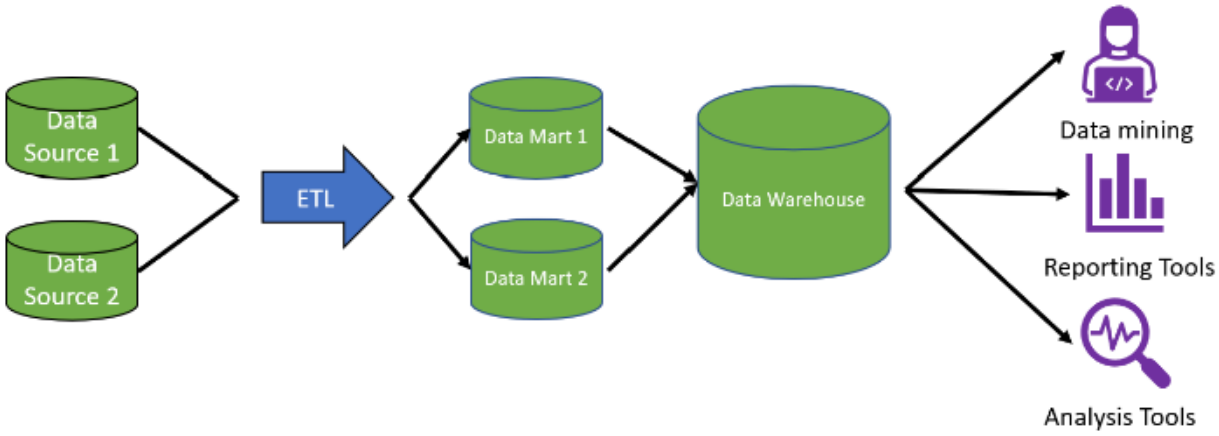
Slika 1: Prikaz izrade skladišta podataka prema Inmonovom pristupu (Izvor: https://www.researchgate.net/figure/Bill-Inmons-Top-Down-approach-to-DWH-design_fig1_328434296)

2.5.2. Bottom Up (Kimball model)

Kimball-ov pristup skladištenju podataka temelji se na dimenzionalnom modeliranju, gdje podaci nisu normalizirani radi brže analize i jednostavnijeg dohvaćanja informacija. Započinje kreiranjem područnih skladišta podataka usmjerenih na specifične poslovne potrebe i na kraju se povezuju kako bi formirali veliko skladište podataka ukupne organizacije.

S obzirom na to da je cijela arhitektura skladišta podataka u ovom pristupu izgrađena na dimenzijskom modeliranju, a u radu koristimo bottom-up pristup, detaljan opis načina modeliranja skladišta podataka nalazi se u poglavlju 2.6. [Dimenzijsko modeliranje skladišta podataka]. U tom poglavlju objašnjene su sve ključne komponente modeliranja

Kimball-ovog modela, kao i način na koji se podaci organiziraju i koriste za analitičke svrhe. [7][8]



Slika 2: Prikaz izrade skladišta podataka prema Kimball-ovom modelu (Izvor:https://campus.datacamp.com/courses/data-warehousing-concepts/data-warehouse-data-modeling?ex=1)

2.5.3. Razlike modela

Inmon model	Kimball model
Izrada skladišta podataka je vremenski zahtjevnija	Izrada skladišta podataka je vremenski kraća
Održavanje skladišta podataka je lakše	Održavanje skladišta podataka je teže
Izrada je kompleksnija	Izrada je efikasnija
Normalizirana struktura	De-normalizirana struktura
Početni troškovi implementacije su veći	Početni troškovi implementacije su niži
Prvo se izrađuje EDW, i tek onda iz njega data mart-ovi	Odmah se izrađuju data mart-ovi

Top-down metoda	Bottom-up metoda
Lakše se prilagođava promjenama	Teže se prilagođava promjenama

Tablica 2: Razlike između Inmonovog i Kimballovog modela

Analizirajući tablicu 2 možemo zaključiti da Inmon-ov i Kimball-ov pristup skladištenju podataka imaju različite prednosti i nedostatke, a odabir između njih ovisi o potrebama, resursima i dugoročnim ciljevima organizacije. Inmon-ov model bolje odgovara velikim sustavima koji zahtijevaju EDW, visoku kvalitetu i integraciju podataka, ali njegova implementacija je složenija i dugotrajnija. S druge strane, Kimball-ov model omogućuje bržu i jednostavniju izgradnju skladišta podataka kroz data martove prilagođene poslovnim funkcijama, ali može dovesti do nekonzistentnosti podataka između odjela.

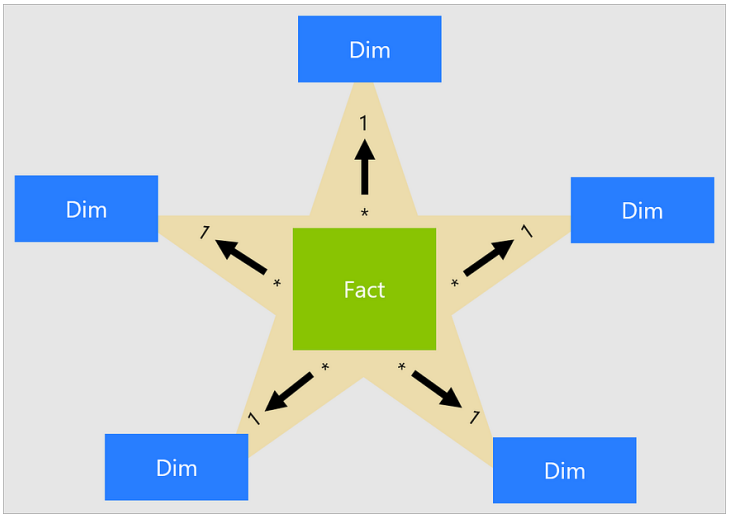
2.6. Dimenzijsko modeliranje skladišta podataka

Dimenzijsko modeliranje je tehnika modeliranja podataka u kojoj su podaci jednostavni, čitljivi i organizirani na način koji omogućava učinkovitu analizu i brzo dohvaćanje informacija. Temelji se na podacima koji nisu normalizirani, tj. sadrži redundantne podatke.

Svaki dimenzijski model sastoji se od jedne glavne tablice, poznate kao tablica činjenica(eng. *Fact table*), koja sadrži kvantitativne podatke i povezuje se s dimenzijskim tablicama putem tehničkih ključeva istih. Uz nju se nalaze dimenzijske tablice (eng. *Dimension tables*), koje sadrže kvalitativan tip podataka i koje pružaju kontekst za analizu podataka. Dimenzijske tablice su obično manje u broju zapisa, ali bogate opisnim

atributima koji omogućuju dublju analizu i grupiranje podataka prema različitim poslovnim kategorijama. Ova organizacija podataka oblikuje zvjezdanu shemu (eng. *Star schema*). [1]

Star shema je najpopularniji i najčešći stil dimenzijskog modeliranja. U star shemi tablica činjenica nalazi se u sredini modela, a okružuju je dimenzijske tablice. Star shema je popularna zbog svoje jednostavnosti i učinkovitosti u analitičkim sustavima. Njena de-normalizirana struktura omogućuje brzo izvršavanje upita, smanjuje potrebu za složenim pridruživanjem tablica i olakšava razumijevanje podataka poslovnim korisnicima. Također, omogućuje jednostavno proširivanje modela dodavanjem novih dimenzija ili mjerenja. [9]



Slika 3: Struktura zvjezdane/star sheme (Izvor: https://medium.com/@marcosanchezayala/data-modeling-the-star-schema-c37e7652e206)

2.6.1.Tablica činjenica (Fact Tables)

Tablica činjenica je glavna i središnja tablica u dimenzijskom modelu u koju se pohranjuju numeričke vrijednosti koje predstavljaju mjerenja ili pokazatelje poslovnih procesa. Uz mjerenja, u tablici činjenica nalaze se i tehnički ključevi dimenzija koji još detaljnije opisuju poslovni proces. [10]

Pojam tablica činjenica koristimo za predstavljanje poslovnih mjera. Poslovna mjera odgovara na pitanje „Što se dogodilo“, tj. bilježi konkretne podatke o događajima u poslovnom procesu, kao što su prodaja proizvoda, ostvareni prihod, broj transakcija ili količina isporučene robe. Na primjer, ako analiziramo temu ovog završnog rada, tablica činjenica predstavlja mjeru o počinjenim zločinima, a njeni podaci bi bili broj svjedoka, informacije da li je osoba bila uhićena (da/ne), informacije da li je zločin povezan sa obiteljskim nasiljem (da/ne)…

Za razliku od dimenzijskih tablica, koje su de-normalizirane i redundantne, tablica činjenica je često normalizirana i često identična odgovarajućoj tablici u transakcijskom sustavu. [10]

2.6.2. Dimenzijske tablice (Dimension Tables)

Dimenzijske tablice okružuju tablicu činjenica i dopunjuju tablicu činjenica tekstualnim opisima poslovanja. Da bi rezultat bio dobro dizajniran dimenzionalni model, dimenzijske tablice bi trebale imati puno stupaca ili atributa koji opisuju retke u tablici dimenzija. Obično dimenzijske tablice sadrže manji broj redaka u usporedbi sa tablicom činjenica.

Svaka dimenzija definirana je svojim jedinstvenim tehničkim ključem (TK, eng.), koji služi za povezivanje sa tablicom činjenica. Veza između tablice činjenica i dimenzijske

tablice je „one-to-many“ ili jedan na više, što znači da jedan zapis u dimenzijskoj tablici može biti povezan na više zapisa u tablici činjenica.

Atributi dimenzija ključni su za filtriranje, grupiranje i označavanje podataka u upitima. U upitima ili zahtjevima za izvještaje atributi se identificiraju riječima „po“, na primjer kada korisnik želi vidjeti zločine po kvartu i po mjesecu, kvart i mjesec moraju biti dostupni kao atributi dimenzija. Atributi određuju upotrebljivost i razumljivost skladišta podataka, a njegova kvaliteta ovisi o dubini i preciznosti tih atributa. Što su atributi detaljniji i bolje definirani, to je skladište podataka korisnije, također najvrjedniji atributi su oni tekstualni i diskretni. [10]

Za razliku od činjenične tablice, koja obično sadrži veliki broj redaka i referencira dimenzijske tablice putem vanjskih ključeva, dimenzijske tablice su de-normalizirane i sadrže redundantne podatke kako bi omogućile jednostavniju analizu. [10]

2.6.3. Degenerirane dimenzije

Degenerirane dimenzije (eng. *Degenerate Dimension*) su posebne vrste dimenzija u skladištima podataka koje se ne mogu smjestiti u posebnu dimenzijsku tablicu, već se njeni atributi nalaze unutar tablice činjenica. Najčešće se radi o atributima koji nemaju svoju prirodnu dimenzionalnu hijerarhiju, ali su korisni za analitičke upite.

