SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

**SEMINAR**

# Početak rada s Jezerskim Skladištem Podataka: Teorija i Tehnologija

*Ivan Derdić*
Voditelj: *Alan Joivić*

Zagreb, travanj 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. U ovom poglavlju dana je definicija Jezerskog Skladišta Podataka, te su opisani sljedeći slojevi:

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

## 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 (Menon, 2022).

# 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. Za detaljniji opis Spark-a vidjeti poglavlje 1 iz (Damji et al., 2020, 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. Za detaljniji opis Delta Lake-a vidjeti (Microsoft, 2023)

# 4. Zaključak

Zaključak.

# 5. Literatura

Jules S. Damji, Brooke Wenig, Tathagata Das, i Denny Lee. *Learning Spark: Lightning-Fast Data Analytics*. O'Reilly Media, 2020.

Pradeep Menon. *Data Lakehouse in Action*. Packt Publishing Pvt Ltd, 2022.

Microsoft. What is delta lake?, 2023. URL `https://learn.microsoft.com/en-us/azure/databricks/delta/`.

# 6. Sažetak

Sažetak.