POSLOVNA SOFTVERSKA OS REŠENJA

školska 2024/2025 godina

Vežba 5: Uvod u rad sa velikim podacima (Big Data)

Big Data se odnosi na **ogromne količine podataka** koji se svakodnevno generišu iz različitih izvora i koji se **ne mogu efikasno obraditi klasičnim alatima** kao što su tradicionalne relacione baze podataka ili Excel. Ovi podaci zahtevaju specijalizovane alate, tehnologije i pristupe za skladištenje, obradu i analizu.

U savremenim **poslovnim softverskim rešenjima**, Big Data omogućava kompanijama da:

- **Prikupe i objedine informacije** iz raznih izvora: veb sajtova, mobilnih aplikacija, IoT uređaja, baza podataka, server logova, društvenih mreža, senzora, transakcija itd.
- **Otkrivaju obrasce ponašanja** korisnika, tržišnih trendova, navika kupovine, rizika, prevara itd.
- **Donose pametnije poslovne odluke** pomoću alata za poslovnu analitiku (BI) i mašinsko učenje.

Na primer:

🔷 **Netflix – Analiza ponašanja gledalaca**
Netflix prikuplja podatke o svakoj vašoj interakciji:

- koje serije gledate,
- kada pauzirate ili prekinete gledanje,
- koje žanrove volite,
- koje uređaje koristite.

Na osnovu toga, sistem koristi **Big Data analitiku** i **mašinsko učenje** kako bi vam predložio sadržaje koje najverovatnije želite da gledate. Ova personalizacija značajno povećava zadovoljstvo korisnika i vreme provedeno na platformi.

🔷 **Amazon – Preporuka proizvoda**

Amazon koristi Big Data da prati **istoriju kupovina**, **pretrage**, **recenzije**, **proizvode koje ste gledali**, kao i ponašanje korisnika koji su slični vama. Na osnovu toga, **mašinski algoritmi analiziraju obrasce** i automatski predlažu proizvode koje biste mogli želeti da kupite.

🔷 **Banke – Detekcija prevara u realnom vremenu**

U bankarstvu, Big Data sistemi analiziraju **milione transakcija u sekundi**.
Ako algoritam otkrije neobičan obrazac – npr.

- transakcija sa nepoznate lokacije,
- neuobičajena suma novca,
- česta kupovina u kratkom vremenskom periodu –
  automatski se aktivira sistem za **detekciju potencijalne prevare**.

🔪 **5V Karakteristike Big Data**

Big Data se definiše pomoću **5 ključnih osobina** – poznatih kao **5V model**:

1. **Volume (Obim podataka)**
   Predstavlja količinu podataka koja može da dostigne terabajte, petabajte, pa i eksabajte. Na primer, Facebook dnevno generiše preko 4 petabajta podataka.
2. **Velocity (Brzina)**
   Podaci se generišu i moraju biti obrađeni u realnom vremenu ili skoro u realnom vremenu. Na primer, transakcije u banci, lajkovanje objava, GPS lokacije u vožnji.
3. **Variety (Raznolikost)**
   Podaci dolaze u mnogim različitim oblicima:
   - Struktuirani (tabele, baze podataka)
   - Nestruktuirani (tekst, slike, video, e-mail)
   - Polustruktuirani (JSON, XML, log fajlovi)
4. **Veracity (Tačnost i pouzdanost)**
   Odnosi se na kvalitet i pouzdanost podataka. Loši ili nepotpuni podaci mogu dovesti do pogrešnih odluka. Na primer, analiza korisničkih komentara sa greškama, spam poruka, nekompletnih zapisa.
5. **Value (Vrednost)**
   Najvažniji aspekt – da li iz podataka možemo izvući korisne informacije koje mogu doneti poslovnu vrednost? Na primer, analiza podataka o kupovinama može pomoći firmi da optimizuje marketing i poveća prodaju.