2.7. ETL proces

Kimball i Caserta [5] tvrde da je ETL (*Extract, Transform, Load*) sustav presudan za uspjeh skladišta podataka. Iako je izgradnja ETL sustava proces koji se odvija u pozadini,

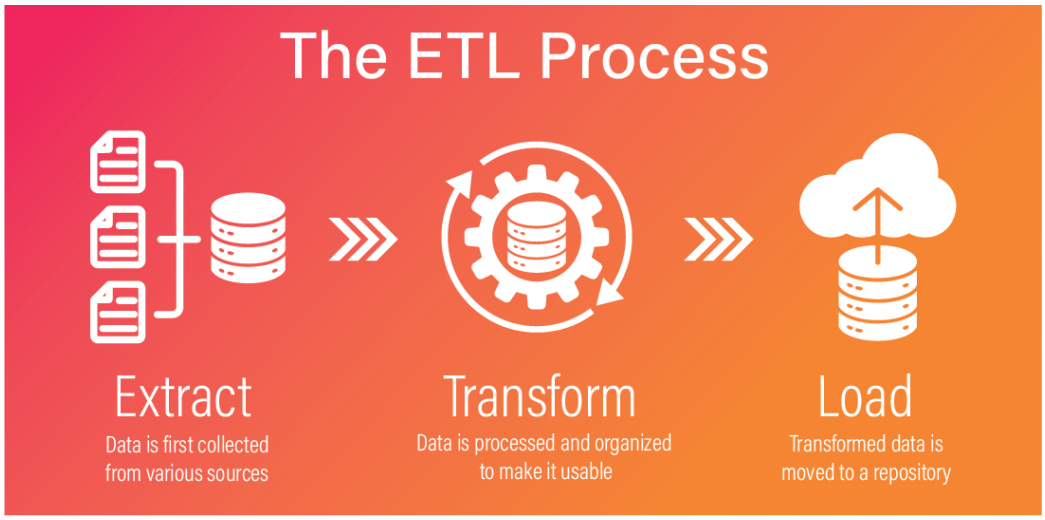
nevidljiv krajnjim korisnicima, on troši 70% resursa potrebnih za implementaciju i održavanje skladišta podataka.

ETL je proces preuzimanja podataka iz različitih izvora, transformacije i čišćenje tih podataka, te učitavanje obrađenih podataka u skladište podataka.

Kimball i Caserta [5] još objašnjavaju da ETL susutav ne služi samo za premještanje podataka iz izvora u skladište podataka, nego da je njegova vrijednost i njegov doprinos puno veći. Navode da ETL sustav:

- Uklanja pogreške i ispravlja nedostajuće podatke,
- Omogućuje dokumentirane mjere pouzdanosti podataka,
- Bilježi tijek transakcijskih podataka radi sigurnog čuvanja,
- Prilagođava podatke iz više izvora kako bi se mogli koristiti zajedno,
- Strukturira podatke tako da budu upotrebljivi za krajnje korisnike i analitičke alate.

Faze ETL procesa su **izdvajanje podataka**(eng. *Extract*)**, transformacija** (eng. *Transform*) **i punjenje** (eng. *Load*)**.**



Slika 4: ETL proces (Izvor: https://www.zuar.com/blog/what-is-etl-pipeline/)

2.7.1. Izdvajanje podataka

Prvi korak ETL procesa je izdvajanje podataka, on je zaslužan za izdvajanje podataka iz transakcijskih ili izvorišnih sustava organizacije. Kimball i Caserta [5] objašnjavaju da izdvajanje podataka mora biti pažljivo planirano kako bi se osiguralo da se podaci u izvorišnom sustavu ne mijenjaju niti brišu te da proces izdvajanja ima minimalan utjecaj na stabilnost sustava. Kako bi se smanjilo oštećenje i opterećenje sustava, podaci se privremeno pohranjuju u pripremno područje(eng. *Staging area*) i tamo ostaju sve do učitavanja podataka u skladište. To znači da ovo područje služi kao neka vrsta međuspremnika u kojoj podaci ostaju sve do trenutka kada su spremni za učitavanje u skladište, prolaze kroz proces čišćenja, validacije i prilagodbe kako bi zadovoljili poslovna pravila i strukturu skladišta podataka.

Izdvajanje podataka se dijeli na potpuno (eng. *Full load*) i djelomično (eng. *Incremental load*). Potpuno izdvajanje podataka podrazumijeva izdvajanje svih podataka iz odabranih izvora i koristi se obično prilikom inicijalnog punjenja skladišta podataka. Sa druge strane djelomično izdvajanje podataka je preuzimanje onih podataka koji su novi ili koji su se promijenili od posljednjeg ETL procesa.

Kako bi se osiguralo točno prepoznavanje promjena u podacima, koristi se tehnika Change Data Capture (CDC), koja omogućuje praćenje i obradu samo onih podataka koji su se promijenili u izvorišnom sustavu koje koristimo. Budući da je CDC ključan dio ovog rada, detaljnije ću ga obraditi u sljedećem poglavlju 2.8. [Change Data Capture (CDC)] , gdje ću objasniti kako funkcionira, koje metode postoje i na koji način se implementiraju.

2.7.2.Transformacija podataka

Drugi korak ETL procesa je transformacija podataka gdje se podaci, koji su prikupljeni u pripremnom području, prilagođavaju kako bi bili točni, čitljivi i usklađeni. Kao što sam spomenula u poglavlju 2.7.2. [Izdvajanje podataka], svaka transformacija, čišćenje i standardizacija podataka izvodi se u pripremnom području. Prema Kimballu i Caserti [5], transformacija je često najzahtjevniji dio ETL procesa, jer se u ovoj fazi osigurava da podaci iz različitih izvora budu međusobno usklađeni i prilagođeni pravilima skladišta podataka.

Inmon [2] tvrdi da ako se podaci ne transformiraju i ne modificiraju na pravi način, skladište podataka može sadržavati nekonzistentne i netočne informacije, što može negativno utjecati na analitičke procese i donošenje poslovnih odluka. Uzimajući to u obzir neki od načina modificiranja i transformiranja podataka su :

- Čišćenje podataka – uklanjanje duplikata, ispravljanje grešaka, popunjavanje *null* vrijednosti
- Standardizacija podataka – pretvaranje podataka u jedinstveni format
- De-normalizacija podataka – kombiniranje podataka i tablica
- Generiranje novih podataka na temelju postojećih – novi podaci se stvaraju na temelju postojećih
- Filtriranje podataka – uklanjanje nepotrebnih ili nerelevantnih podataka

2.7.3. Punjenje skladišta podataka

Treći korak ETL procesa je punjenje podataka u skladište podatka. Sređeni i validirani podaci tada napokon napuštaju pripremno područje. Ovisno o specifičnom procesu, punjenje može biti odjednom (eng. *Bulk load*), što je brže i koristi se za velike količine

podataka, ili zapis po zapis (eng. *Row-by-row load*), što omogućuje preciznije upravljanje podacima, ali je sporije.

Tijekom procesa punjenja podataka u skladište podataka provode se tri ključne operacije. Prva je dodavanje novih zapisa (eng. *Insert*), što se odnosi na unos podataka koji prethodno nisu postojali u skladištu, primjerice novih transakcija ili novih korisnika. Druga operacija je ažuriranje postojećih podataka (eng. *Update*), gdje se zapisi koji su već pohranjeni osvježavaju novim informacijama, poput promjene adrese korisnika ili ažuriranja statusa narudžbe. Treća operacija je brisanje zastarjelih ili nepotrebnih podataka (eng. *Delete*), čime se osigurava da skladište podataka ostane optimizirano i ne sadrži irelevantne informacije, poput neaktivnih korisnika ili zastarjelih poslovnih zapisa.[5]

2.8. Change Data Capture (CDC)

CDC je tehnika za identifikaciju, praćenje i prijenos promjena u podacima baze podataka u druge sustave, poput skladišta podataka. Umjesto ponovnog učitavanja cijelog skupa podataka, CDC osigurava da se u odredišni sustav prenose samo novi, izmijenjeni ili obrisani zapisi.

U tradicionalnim pristupima ažuriranja podataka, cijeli skup podataka često se učitavao iznova, što je uzrokovalo značajno opterećenje na izvorne sustave i smanjivalo učinkovitost obrade. CDC rješava taj problem identificiranjem samo onih podataka koji su se promijenili od posljednje obrade te ih selektivno prenosi u odredišni sustav.

CDC je komponenta ETL procesa koja se kontinuirano izvodi primjenjuje u fazi izdvajanja podataka, gdje identificira i dohvaća samo one podatke koji su se promijenili od posljednjeg učitavanja. Nakon učitavanja promijenjenih podataka, podaci prolaze kroz fazu transformacije podataka, gdje se čiste, filtriraju i prilagođavaju pravilima skladišta

podataka. U završnoj fazi, CDC omogućuje inkrementalno ažuriranje podataka, koristeći tehnike poput *Insert, Update, Delete* operacija, čime se eliminira potreba za ponovnim učitavanjem cijele baze.

