SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Početak rada s Jezerskim Skladištem Podataka: Teorija i Tehnologija

Ivan Derdić
Voditelj: *Alan Joivić*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod

Uvod rada. Nakon uvoda dolaze poglavlja u kojima se obrađuje tema.

2. Teorija

U ovom poglavlju se obrađuje teorija Jezerskog Skladišta Podataka. Ne obrađuje se cijela arhitektura nego samo odabrani slojevi te se daje definicija samog Jezerskog Skladišta Podataka. Odabrani slojevi su:

- Sloj Unosa Podataka,
- Sloj Pohrane Podataka,
- Sloj Obrade Podataka.

Odabirom određenih slojeva, u ovom radu se obrađuje samo dio arhitekture Jezerskog Skladišta Podataka, ali se obrađuje teorija koja je potrebna za razumijevanje tehnologije koja se koristi u praktičnom dijelu rada.

2.1. Jezersko Skladište Podataka

Jezersko Skladište Podataka (eng. Data Lakehouse) je arhitektura skladištenja podataka koja kombinira karakteristike Jezera Podataka (eng. Data Lake) i Skladišta Podataka (eng. Data Warehousea). To je centralizirano skladište podataka koje omogućava analizu velike količine strukturiranih i nestrukturiranih vrsta podataka u realnom vremenu ili u kasnijem trenutku. Jezersko skladište podataka omogućuje integraciju podataka iz različitih izvora, olakšava upravljanje podacima, smanjuje troškove i vrijeme potrebno za pripremu podataka za analizu. Ova arhitektura skladištenja podataka postaje sve popularnija u posljednje vrijeme jer olakšava analizu podataka, izvještavanje i donošenje odluka u stvarnom vremenu. Za detaljniji opis Jezerskog Skladišta Podataka vidjeti [1].

2.2. Sloj Unosa Podataka

Sloj Unosa Podataka (engl. Data Ingestion Layer) u Jezerskom Skladištu Podataka je sloj koji se koristi za prikupljanje i spremanje podataka iz različitih izvora u Jezersko

Skladište Podataka. Ovaj sloj obuhvaća dva načina prikupljanja podataka:

1. serijsko prikupljanje podataka,
2. strujno prikupljanje podataka.

Sloj Unosa Podataka omogućuje podatkovnim inženjerima da učinkovito prikupe podatke iz različitih izvora i formata, poput baza podataka, datoteka ili senzorskih uređaja, te ih jednostavno učitaju u Jezersko Skladište Podataka. Ovaj sloj obično uključuje alate za obradu velikih količina podataka, poput Apache Spark-a, kako bi se omogućilo brzo i učinkovito prikupljanje i spremanje velikih količina podataka. Za detaljni opis Sloja Unosa Podataka vidjeti [1].

2.3. Sloj Pohrane Podataka

U Jezerskom Skladištu Podataka, Sloj Pohrane Podataka obuhvaća skup tehnologija i alata za pohranu velikih količina podataka u različitim formatima, kao što su Apache Hadoop Distributed File System (HDFS), Amazon S3, Azure Blob Storage i Google Cloud Storage.

Osim toga, Delta Lake tehnologija može se koristiti kao sloj pohrane podataka u Jezerskom Skladištu, jer omogućuje verzioniranje podataka, upravljanje transakcijama i omogućuje pohranu podataka u strukturiranom obliku, čime se olakšava proces analize.

Podaci u Sloju Pohrane Podataka se dijele na 3 razine:

- **Sirovi podaci/Brončani sloj** - podaci koji su prikupljeni iz izvora podataka,
- **Obrađeni podaci/Srebrni sloj** - podaci koji su transformirani i pripremljeni za analizu,
- **Analizirani podaci/Zlatni sloj** - podaci koji su analizirani i spremljeni u agregiranom obliku.

Sloj pohrane podataka u Jezerskom Skladištu Podataka mora biti dizajniran i konfiguriran na način koji omogućuje brzi i jednostavan pristup podacima za analizu, uz osiguravanje pouzdanosti, sigurnosti i skalabilnosti skladišta. Za detaljni opis Sloja Pohrane Podataka vidjeti [1].