🛠️ Ključni alati i tehnologije za Big Data

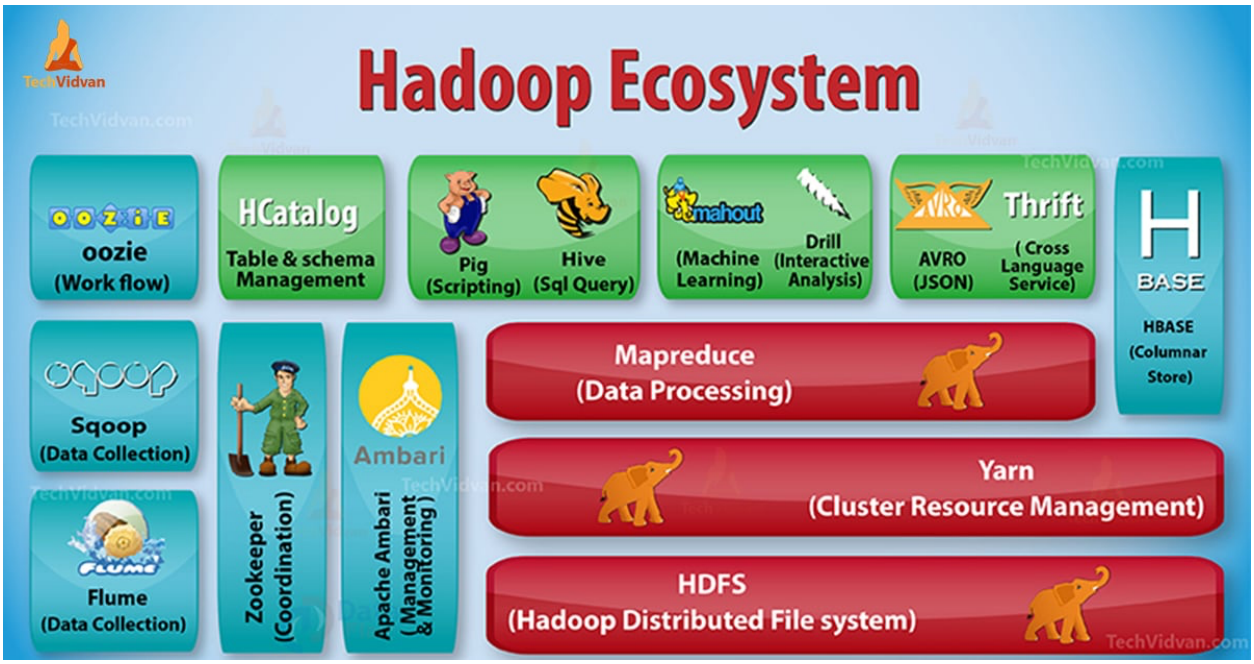
⚙️ 1. Hadoop – Osnovna platforma za Big Data obradu

- **Hadoop** je open-source sistem dizajniran za **distribuiranu obradu i skladištenje ogromnih količina podataka** preko više računara (čvorova).
- Zasnovan je na **MapReduce modelu**, gde se obrada deli na:
  - **Map** (rasparčavanje i lokalna obrada),
  - **Reduce** (spajanje i analiza rezultata).

📌 Ključne komponente:

- **HDFS (Hadoop Distributed File System)** – Omogućava skladištenje velikih fajlova na više računara tako da sistem automatski pravi kopije (redundanciju) i štiti podatke.
- **YARN (Yet Another Resource Negotiator)** – Upravlja resursima sistema i raspoređuje zadatke između čvorova.
- **MapReduce** – Mehanizam za paralelnu obradu podataka, posebno efikasan za batch obradu (veliki setovi podataka obrađeni odjednom).

📍 *Primer:* Hadoop se koristi za paralelnu obradu ogromnih količina log fajlova sa servera koji beleže ponašanje korisnika. Na ovaj način, kompanija može brzo identifikovati najposećenije stranice, uobičajene rute korisnika kroz sajt ili potencijalne greške u sistemu.