Postoji nekoliko metoda i tehnika CDC implementacije, ali najčešće korištene su:

- **Dodavanje vremenske oznake** (eng. *Timestamp-Based CDC*) – ova metoda koristi dodatni stupac u tablicama transakcijskog sustava, najčešće nazvan „last_modified“, koji bilježi vrijeme posljednje izmjene retka. Prilikom svakog unosa ili ažuriranja podataka, stupac se automatski popunjava trenutnom vremenskom oznakom. Kod svakog učitavanja u skladište podataka, ETL proces dohvaća samo one zapise čija je vremenska oznaka veća od one zabilježene pri posljednjem učitavanju. Iako je implementacija jako jednostavna, glavni nedostatak ove tehnike je to što ne može detektirati brisanja podataka. Ako se redak izbriše iz transakcijske baze, ne postoji vremenska oznaka koja bi označila tu promjenu. U ovome završnom radu koristi se ova metoda CDC implementacije, budući da je relativno jednostavna i učinkovita u kontekstu podataka o zločinima u Chicagu.
- **Korištenje dnevnika transakcija** (eng. *Log-Based CDC*) – koristi transakcijski zapisnik baze podataka kako bi detektirala promjene nad podacima. Svaka operacija dodavanja, ažuriranja ili brisanja u izvornoj bazi automatski se bilježi u dnevniku transakcija. Log-Based CDC analizira taj dnevnik i identificira samo one zapise koji su se promijenili i generira nove verzije tih podataka u skladištu podataka. Kompleksniji je za implementaciju, no vrlo učinkovit i minimalno opterećuje izvorni sustav.
- **Korištenje okidača** (eng. *Trigger-Based CDC*) – koristi SQL okidače za bilježenje promjena u bazi podataka. Okidači su postavljeni da se aktiviraju prilikom izvršavanja operacija INSERT, UPDATE i DELETE, pri čemu automatski upisuju odgovarajuće zapise u CDC tablicu, a CDC ih onda analizira i kasnije u ETL procesu obrađuju promjene. Ova metoda osigurava visoku točnost podataka, ali

implementacija zahtijeva dodavanje i održavanje SQL okidača, što može povećati kompleksnost kod velikih baza podataka

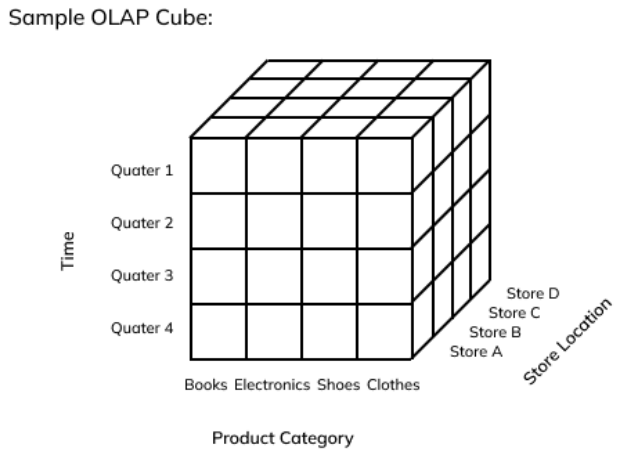
- **Razlika snimki stanja** (eng. *Snapshot-Based CDC*) – koristi metodu uspoređivanja trenutne snimke podataka s prethodnom koja je sačuvana u pripremnom području, analizirajući zapise jedan po jedan kako bi se otkrile nove, ažurirane ili obrisane vrijednosti. Ova metoda je jednostavna za implementaciju i ne zahtijeva promjene u strukturi baze, ali troši značajne resurse i vrijeme.

2.9. OLAP

OLAP (Online analitička obrada podataka, eng. *Online Analytical Processing*) je jedna od ključnih tehnologija poslovne inteligencije (BI, eng. *Business Intelligence*) i ona omogućuje brzo pretraživanje, analizu i vizualizaciju podataka iz skladišta podataka. [6]

OLAP sustavi bi trebali osigurati brzu obradu upita, multidimenzionalnu analizu podataka, fleksibilne analitičke operacije i podršku za donošenje odluka.

Kako bi se ostvarila detaljna analiza podataka, OLAP koristi različite analitičke operacije koje omogućuju korisnicima da podatke istraže iz različitih perspektiva kroz multidimenzionalni model podataka, poznat kao OLAP kocka. OLAP kocka omogućuje organizaciju podataka u više dimenzija, pomoću čega korisnici mogu izdvajati određene dijelove podataka, mijenjati razinu detalja prikazanih informacija itd.



Slika 5: OLAP kocka (Izvor:https://www.sprinkledata.com/blogs/what-is-olap-cube)

Operacije koje se izvode nad OLAP kockom za OLAP analizu su:

- **„Slice and Dice“** – OLAP kocka se „reže“ što znači da „Slice“ izdvaja podatke prema određenoj dimenziji, dok „Dice“ filtrira podatke prema više dimenzija istovremeno i OLAP kocka se sužava na određene kriterije
- **„Drill-down“ –** OLAP kocka se „produljuje“ što znači da podaci prelaze sa grupiranih na detaljnije prikaze i spuštaju na niže razine hijerarhije
- **„Roll-up“ –** OLAP kocka se „skuplja“ podaci se grupiraju i prikazuju na višoj razini hijerarhije
- **„Pivot“ –** OLAP kocka se rotira kako bi se promijenila perspektiva prikaza podataka
- **„Drill-across“ -** OLAP kocka povezuje podatke iz više tablica činjenica

3. Izrada skladišta podataka - Case study: Crimes in Chicago

3.1. Skup podataka

Za ovaj završni rad koristila sam skup podataka „Crimes In Chicago (2001 to 2023)“. Skup podataka je preuzet sa online platforme Kaggle koja omogućuje pristup širokom rasponu javno dostupnih skupova podataka. [14]

Izabrani skup podataka sadrži zapise o svim počinjenim zločinima u gradu Chicagu, a podaci su prikupljeni od strane policijske uprave u Chicagu. Kako bi se zaštitila privatnost žrtava zločina, adrese počinjenih zločina su upisali na razini kvartova ili ulica bez brojeva. [14]

Originalan skup podataka sadržavao je 22 stupca, odnosno atributa, koji uključuju osnovne informacije poput identifikacijskog broja zločina, datuma počinjenog zločina, adrese gdje je počinjen zločin, opis zločina, da li je osoba uhićena te dodatne atribute vezane uz okolnosti počinjenog zločina. Skup je sadržavao i 7.784.665 redaka, pri čemu svaki redak predstavlja jedan počinjeni zločin. Za potrebe ovog rada iskoristila sam 8 ključnih stupaca i 1.048.575 redaka.

ID ID	Date Date	Block Area	Primary Type Type	Description Description	Location Description Location	Arrest Arrest: True/False	Domestic Domestic
634 13.1m	2001-01-01 2023-04-22	62845 unique values	THEFT 21% BATTERY 18% Other (4719603) 61%	SIMPLE 12% $500 AND UNDER 8% Other (6240104) 80%	STREET 26% RESIDENCE 17% Other (4451993) 57%	true 2.03m 26% false 5.75m 74%	true 1.08m 14% false 6.71m 86%
10224738	09/05/2015 01:30:00 PM	043XX S WOOD ST	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	false	true
10224739	09/04/2015 11:30:00 AM	008XX N CENTRAL AVE	THEFT	POCKET-PICKING	CTA BUS	false	false
11646166	09/01/2018 12:01:00 AM	082XX S INGLESIDE AVE	THEFT	OVER $500	RESIDENCE	false	true
10224740	09/05/2015 12:45:00 PM	035XX W BARRY AVE	NARCOTICS	POSS: HEROIN(BRN/TAN)	SIDEWALK	true	false

Slika 6: Prikaz odabranih atributa iz originalnog skupa podataka (Izvor: https://www.kaggle.com/datasets/utkarshx27/crimes-2001-to-present)

Slika 6 nam prikazuje odabrane atribute iz skupa podataka. Iako su podaci korisni za manje i osnovne analize, za potrebe ovog završnog rada bilo ih je potrebno dodatno obraditi, nadopuniti kvantitativnim informacijama i prilagoditi njihovu strukturu kako bi se omogućila učinkovitija analiza i opširnost podataka.

Za obradu, generiranje i transformaciju podataka koristila sam programski jezik Python, koji sam odabrala zbog njegove jednostavnosti i brzine u manipulaciji podacima. Podaci su izvorno bili u CSV formatu (zarezom odvojene vrijednosti, eng. Comma Separated Values), a za njihovu obradu najviše su mi pomogle biblioteke pandas, csv i random.

Za čitanje i zapisivanje CSV datoteka koristila sam biblioteku csv, dok sam za učitavanje, obradu i razdvajanje podataka koristila pandas. Pomoću pandas biblioteke razdvojila sam atribute datum i vrijeme, dodala novi atribut strana_grada, koji se određuje na temelju adrese, te atribut cijena_zločina, koji predstavlja cijenu svake vrste zločina. Biblioteka random korištena je za generiranje dodatnih atributa kako bi podaci dobili više kvantitativnih i demografskih informacija. Pomoću nje generirani su atributi količina_svjedoka, koji dodaje broj svjedoka za svaki zločin, spol_počinitelja, prijašnje_kažnjavanje i dobna_skupina, koji dopuštaju bolju analizu profila počinitelja zločina.