2.4. Sloj Obrade Podataka

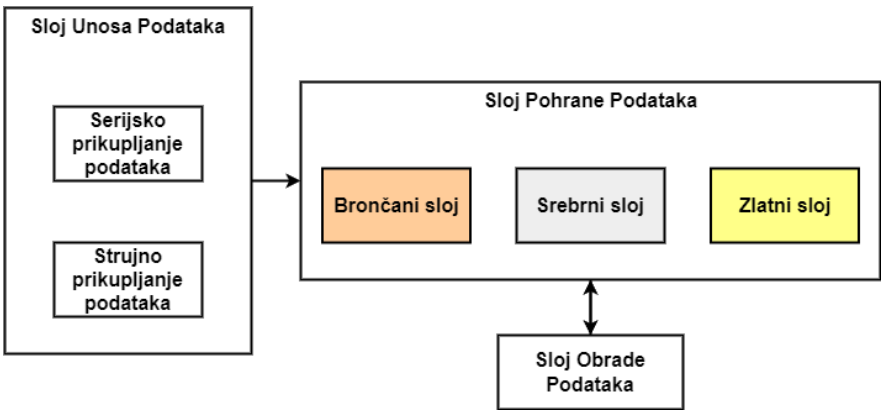
Sloj Obrade Podataka (eng. Data Processing Layer) u Jezerskom Skladištu Podataka odnosi se na skup tehnologija i alata koji omogućuju obradu velikih količina podataka pohranjenih u Jezerskom Skladištu Podataka. Ovaj sloj obično uključuje distribuirane obradne okvire, poput Apache Sparka, Apache Flinka i Apache Beam-a.

Sloj Obrade Podataka je dizajniran kako bi omogućio izvođenje različitih operacija na podacima, uključujući čišćenje, transformiranje, spajanje i agregiranje. Ovaj sloj omogućuje korisnicima da lako i učinkovito izvode složenu obradu velikih skupova podataka, koristeći alate za distribuiranu obradu.

Sloj Obrade Podataka u Jezerskom Skladištu Podataka igra ključnu ulogu u omogućavanju pouzdane, skalabilne i brze obrade podataka pohranjenih u Jezerskom Skladištu Podataka, što omogućuje korisnicima da izvuku vrijednost iz podataka i donose informirane poslovne odluke. Za detaljni opis Sloja Obrade Podataka vidjeti [1].

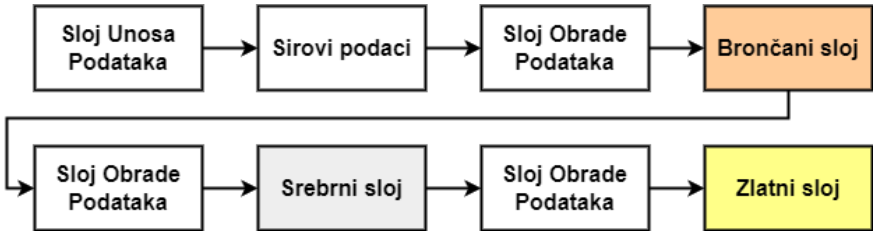
2.5. Arhitektura i tok podataka

Arhitektura Jezerskog Skladišta Podataka prikazana je na slici (2.1). Ona proizlazi iz slojeva opisanih u poglavljima (2.2), (2.3) i (2.4). Sloj Unosa Podataka samo upisuje podatke u Sloj Pohrane Pohrane Podataka, dok Sloj Obrade Podataka čita i piše podatke u Sloj Pohrane Podataka. Sloj Obrade Podataka čita i piše podatke jer u tom sloju čišćenjem, transformiranjem, spajanjem i agregiranje podataka ostvaruju podslojevi (Brončani, Srebrni i Zlatni Sloj) Sloja Pohrane Podataka.



Slika 2.1: Arhitektura modela Jezerskog Skladišta Podataka.