⚡ 2. Apache Spark – Brži i fleksibilniji naslednik MapReduce-a

- Apache Spark je moderno rešenje koje koristi **in-memory obradu**, što znači da drži podatke u RAM-u umesto na disku – **što je znatno brže**.
- Podržava i **batch** i **real-time obradu podataka**.

🔧 Mogućnosti:

- **Spark SQL** – upiti nad velikim skupovima podataka koristeći SQL sintaksu.
- **MLlib** – biblioteka za **mašinsko učenje** (klasifikacija, regresija, klasterovanje).
- **Spark Streaming** – obrada podataka u realnom vremenu (npr. analiza tvitova dok se objavljuju).
- **GraphX** – obrada grafova (npr. analiza društvenih mreža).

📍 *Primer:* Analiza podataka sa senzora u realnom vremenu (temperatura, pritisak, kretanje).

🗃️ 3. NoSQL baze podataka – Rad sa fleksibilnim i raznovrsnim podacima

- Za razliku od klasičnih (relacionih) baza, **NoSQL baze** su dizajnirane da lako obrađuju **nestrukturirane i polustrukturirane podatke** – kao što su JSON, XML, slike, video zapisi, itd.
- Idealne su kada podaci nisu uvek istog formata i kada sistem mora brzo da se skalira.

🧩 Glavne vrste NoSQL baza:

- **Dokument-orijentisane** (npr. *MongoDB*) – podatak se čuva kao dokument (npr. JSON), fleksibilno i lako za rad sa različitim strukturama.
- **Kolonijalne** (npr. *Cassandra*) – optimizovane za brzo čitanje/pisanje velikih količina podataka, često korišćene u telekomunikacijama.
- **Key-Value baze** (npr. *Redis*) – izuzetno brze, koriste se za keširanje i upravljanje sesijama.
- **Graf baze** (npr. *Neo4j*) – čuvaju podatke u obliku čvorova i veza, idealne za društvene mreže i preporuke.

📍 *Primer:* E-commerce platforme koriste MongoDB za čuvanje podataka o proizvodima različitih kategorija (koji nemaju istu strukturu), a Redis za brzo keširanje informacija o

🧪 Big Data u Pythonu – šta sve možemo da radimo?

Python je jedan od najpopularnijih jezika za rad sa podacima zbog svojih moćnih biblioteka. Kada je u pitanju **obrada velikih podataka**, Python može da radi sledeće:

🧹 1. Čišćenje i obrada velikih datasetova

📦 **Biblioteke**: pandas, dask, pyarrow, numpy

```
# Učitavanje Pandas biblioteke za rad sa podacima u DataFrame-u
import pandas as pd

# Učitavanje CSV fajla od više GB
chunks = pd.read_csv("big_dataset.csv", chunksize=100000)

for chunk in chunks:
    cleaned = chunk.dropna()  # ukloni prazne redove
    # dodatna obrada...
```

📌 **Napomena**: Kada dataset ne može da stane u RAM, koristi se chunksize ili biblioteka **Dask**, koja omogućava rad sa podacima većim od memorije koristeći paralelnu obradu.

📌 **Modin** je još jedna biblioteka koja koristi više CPU jezgara automatski, ali sa istom pandas sintakso

🔄 2. Distribuirana obrada sa PySpark

📦 **PySpark** je Python API za Apache Spark. Omogućava obradu gigantskih količina podataka raspodeljeno – na više mašina.

```
# Učitavanje neophodne biblioteke i kreiranje Spark instance
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("BigDataExample").getOrCreate()

# Učitavanje velike CSV datoteke
df = spark.read.csv("big_data.csv", header=True, inferSchema=True)

# Primer obrade: filtriraj zaposlene sa platom većom od 5000
df_filtered = df.filter(df.salary > 5000)
df_filtered.show()
```

✅ Spark koristi RAM efikasnije od tradicionalnih pristupa, omogućava analitiku u realnom vremenu, i može da radi sa podacima iz raznih izvora (HDFS, S3, lokalni disk, itd.).