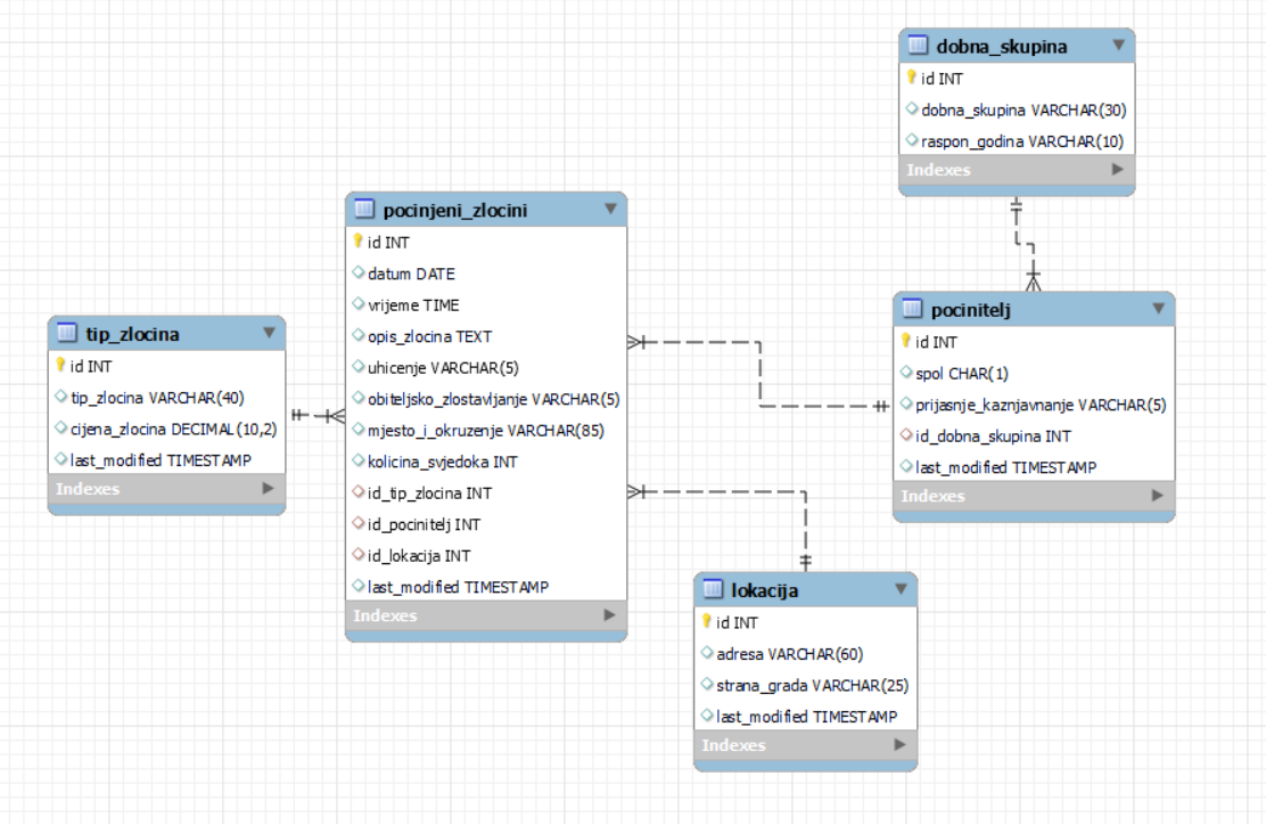
	ID	Primary Type	Description	Location Description	Arrest	Domestic	Date	Time	Block	Witnesses	spol	prijasnje_kaznjavanje	dobna_skupina	raspon_godina	price	Strana_svijeta
1	10224788	OTHER OFFENSE	GUN OFFENDER : ANNUAL REGISTRA	POLICE FACILITY/VEH PARKING LOT	True	False	2014-03-27	0:01	035XX S MICHIGAN AVE	4	F	True	Teen	13-19	300	SOUTH
2	11034701	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $	RESIDENCE	False	False	2001-01-01	11:00	016XX E 86TH PL	11	M	True	Teen	13-19	400	EAST
3	11645601	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $	RESIDENCE	False	False	2014-06-01	0:01	087XX S SANGAMO N ST	5	F	False	Senior	50+	400	SOUTH
4	11645833	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $		False	False	2012-05-05	12:25	057XX W OHIO ST	3	M	True	Young adult	20-29	400	WEST
5	11227517	CRIMINAL SEXUAL	PREDATOR Y	RESIDENCE	False	True	2013-02-10	0:00	071XX S LAFAYETTE	0	F	True	Teen	13-19	2300	SOUTH

Slika 7: Prikaz atributa nakon svih obrada i transformacija - U finalnom obliku

3.2. Izrada baze podataka

Kako bi se izradilo skladište podataka, prvo je potrebno izraditi bazu podataka koja služi kao izvorni sustav za skladište podataka.

Izrada baze podataka započela je analizom skupa podataka te identifikacijom ključnih entiteta i odnosa među njima. Na temelju toga izrađen je prošireni relacijski model entiteta i veza (EER, eng. *Enhanced entity-relationship*) dijagram, koji definira strukturu podataka i veze između entiteta.



Slika 8: EER dijagram

Ovaj EER dijagram prikazuje 5 entiteta: pocinjeni_zlocini, dobna_skupina, pocinitelj, tip_zlocina i lokacija. Entitet u sredini, pocinjeni_zlocini, sadrži podatke o svakom počinjenom zločinu. Atributi koji se nalaze u tom entitetu su datum, vrijeme, kolicina_svjedoka, opis_zlocina, obiteljsko_zlostavljanje(True ili False),

mjesto_i_okruzenje, uhicenje(True ili False), svi strani ključevi povezanih entiteta(id_tip_zlocina, id_pocinitelj, id_lokacija) i last_modified koji bilježi kada se zadnji put dogodila promjena ili prvi unos u tablici. Njegova povezanost sa ostalim entitetima pruža dodatne informacije o svakom zabilježenom slučaju.

Entitet tip_zlocina sadrži podatke o vrstama zločina. Unutar ovog entiteta nalaze se tip_zlocina, cijena_zlocina i last_modified. S obzirom da se jedan tip zločina može pojaviti u više različitih počinjenih zločina, ali svaki pojedinačni zločin pripada samo jednom tipu zločina, veza između ova dva entiteta definirana je kao 1:N (jedan na više).

Entitet lokacija sadrži podatke o lokaciji na kojoj su zločini počinjeni. Njegovi atributi su adresa, strana_grada i last_modified. Svaki počinjeni zločin mora biti povezan sa jednom lokacijom, dok se na istoj lokaciji može dogoditi više zločina. Veza između ta 2 entiteta definirana je kao 1:N(jedan na više).

Entitet pocinitelj sadrži podatke o osobama koje su počinile kaznena dijela. Atributi ovog entiteta uključuju spol(F ili M), prijasnje_kaznjavanje (True ili False) i id_dobna_skupina kao strani ključ entiteta dobna_skupina i last_modified. Veza između entiteta pocinitelj i entiteta pocinjeni_zlocini je 1:N(jedan na više), jer svaki počinitelj može biti odgovoran za više kaznenih dijela , dok svaki pojedinačni zločin može imati samo jednog počinitelja. Također, veza između entiteta pocinitelj i entiteta dobna_skupina je 1:N(jedan na više), jer svaki počinitelj pripada samo jednoj dobnoj skupini, dok jednoj dobnoj skupini može pripadati više počinitelja.

Entitet dobna_skupina sadrži dobne kategorije po rasponu godina i njegovi atributi su dobna_skupina i raspon_godina.

Nakon definiranja strukture baze podataka kroz EER dijagram, za implementaciju baze podataka odabran je MySQL koji je vrlo popularan sustav za upravljanje relacijskim bazama podataka otvorenog koda, također je brz, pouzdan i jednostavan za korištenje. [15]

```
CREATE TABLE pocinitelj (
 id INTEGER AUTO_INCREMENT,
 spol CHAR(1),
 prijasnje_kaznjavnanje VARCHAR(5),
 id_dobna_skupina INT,
 last_modified TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP,
 CONSTRAINT pocinitelj_pk PRIMARY KEY(id),
 CONSTRAINT dobna_skupina_fk FOREIGN KEY (id_dobna_skupina) REFERENCES dobna_skupina (id)
);
```

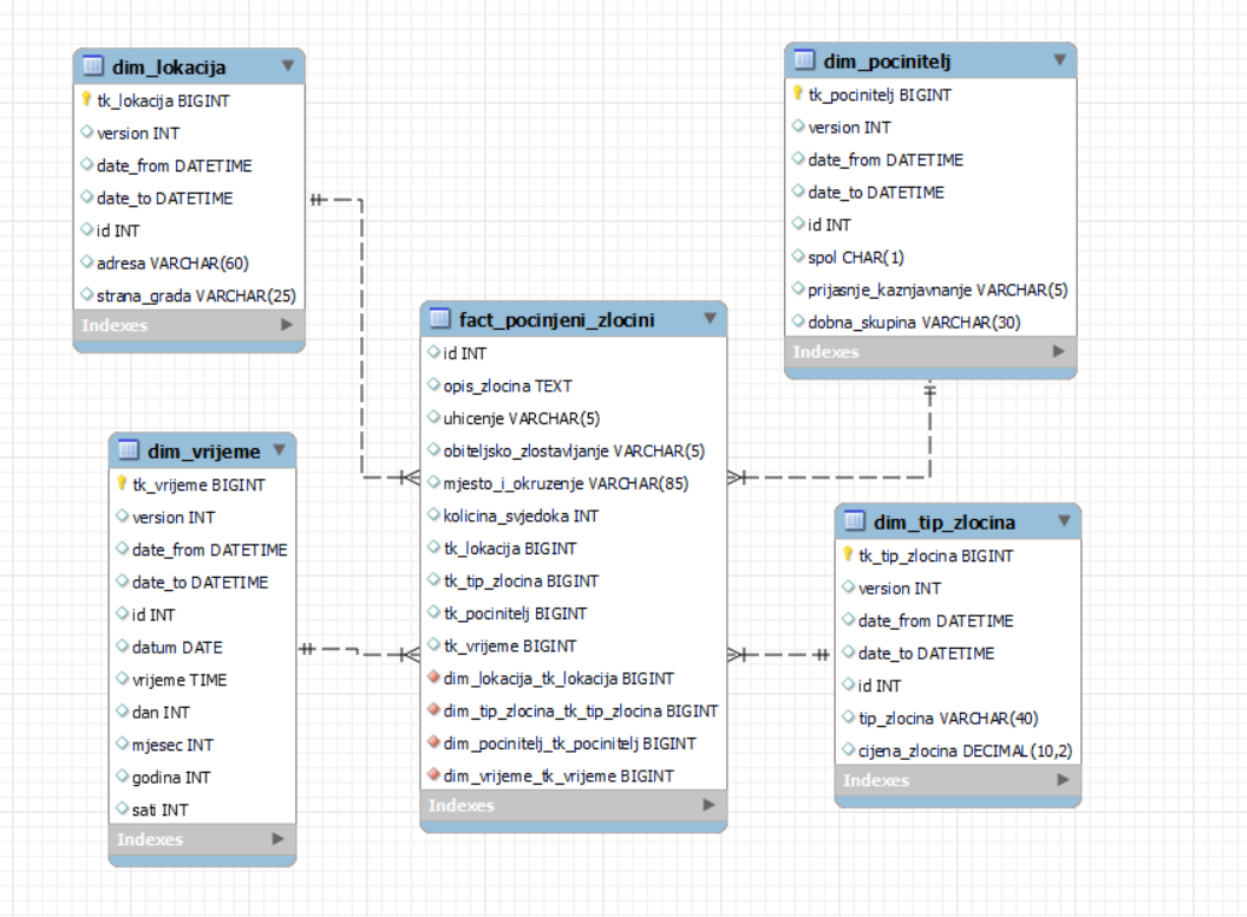
Slika 9: Primjer izrade tablice u MySQL

Kao primjer implementacije tablica u MySQL sustavu, prikazana je izrada tablice pocinitelj. Ova tablica definira više atributa, uključujući primarni ključ (id), koji se automatski generira pomoću AUTO_INCREMENT, te atribute spol, koji je definiran sa CHAR tipom podataka, i prijasnje_kaznjavanje, koji je definiran sa VARCHAR tipom podataka. Također u ovoj strukturi koristi se strani ključ (id_dobna_skupina), tipa podataka INT, koji povezuje svakog počinitelja s njegovom dobnom skupinom u tablici dobna_skupina. I na kraju, atribut last_modified koji je TIMESTAMP i koji poprimi trenutačno vrijeme pri unosu podataka u tablicu i prilikom mijenjanja podataka u tablici zbog pravila DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT TIMESTAMP. Ograničenje osigurava da je primarni ključ jedinstven za svaki zapis u tablici, dok ograničenje za strani ključ osigurava da svaki zapis u tablici mora odgovarati postojećem zapisu u referenciranoj tablici.