Sa slike (2.2) se vidi da postoji tok podataka između Slojeva Unosa, Pohrane i Obrade Podataka. Sloj Unosa Podataka upisuje podatke iz vanjske okoline (baze po-



Slika 2.2: Tok podataka u Jezerskom Skladištu Podataka.

dataka, SFTP serveri, Jezera Podataka) u Sloj Pohrane Podataka. Unosom podataka se dobivaju Sirovi podaci koje Sloj Obrade Podataka dohvaća i unosi u Brončani sloj s najmanjim mogućim brojem transformacija. Sljedeće, Sloj Obrade Podataka dohvaća podatke iz Brončanog sloja te ih čisti, transformira i spaja. Obrađeni podaci Brončanog sloja se unose u Srebrni sloj. Na kraju, Sloj Obrade Podataka dohvaća podatke iz Srebrnog sloja te ih agregira i unosi u Zlatni sloj.

3. Tehnologija

U ovom poglavlju se obrađuju tehnologije koje se koriste za ostvarivanje slojeva Jezerskog Skladišta Podataka. U ovom poglavlju opisane su sljedeće tehnologije:

- Apache Spark,
- Delta Lake.

3.1. Spark

Spark je distribuirani okvir za obradu podataka otvorenog koda koji je dizajniran za brzu i jednostavnu obradu velikih količina podataka. Spark omogućava programerima da razviju složene aplikacije za obradu podataka i analizu u nekoliko programskih jezika, uključujući Java, Python, Scala i R.

Spark je popularan zbog svoje sposobnosti da brzo i jednostavno obradi velike količine podataka izvršavanjem na grozdu. Osim toga, Spark pruža mnoge biblioteke i alate za obradu podataka, uključujući Spark SQL, Spark Streaming, MLlib i GraphX.

Spark se često koristi u industriji za obradu podataka u stvarnom vremenu, strojno učenje, obradu teksta, obradu slika i druga područja primjene. Spark je postao popularan zbog svoje brzine obrade podataka, skalabilnosti i fleksibilnosti u korištenju.

Spark se u Jezerskom Skladištu Podataka može koristiti za ostvarivanje Sloja Unosa Podataka (vidi poglavlje (2.2)) i Sloja Obrade Podataka (vidi poglavlje (2.4)). Za detaljniji opis Spark-a vidjeti poglavlje 1 iz [2, c.1].

3.2. Delta Lake

Delta Lake je open source projekt koji se temelji na Apache Spark-u, a namijenjen je upravljanju i obradi podataka u velikim i složenim analitičkim aplikacijama. Delta Lake kombinira karakteristike Data Lake-a i skladišta podataka (Data Warehouse-a) u

jednoj platformi koja je skalabilna, otporna na pogreške i sposobna za rad u realnom vremenu.

Delta Lake omogućuje pohranu podataka u obliku tabela, s podrškom za transakcije i verzioniranje. To omogućuje korisnicima da jednostavno dodaju, ažuriraju ili brišu podatke, dok se istovremeno održava povijest promjena.

Delta Lake također pruža podršku za naprednu obradu podataka, poput upravljanja s vremenom, verzioniranja shema i upravljanja sinkronizacijom podataka. Ove značajke olakšavaju integraciju s drugim alatima i aplikacijama, što je korisno u velikim poduzećima s kompleksnim IT okruženjima.

Delta Lake se često koristi u poslovima koji zahtijevaju brzu obradu podataka u stvarnom vremenu i velike količine podataka, poput bankarstva, telekomunikacija, e-trgovine i drugih industrija koje se bave velikim količinama podataka.

Delta Lake je tehnologija za upravljanje i obradu podataka koja se može koristiti u sklopu Jezerskog Skladišta Podataka. Može se koristiti za ostvarivanje Sloja Pohrane Podataka (vidi poglavlje (2.3)). Za detaljniji opis Delta Lake-a vidjeti [3]

4. Zaključak

Zaključak.

5. Literatura

[1] P. Menon, *Data Lakehouse in Action*. Packt Publishing Pvt Ltd, 2022.

[2] J. S. Damji, B. Wenig, T. Das, and D. Lee, *Learning Spark: Lightning-Fast Data Analytics*. O'Reilly Media, 2020.

[3] Microsoft, "What is delta lake?," 2023.

6. Sažetak

Sažetak.