📊 3. Analitika i vizualizacija

📦 **Biblioteke:** matplotlib, seaborn, plotly, dash

Nakon obrade i filtriranja podataka, vizuelizacija pomaže da se prepoznaju obrasci, trendovi i anomalije u podacima. Ona je ključna za poslovno izveštavanje i donošenje odluka.

📌 Možemo crtati grafikone za:

- Trendove u vremenu (npr. porast prodaje)
- Grupisane poređenja (npr. prosečne plate po sektoru)
- Korelacije između atributa (scatter grafici)

```
# Učitanje biblioteka za crtanje dijagrama
import matplotlib.pyplot as plt
import seaborn as sns

# Crtanje histograma za plate
df_pandas.salary.hist(bins=30)
plt.title("Distribucija plata")
plt.xlabel("Plata")
plt.ylabel("Broj zaposlenih")
plt.show()

# Korelacioni dijagram
sns.scatterplot(data=df_pandas, x="age", y="salary", hue="department")
plt.title("Odnos godina i plata")
plt.show()
```

🤖 4. Machine Learning na velikim podacima

📦 **Biblioteke:** MLlib (Spark), Scikit-learn, XGBoost

Mašinsko učenje omogućava da iz ogromnih količina podataka automatski prepoznajemo obrasce i donosimo inteligentne poslovne odluke bez eksplicitnog programiranja pravila. Ova tehnologija koristi se u skoro svim industrijama kako bi se povećala efikasnost i unapredila korisnička iskustva.

Sa dovoljno podataka možemo:

- Predviđati ponašanje korisnika
- Otkrivati anomalije u transakcijama
- Automatizovati preporuke
- Optimizovati lance snabdevanja analizom potražnje u realnom vremenu
- Procenjivati kreditni rizik kod klijenata na osnovu istorije i ponašanja

```
from pyspark.ml.classification import LogisticRegression

# Primer ML modela u Spark-u
from pyspark.ml.feature import VectorAssembler
vec = VectorAssembler(inputCols=["age", "salary"], outputCol="features")
data = vec.transform(df_filtered)
lr = LogisticRegression(labelCol="churned", featuresCol="features")
model = lr.fit(data)
```

🌐 5. Real-time obrada podataka

📦 Alati: **Kafka + Spark Streaming**

Real-time obrada podataka omogućuje analiziranje podataka dok se oni generišu, što je ključno za donošenje brzih odluka. U kombinaciji sa Kafkom i Spark Streamingom, Python postaje moćan alat za obradu podataka u stvarnom vremenu. Kafka pruža efikasan prenos podataka između sistema, dok Spark Streaming obrađuje podatke odmah kako pristižu.

Primeri upotrebe:

- **Notifikacije u realnom vremenu**: Na osnovu korisničkog ponašanja ili sistemskih događaja, odmah šaljemo obaveštenja korisnicima.
- **Live dashboard**: Vizualizacija podataka u stvarnom vremenu za menadžere i analitičare, omogućavajući brzo donošenje odluka.
- **Fraud detection**: Analiza transakcija dok se dešavaju kako bi se odmah otkrile sumnjive aktivnosti i sprečile prevare.

📈 Business Inteligence

Takođe, korišćenjem Big Data tehnologija, preduzeća mogu i:

- **Analizirati ponašanje korisnika**: Personalizacija ponuda i poboljšanje korisničkog iskustva, što vodi većoj lojalnosti.
- **Automatizovati procese**: Preporuke proizvoda ili usluga na osnovu podataka, čime se povećava efikasnost.
- **Prepoznavanje obrazaca i predviđanje trendova**: Prediktivna analitika za bolju strategiju i donošenje pravovremenih odluka.
- **Poboljšanje donošenja odluka**: Brža i tačnija poslovna odluka uz aktuelne podatke, što smanjuje rizik.