Izrada svih tablica je bilo identično samo uz izmijenjene atribute, njihove tipove podataka i njihove strane ključeve.

3.3. Izrada dimenzijskog modela

Nakon što je kreirana baza podataka koja služi kao izvorni sustav za skladište podataka, potrebno je izraditi model prema kojem će se skladište podataka izgraditi i funkcionirati. Za implementaciju skladišta podataka koristi se Kimballova „bottom-up“ metoda, što znači da je model skladišta podataka izrađen koristeći star shemu.



Slika 10: Star schema

Ova star shema prikazuje 1 tablicu činjenica, 4 dimenzijske tablice i 5 degeneriranih dimenzija.

Tablica činjenica *fact_pocinjeni_zlocini* predstavlja tablicu činjenica bez činjenice, što znači da ne sadrži numeričke mjere(mjere promatranja), već bilježi isključivo tehničke ključeve dimenzija i degenerirane dimenzije koje opisuju događaj. Njezina glavna svrha je omogućiti analizu učestalosti počinjenih zločina kroz različite dimenzije, pri čemu se podaci interpretiraju na temelju broja zapisa u tablici. U njoj se nalaze 5 degeneriranih dimenzija opis_zlocina, uhicenje, obiteljsko_zlostavljanje, mjesto_i_okruzenje i kolicina_svjedoka, te dimenzije su izravno povezane sa počinjenim zločinom i imaju nisku kardinalnost što znači da nema potrebe da se stvaraju zasebne dimenzijske tablice za njih.

Dimenzijska tablica dim_vrijeme je vremenska dimenzija koja sadrži hijerarhiju datuma i vremena počinjena nekog zločina, to znači da se podaci mogu analizirati po satu, danu, mjesecu i godini počinjenog zločina.

Dimenzijska tablica dim_lokacija je lokacijska dimenzijska tablica u kojoj nalazimo stranu grada i adresu počinjenog zločina.

Dimenzijska tablica dim_tip_zlocina je dimenzija vrste zločina i u njoj se nalazi tip zločina i cijena svake vrste zločina.

Dimenzijska tablica dim_pocinitelj je dimenzija počinitelja zločina i u njoj se nalaze informacije o počinitelju kao spol, prijašnje kažnjavanje i dobnu skupinu počinitelja.

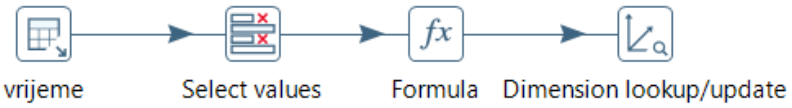
Sve dimenzijske tablice također sadrže tehnički ključ sa kojim su spojene na tablicu činjenica, verziju unosa koja nam govori koliko je zapis puta bio mijenjan i primarni ključ zapisa, sadrže i atribute date_from i date_to koji nam govori od kada do kada je zapis bio aktivan i relevantan.

3.4. ETL proces

Nakon izrade dimenzijskog modela, sljedeći korak je ETL proces, čija je svrha preuzimanje podataka iz izvornog sustava, njihova obrada i učitavanje u skladište podataka. Za realizaciju ETL procesa korišten je Pentaho Data Integration, alat koji omogućuje vizualno kreiranje i automatizaciju ETL procesa putem jednostavnog grafičkog sučelja. [16]

ETL proces je proveden kroz jedno inicijalno punjenje podataka, nakon čega je uslijedilo deset inkrementalnih punjenja koja su ažurirala podatke u skladištu koristeći CDC tehniku. Detaljna implementacija CDC metode opisana je u zasebnom poglavlju.

3.4.1. Dimenzijske tablice



Slika 11: Primjer ETL procesa za vremensku dimenziju u Pentaho Data Integration

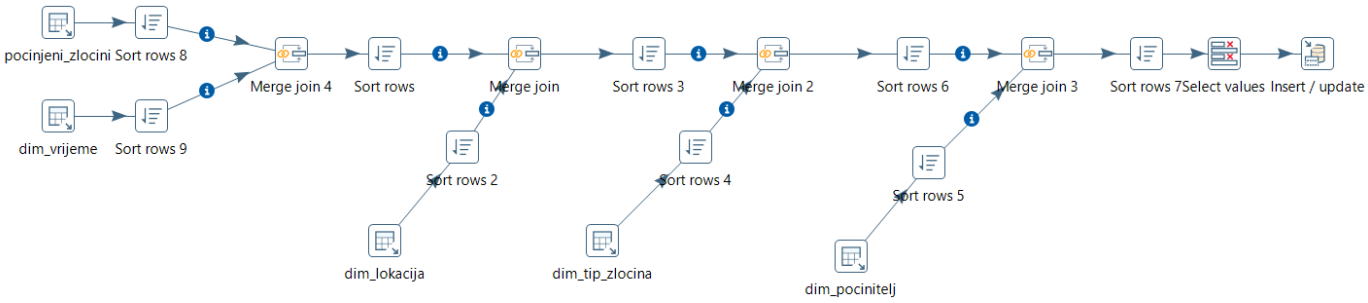
Na slici 11 prikazan je ETL proces za dimenziju vremena, koji se sastoji od tri glavne faze: prikupljanje podataka, transformacija i punjenje podataka u skladište.

Prvi korak u ETL procesu je učitavanje podataka iz izvora, što je na slici prikazano početnom točkom označenom kao "vrijeme". Ovaj korak zapravo predstavlja "Table input", gdje se unosi SQL upit koji dohvaća podatke iz povezane baze podataka. Rezultat upita su podaci koji se spremaju u pripremno područje unutar ETL procesa. U ovoj fazi primjenjuje se CDC tehnika, koja omogućuje inkrementalno punjenje podataka

Nakon prikupljanja podataka, potrebno ih je prilagoditi kako bi bili spremni za skladište podataka. Prvi korak u ovoj fazi je "Select values", gdje se podaci poput datuma i vremena formatiraju da bi nastavili obradu i gdje se eliminiraju nepotrebni podaci. Sljedeći korak u transformaciji je "Formula", gdje se iz postojećih vremenskih podatka izdvajaju pojedine vremenske komponente, poput sata, dana, mjeseca i godine. Kod drugih dimenzija, ovisno o potrebama, može se dodati "Sort rows" kako bi se podaci pravilno organizirali prije spajanja, kao i druge metode transformacije.

Zadnji korak prikazan na slici je "Dimension lookup/update". U njemu se puni skladište podataka i implementira CDC tehnika, osiguravajući inkrementalno ažuriranje podataka u skladištu bez nepotrebnog ponovnog učitavanja svih podataka.

3.4.2. Tablica činjenica



Slika 12: Primjer ETL procesa za tablicu činjenica

ETL proces, prikazan na slici 12, započinje prikupljanjem podataka o počinjenim zločinima iz tablice pocinjeni_zlocini, koja je slična po strukturi tablice fact_pocinjeni_zlocini. Istovremeno se dohvaćaju podaci iz dimenzijskih tablica dim_vrijeme, dim_lokacija, dim_pocinitelj i dim_tip_zlocina, koje su već prošle ETL proces prije tablice činjenica.

U fazi transformacije, podaci se moraju sortirati prije koraka „Merge join“ zato što on očekuje da su podaci sortirani prema ključevima po kojima se spajaju, ako nisu mogu se dogoditi pogreške pri spajanju tablica. Pomoću "Merge join", podaci se spajaju s dimenzijskim tablicama korištenjem stranih ključeva iz tablice pocinjeni_zlocini i pripadajućih id-ova iz dimenzijskih tablica. Tim postupkom svaki zapis u pripremnom području pohrani sve atribute iz dimenzijske tablice po njegovom id-ju. Nakon svakog spajanja potrebno je ponovno potrebno provesti još jedno sortiranje.

Nakon što su svi podaci povezani s dimenzijama, u koraku "Select values" iz transformiranih podataka odabiru se samo oni atributi koji su potrebni za tablicu činjenica. Ovim postupkom eliminiramo sve nepotrebne podatke iz dimenzijskih tablica, spremljene u pripremnom području, i ostavljamo samo tehničke ključeve.

Posljednji korak ETL procesa je punjenje podataka u tablicu činjenica pomoću koraka "Insert/Update", koji provjerava postoji li zapis već u bazi. U ovoj fazi implementiramo CDC tehniku. To znači da se novi zapisi dodaju samo ako prethodno ne postoje, a postojeći zapisi ažuriraju bez dupliciranja podataka, što nam dozvoljava inkrementalno punjenje tablice činjenica.

3.5. Implementacija CDC tehnike

Za osiguravanje inkrementalnog punjenja podataka uz praćenje povijesnih promjena, primijenjena je CDC timestamp metoda u procesu izgradnje skladišta podataka. Proces je započet inicijalnim punjenjem, pri čemu su svi podaci iz izvornog sustava prvi put učitani u skladište podataka. Nakon toga, kroz deset inkrementalnih punjenja, u skladište su se unosili samo novi i ažurirani podaci.

Implementacija CDC tehnike započinje izmjenom strukture baze podataka, koja uključuje dodavanje dodatnog stupca „last_modified“ u svaku relevantnu tablicu. Ovaj stupac

omogućuje praćenje promjena nad podacima te optimizira inkrementalno ažuriranje skladišta podataka.

```
CREATE TABLE tip_zlocina (
 id INTEGER AUTO_INCREMENT,
 tip_zlocina VARCHAR(40) UNIQUE,
 cijena_zlocina DECIMAL(10,2),
 last_modified TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP,
 CONSTRAINT tip_zlocina_pk PRIMARY KEY(id)
);
```

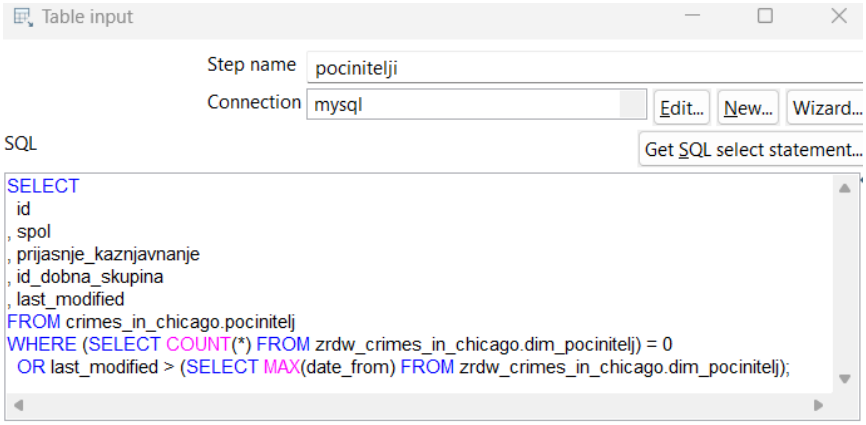
Slika 13: Prikaz atributa "last_modified"

Tijekom INSERT operacije, „last_modified“ automatski preuzima trenutnu vremensku oznaku, dok se pri UPDATE operaciji njegova vrijednost ažurira na vrijeme izvršenja promjene. Ovim pristupom osigurava se kontinuirano praćenje svih izmjena u podacima, čime se smanjuje opterećenje sustava i poboljšava učinkovitost obrade podataka.

Nakon izmjene i punjenja baze podataka sa Pythonom, slijedi implementacija CDC-a u ETL proces.

3.5.1. Dimenzijske tablice

Za implementaciju CDC tehnike u dimenzijske tablice, u ETL procesu, CDC prvo primjenjujemo u fazi prikupljanja.



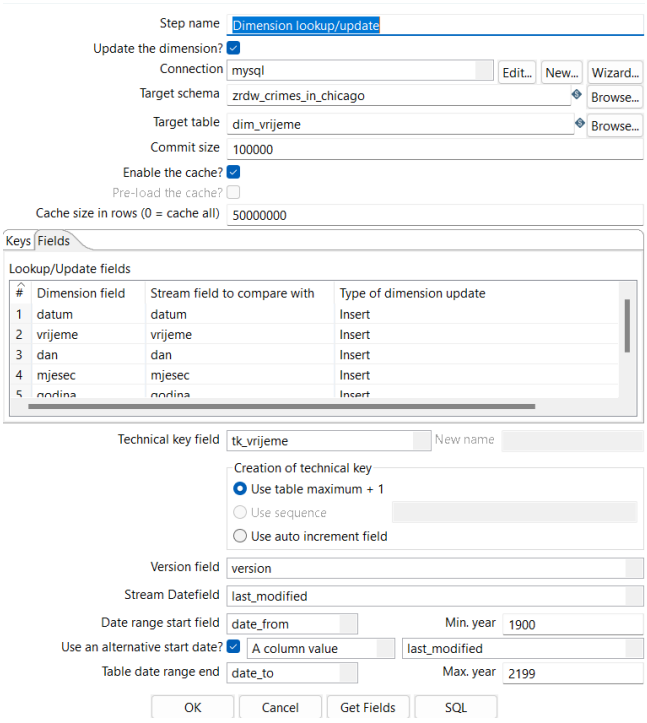
Slika 14:Prikaz Table input-a za dimenziju dim_pocinitelj (faza prikupljanja)

Slika 14 prikazuje SQL upit unutar Pentaho Data Integration alata, kojim se ostvaruje prikupljanje podataka iz tablice „pocinitelj“. Upit koristi CDC timestamp tehniku kako bi osigurao inkrementalno dohvaćanje podataka.

U upitu se najprije provjerava stanje u dimenzijskoj tablici „dim_pocinitelj“. Ako je tablica prazna, svi podaci iz izvorišne baze prenose se u skladište, što predstavlja inicijalno punjenje. U suprotnom, dohvaćaju se samo zapisi kod kojih je vrijednost „last_modified“ veća od najveće vrijednosti „date_from“ u dimenzijskoj tablici „dim_pocinitelj“, čime se utvrđuje vrijeme posljednjeg unesenog podataka u skladište, odnosno prati vrijeme kada je posljednji put izvršeno punjenje skladišta podataka. Na taj način u skladište se unose isključivo novi ili ažurirani zapisi, dok se ne obrađuju podaci koji su ostali nepromijenjeni.

Ovaj proces se koristi i implementira za svaku dimenzijsku tablicu i ovim pristupom možemo razlikovati nove i ažurirane podatke od onih koji su već u skladištu podataka nepromijenjeni.

Nakon prikupljanja podataka i transformiranja istih, slijedi punjenje podataka u skladište podataka. Sa „Dimension lookup/update “ korakom provjeravamo da li zapisi koji su u pripremnom području već postoje.



Slika 15: Prikaz "Dimension Lookup/Update" koraka za dim_vrijeme(faza punjnenja)

Ako zapis već postoji, sustav ažurira njegov atribut „date_to“, postavljajući ga na vrijednost atributa „last_modified“ novog podatka. Nakon toga, dodaje se novi redak u tablicu s ažuriranim zapisom, povećanim atributom „version“ za jedan i novim tehničkim ključem, dok se atribut „date_from“ tog zapisa postavlja na istu vrijednost „last_modified“, a „date_to“ na zadanu vrijednost „2200-01-01 00:00:00“.

tk_lokacija	version	date_from	date_to	id	adresa	strana_grada
1	1	2024-10-31 01:29:58	2024-10-31 01:32:08	1	043XX S WOOD ST	SOUTH
40173	2	2024-10-31 01:32:08	2024-10-31 01:35:08	1	043XX S WOOD STA	SOUTH
40174	3	2024-10-31 01:35:08	2200-01-01 00:00:00	1	043XX S WOOD ST	SOUTH
NULL	NULL	NULL	NULL	NULL	NULL	NULL

Slika 16: Prikaz implementacije CDC tehnike kada se zapis ažurira u lokacijskoj dimenziji

Ako zapis ne postoji, dodaje se novi redak u tablicu dimenzije vremena.

	tk_lokacija	version	date_from	date_to	id	adresa	strana_grada
▶	2	1	2024-10-31 01:29:58	2200-01-01 00:00:00	2	008XX N CENTRAL AVE	NORTH
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Slika 17: Prikaz implementacije CDC metode kada je zapis nov u lokacijskoj dimenziji

Na slici 16 vidimo zapis o lokaciji koja je kroz vrijeme prolazila promjene.

Prvi zapis prikazuje početnu verziju lokacije s tk_lokacija = 1, koja je bila aktivna od 2024-10-31 01:29:58 do 2024-10-31 01:32:08. Nakon toga, nova verzija lokacije (tk_lokacija = 40173) preuzima njeno mjesto, pri čemu se date_to prvog zapisa postavlja na trenutak kada je ažurirana verzija stupila na snagu.

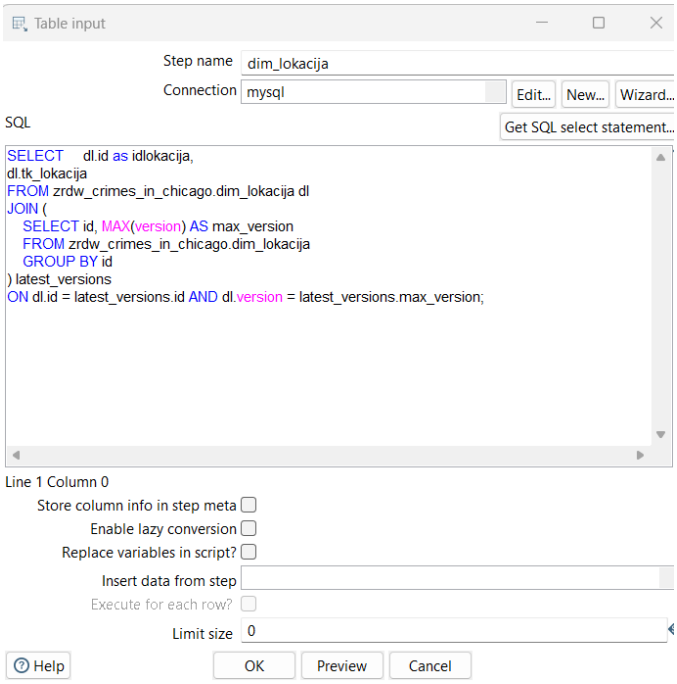
Treća verzija (tk_lokacija = 40174) unosi se 2024-10-31 01:35:08, čime se prethodni zapis zatvara (date_to = 2024-10-31 01:35:08). Ova verzija ostaje aktivna sve do daljnjih promjena, što je označeno postavljanjem date_to = 2200-01-01 00:00:00, što se često koristi kao oznaka "trenutno važećeg" zapisa.

Kroz ove promjene možemo pratiti kako su se podaci o lokaciji ažurirali kroz vrijeme bez gubitka povijesnih podataka. Ova metoda omogućava detaljno analiziranje promjena kroz vremenske periode, što je ključno za skladišta podataka i analitičke svrhe.

3.5.2. Tablica činjenica

Za tablice činjenica je drugačiji proces, jer se uglavnom ne mijenjaju, već se u njih dodatno unose novi podaci. Međutim, u ovome slučaju ažuriramo samo tehničke ključeve dimenzijskih tablica.

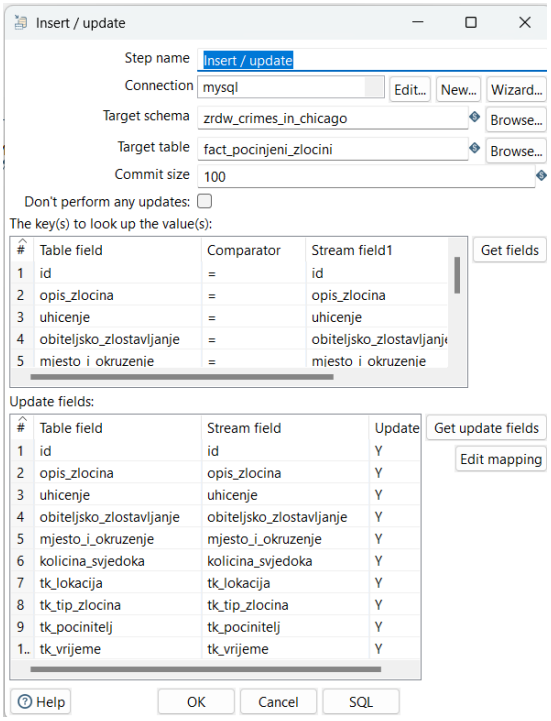
Nakon dodavanja svih zapisa iz tablice „pocinjeni_zlocini“ u pripremno područje, spajamo i sortiramo ih sa dimenzijskim tablicama kako bi dobili tehničke ključeve za svaku dimenziju.



Slika 18: Prikaz dohvaćanja tehničkog ključa dimenzijske tablice "dim_lokacija"

Slika 18 prikazuje dohvaćanje najnovijih verzija zapisa iz dimenzijske tablice dim_lokacija. Prvo se u podupitu identificira najveća vrijednost verzije za svaku lokaciju, ona osigurava da se za svaku lokaciju uzme samo najnoviji zapis. Zatim se glavni upit spaja s dimenzijskom tablicom dim_lokacija na temelju id i najveće verzije, čime se filtriraju samo trenutno aktualni zapisi. Ovaj upit osigurava da tablica činjenica uvijek koristi ažurirane tehničke ključeve, sprječavajući prisustvo zastarjelih podataka u tablici činjenica.

Nakon prikupljanja i transformacije podataka započinje proces punjenja tablice činjenica. Sa „Insert/Update“ korakom dodajemo nove podatke u tablicu činjenica, dok se postojeći zapisi ne prepravljaju, osim ažuriranja tehničkih ključeva dimenzijskih tablica.



Slika 19: Prikaz "Insert/Update" koraka koji puni tablicu činjenica

Ako zapis ne postoji, novi redak se dodaje u tablicu činjenica s pripadajućim tehničkim ključevima za lokaciju, tip zločina, počinitelja i vrijeme, koji su prethodno dohvaćeni spajanjem podataka s odgovarajućim dimenzijama.

id	opis_zlocina	uhicenje	obiteljsko_zlostavljanje	mjesto_i_okruzenje	kolicina_svjedoka	tk_lokacija	tk_tip_zlocina	tk_pocinitelj	tk_vrijeme
1048575	SIMPLE	True	False	OTHER	24	3774	4	1048575	1048575

Slika 20: Prikaz zadnje dodanog zapisa u tablicu činjenica

Ako zapis već postoji, znači da se neka od dimenzija promijenila i da je tehnički ključ promijenjen. Onda se ažuriraju samo tehnički ključevi dimenzijskih tablica koji su mijenjani, kako bi se osiguralo da tablica činjenica referencira najnovije verzije podataka iz dimenzija.

id	opis_zlocina	uhicenje	obiteljsko_zlostavljanje	mjesto_i_okruzenje	kolicina_svjedoka	tk_lokacija	tk_tip_zlocina	tk_pocinitelj	tk_vrijeme
1	DOMESTIC BATTERY SIMPLE	False	True	RESIDENCE	24	1	1	1	1048576

Slika 21: Prikaz zapisa tablice činjnenica prije promjene dimenzije lokacije (Slika 16)

id	opis_zlocina	uhicenje	obiteljsko_zlostavljanje	mjesto_i_okruzenje	kolicina_svjedoka	tk_lokacija	tk_tip_zlocina	tk_pocinitelj	tk_vrijeme
1	DOMESTIC BATTERY SIMPLE	False	True	RESIDENCE	24	40174	1	1	1048576

Slika 22: Prikaz tablice čijenice nakon promjene dimenzije lokacije (Slika 16)

Slika 21 i 22 prikazuju promjenu tk_lokacija u tablici činjenica, koja je prouzrokovana promjenom adrese u lokacijskoj dimenziji „dim_lokacija“ sa slike 16. Ova promjena dogodila se jer je u dimenzijskoj tablici dodan novi red s ažuriranom adresom, čime je kreiran novi tehnički ključ. Budući da se u tablici činjenica koriste tehnički ključevi iz dimenzija, na sljedećem inkrementalnom punjenju tablice činjenica bilo je potrebno ažurirati tk_lokacija kako bi se osigurala ispravna referenca na novu verziju lokacijskog zapisa.

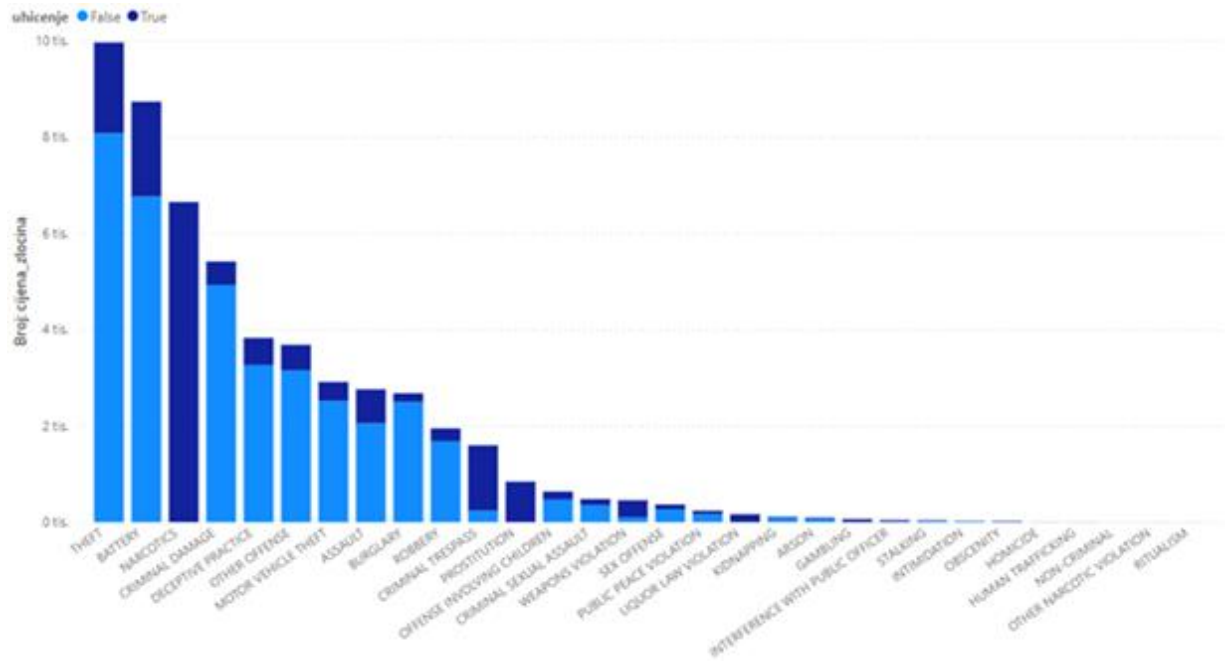
Implementacija CDC timestamp metode omogućila je inkrementalno punjenje skladišta podataka uz praćenje povijesnih promjena bez potrebe za ponovnim učitavanjem cijelog skupa podataka. Korištenjem stupca "last_modified" u izvornoj bazi te CDC timestamp pristupa u dimenzijama i tablici činjenica, osigurana je točnost i konzistentnost podataka kroz vrijeme. Ovaj pristup optimizira ETL proces, smanjuje opterećenje sustava i omogućuje preciznu analitičku obradu ažuriranih podataka.

3.6. OLAP analiza

Nakon stvaranja skladišta podataka, potrebno je napraviti vizualizaciju tih podataka kako bi se dobio bolji uvid u informacije i lakše analizirali ključni trendovi. Vizualizacija podataka pomaže u prepoznavanju obrazaca i odstupanja koja bi inače bila teže uočljiva

analizom sirovih podataka. U ovom radu OLAP analiza se provodi pomoću alata Power BI, on je Microsoftov alat za poslovnu inteligenciju koji služi za jednostavnu analizu i vizualizaciju podataka. [17]

U nastavku su prikazani grafikoni koji vizualiziraju različite aspekte zločina, od učestalosti zločina prema tipu do učestalosti zločina prema spolovima.



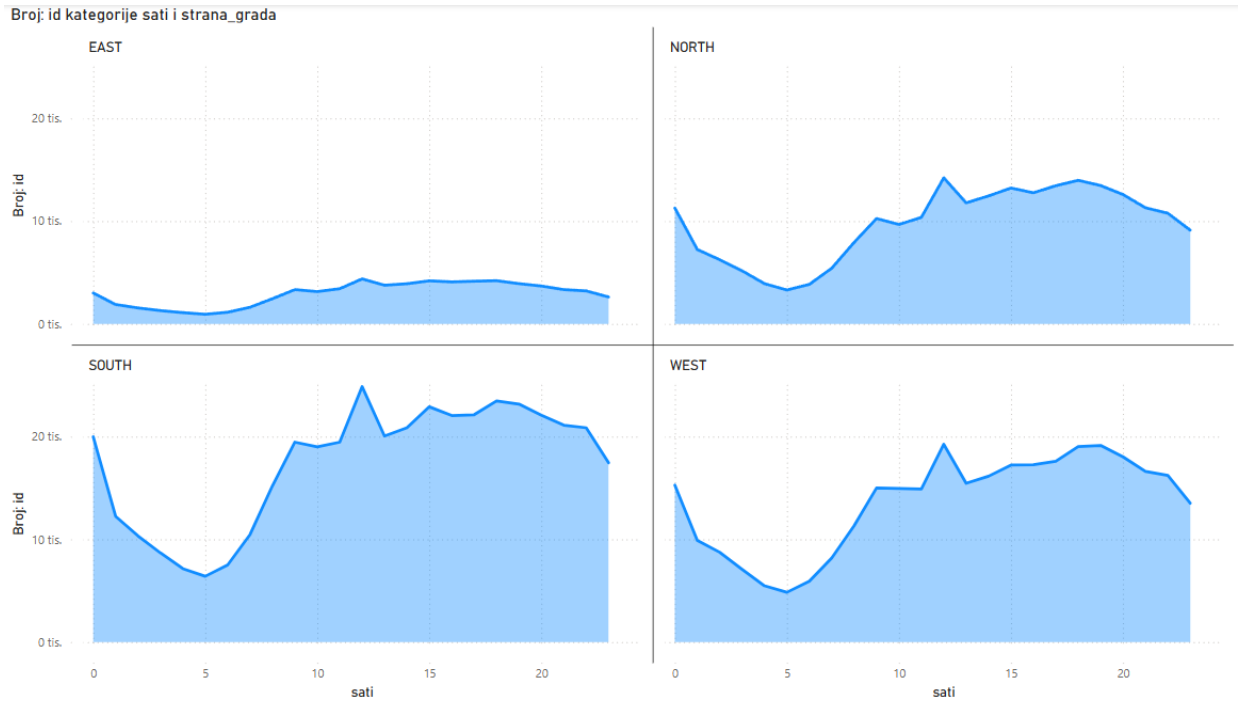
Slika 23: Vizualizacija kriminalnih aktivnosti prema tipu zločina i uhićenja za ista

Slika 23 prikazuje broj počinjenih zločina prema njihovom tipu te uspoređuje slučajeve u kojima je počinitelj uhićen s onima gdje nije došlo do uhićenja. Vidljivo je da su najčešće kaznena djela krađa, nasilje i kaznena djela povezana s drogom. Krađa se ističe kao najzastupljeniji oblik kriminala, no većina takvih slučajeva ne završava uhićenjem. S druge strane, kaznena djela povezana s drogom imaju ogromnu stopu uhićenja, što može biti posljedica aktivne provedbe zakona i smišljene policijske strategije.

Nasilna kaznena djela, poput fizičkog napada (eng. *Battery*) i prijetnje napada (eng. *Assault*), imaju veću stopu uhićenja u usporedbi s imovinskim zločinima, što upućuje na bržu reakciju policije u situacijama koje predstavljaju prijetnju fizičkoj sigurnosti.

Ozbiljniji zločini, poput trgovine ljudima i ubojstva rjeđi su, ali gotovo uvijek rezultiraju uhićenjem počinitelja.

Ova analiza ukazuje na razlike u procesuiranju različitih vrsta kaznenih djela. Dok su uhićenja učestalija kod nasilnih i teških kaznenih djela, imovinski zločini, unatoč visokoj zastupljenosti, često ostaju bez identifikacije i privođenja počinitelja, što može upućivati na izazove u provođenju istraga.



Slika 24: Vizualizacija kriminalnih aktivnosti tijekom dana u različitim dijelovima grada

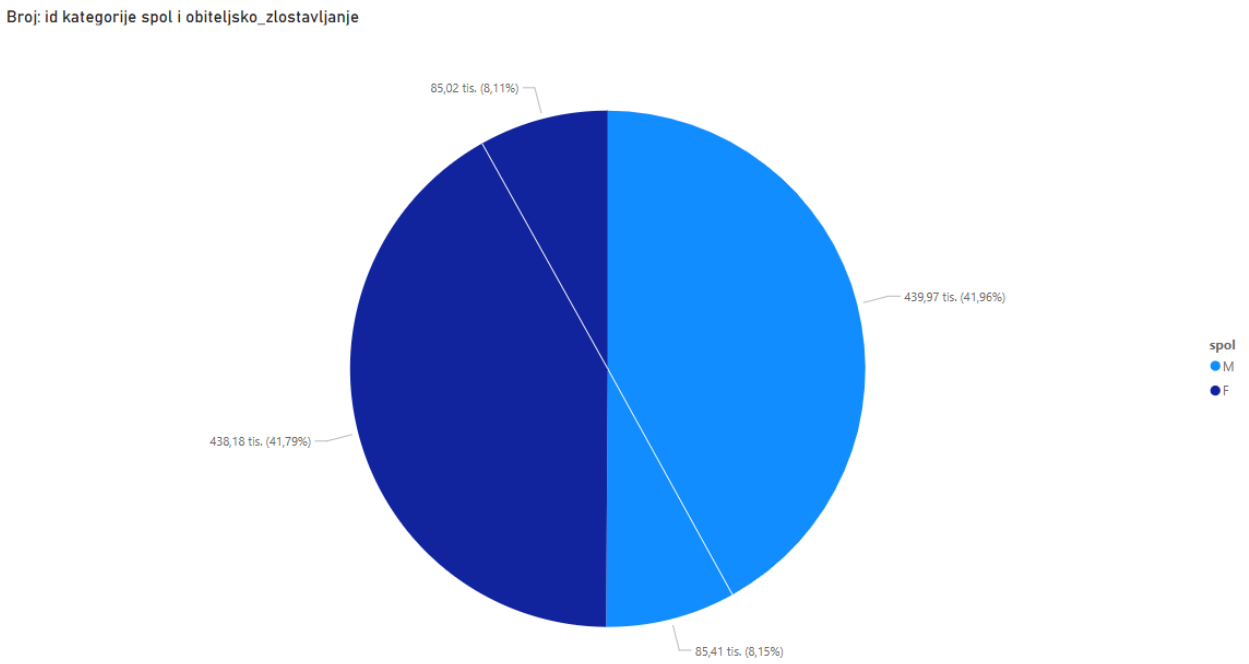
Slika 24 prikazuje distribuciju kriminalnih aktivnosti tijekom dana u različitim dijelovima grada. Podaci su prikazani odvojeno za istok, sjever, jug i zapad. Na istoku grada kriminal je tijekom cijelog dana relativno miran.

Pored toga, na sjeveru kriminalna aktivnost je najniža u jutarnjim satima, nakon čega raste tijekom dana i doseže vrhunac u poslijepodnevnim satima.

Južni dio grada pokazuje drugačiji obrazac, broj zločina u jutarnjim satima je najveći od svih strana grada i značajno raste sredinom dana i ostaje visok do kasnih večernjih sati.

Također, na zapadu kriminal je slično prisutan u ranim jutarnjim satima, ali postupno raste i ostaje na visokoj razini tijekom većeg dijela dana.

Ovi podaci pokazuju da jug i zapad bilježe najviše kriminalnih aktivnosti u popodnevnim i večernjim satima, dok je istok najsmireniji tijekom cijelog dana. Takve informacije mogu biti korisne u planiranju policijskih aktivnosti i preventivnih mjera.



Slika 25: Vizualizacija učestalosti zločina po spolu i po obiteljskom nasilju prema spolu

Slika 25 nam prikazuje odnos između spola i slučajeva obiteljskog zlostavljanja. Vidljivo je da su muškarci i žene gotovo podjednako zastupljeni kao počinitelji, pri čemu muškarci čine 41,96% prijavljenih slučajeva, dok su žene odgovorne za 41,79%. Udio obiteljskog zlostavljanja iz i jedne i druge strane iznosi 8%.

Ovakvi podaci pokazuju da obiteljsko zlostavljanje nije isključivo povezano s jednim spolom, već da ga počine i muškarci i žene. Međutim, važno je napomenuti da ova statistika prikazuje samo prijavljene slučajeve, dok mnogi incidenti ostaju neprijavljeni zbog različitih društvenih i pravnih faktora. Strah od agresora, nepovjerenje u institucije i emocionalna povezanost sa zlostavljačem mogu dovesti do toga da žrtve ne prijave nasilje, što može uzrokovati iskrivljenu sliku o stvarnom omjeru obiteljskog nasilja u počinjenih zločinima.

Zaključak

U ovom radu istražena je i implementirana izrada skladišta podataka o zločinima u Chicagu uz primjenu CDC tehnike za inkrementalno punjenje. Kroz teorijski dio objašnjeni su koncepti skladištenja podataka, razlike između OLTP i OLAP sustava, te ključne metodologije dimenzionalnog modeliranja, pri čemu su obrađeni Inmonov i Kimballov pristup. Nakon toga, detaljno je opisan ETL proces, CDC tehnike i OLAP analiza.

Implementacija je uključivala izradu relacijske baze podataka koja je zatim transformirana u skladište podataka korištenjem star sheme. Primjenom Pentaho Data Integration alata uspješno je kreiran ETL proces koji osigurava kvalitetno prebacivanje podataka iz operativnih izvora u skladište podataka i implementaciju CDC tehnike koja značajno doprinosi učinkovitosti skladišta podataka, bržu obradu promjena i smanjenje nepotrebnih ponovnih učitavanja. Nakon toga, analiza podataka provedena je pomoću OLAP tehnologija.

Primjena vizualizacijskih alata, poput Power BI-a omogućili su dublji uvid u obrasce kriminala u Chicagu te su pokazali potencijal skladišta podataka u analitičkim procesima. Analizom učestalosti zločina po vremenskim periodima, tipovima i lokacijama u Chicagu, identificirani su trendovi koji mogu biti korisni za buduće preventivne mjere i sigurnosne strategije.

Literatura

1. Bhatia, P. (2019). *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press.
2. Inmon, W. H. (2005). *Building the Data Warehouse*. Wiley.
3. „What is OLTP?“ Link na stranicu: https://www.oracle.com/hr/database/what-is-oltp/
4. „OLTP vs OLAP“ Link na stranicu: https://www.integrate.io/blog/oltp-vs-olap/
5. Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Wiley.
6. Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley.
7. „Data Warehouse Concepts: Kimball vs. Inmon Approach“ Link na stranicu: https://www.astera.com/type/blog/data-warehouse-concepts/
8. „Kimball vs Inmon: Which approach should you choose when designing your data warehouse architecture?“ Link na stranicu: https://www.keboola.com/blog/kimball-vs-inmon
9. Turban, E., Sharda, R., Delen, D., & King, D. (2010). *Business intelligence: A .managerial approach* (2nd ed.). Pearson College Div.
10. Kimball, R., & Ross, M. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling* (2nd ed.). John Wiley & Sons.
11. Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies.* McGraw-Hill.
12. „What Is Change Data Capture (CDC): Methods, Benefits, and Challenges“ Link na stranicu : https://www.astera.com/type/blog/change-data-capture-cdc/
13. „What is OLAP?“ Link na stranicu: https://www.ibm.com/think/topics/olap
14. „Crimes In Chicago (2001 to 2023)“ Link na stranicu: https://www.kaggle.com/datasets/utkarshx27/crimes-2001-to-present
15. „Introduction to MySQL“ Link na stranicu: https://www.w3schools.com/mySQl/mysql_intro.asp

16. „Pentaho Data Integration“ Link na stranicu: https://pentaho.com/products/pentaho-data-integration/
17. „Power BI“ Link na stranicu: https://www.microsoft.com/en-us/power-platform/products/power-bi

Popis slika

Popis tablica