Fakultet tehničkih nauka

Univerzitet u Novom Sadu


# Skladištenje velike količine frekventnih i nestruktuiranih podataka

Seminarski rad iz predmeta

Big Data u infrastrukturnim sistemima

MENTOR :
Prof. dr. Aleksandar Kupusinac

STUDENT :
Nikola Miljković E5-3/2023

ASISTENT :
Bojana Samardžić

Novi Sad, decembar, 2023.

## SADRŽAJ

# UVOD

Iako u doslovnom prevodu znači „velika količina podataka“, *Big Data* predstavlja složeniju pojavu. Ukoliko bismo ovaj doslovni prevod uzeli kao definiciju, napravili bismo grešku, s obzirom na to da *Big Data* nije samo tehnologija već da u sebi sadrži određeni inovativni potencijal.

U okruženju informacionih tehnologija koje se stalno razvija, pojava tehnologija velikih podataka dovela je do revolucionarnih promena u načinu na koji organizacije prikupljaju, čuvaju i analiziraju podatke. Ovaj rad se fokusira na izazove i strategije skladištenja velikih količina frekventnih i nestruktuiranih podataka u kontekstu *Big Data* tehnologije.

Veliki podaci u infrastrukturnim sistemima se odnose na upotrebu tehnologija i metoda za prikupljanje, skladištenje, obradu i analizu velikih količina podataka unutar infrastrukture organizacije. Infrastrukturni sistemi obično uključuju mrežnu opremu, serverske resurse, skladištenje podataka, servise aplikacija i sve ostalo što podržava IT operacije organizacije.

Veliki podaci u infrastrukturnim sistemima igraju ključnu ulogu u donošenju informisanih odluka, poboljšanju performansi, povećanju efikasnosti i pružanju konkurentske prednosti organizacijama. Važno je uskladiti tehnologiju i strategije velikih podataka sa specifičnim potrebama i ciljevima organizacije.

U eri *Big Data* tehnologija, skladištenje ogromne količine podataka vezanih za infrastrukturne sisteme postaju ključni faktor. Razumevanje izazova i primena odgovarajućih strategija je od ključnog značaja za postizanje punog potencijala analitike velikih podataka u organizacijama. U narednim delovima rada istražićemo detalje ovih strategija i njihov doprinos uspešnom upravljanju velikim količinama frekventnih i nestruktuiranih podataka.

# Big Data

## Definicija Big Data

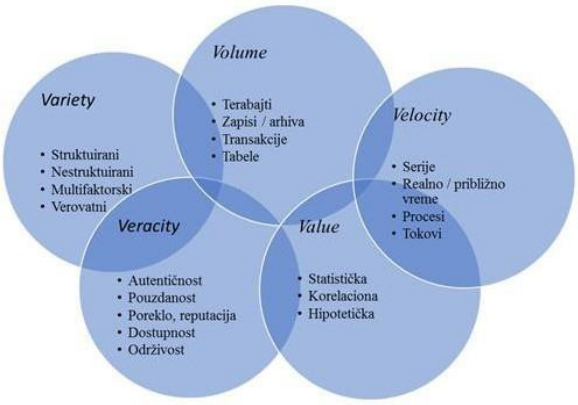
**Big Data** je pojam koji označava velike i kompleksne setove podataka, kod kojih tradicionalne aplikacije za obradu podataka nisu primenljive. Te skupove podataka karakterišu raznovrsnost formata, velike brzine obrade i pristupa, i veliki obim informacija. Izazovi uključuju projektovanje i realizaciju infrastrukture i servisa za skladištenje velikih količina podataka, njihovu pretragu, analizu, deljenje i vizuelizaciju. Termin *big data* se često odnosi na upotrebu predikativne analitike ili drugih naprednih metoda za izdvajanje vrednosti iz podataka, a ne samo na određenu veličinu skupa podataka.

*Big data* koncept karakteriše prelazak sa relacionih na **nerelacione baze podataka**, kao što su na primer Guglov Bigtable [1]i Amazonov Dynamo [2]. Jedno od rešenja za infrastrukturu *Big Data* je *Hadoop*[3], softver otvorenog koda. *Big Data* pruža mogućnost obrade podataka u realnom vremenu, a pretraga se vrši korišćenjem *Map reduce (link)* algoritma. Na primer, rezultati pretrage u Gugl pretraživaču se dobijaju u milisekundama upravo zahvaljujući ovim tehnologijama.

Ovaj koncept obuhvata tri ključna svojstva poznata kao „**3V“** (**Veliki obim – Volume**, **Visoka brzina – Velocity** i **Velika raznovrsnost – Variety**). Kasnije su dodata još dva svojstva – **Verodostojnost – Veracity** i **Vrednost – Value** formirajući „**5V“** model :

- **Veliki obim (Volume)** - *Big Data* karakteriše ogroman obim podataka. Tradicionalni sistemi za skladištenje i analizu podataka često nisu dovoljni za rukovanje takvim količinama informacija.
- **Visoka brzina (Velocity)** - Podaci u *Big Data* okruženju generišu se i pristižu brzo, često u stvarnom vremenu. Ovo svojstvo naglašava potrebu za brzom analizom i obradom podataka.
- **Velika raznovrsnost (Variety)** - *Big Data* uključuje raznolike vrste podataka, uključujući struktuirane podatke (npr. podaci iz baza podataka), nestruktuirane podatke (npr. tekstualni dokumenti, slike, videozapisi) i polustruktuirane podatke (npr. *XML* datoteke).
- **Verodostojnost (Veracity)** - Odnosi se na pouzdanost i tačnost podataka. *Big data* podaci često potiču iz različitih izvora, što može dovesti do pitanja o tačnosti, integritetu i verodostojnosti podataka
- **Vrednost (Value)** - Konačni cilj obrade *Big Data* je izvlačenje vrednosti iz informacija, što uključuje dobijanje novih uvida, informacija ili donošenje informisanih poslovnih odluka.

[1] Google Bigtable - distribuirana, visoko performantna baza podataka koja se koristi za obradu ogromnih količina podataka u stvarnom vremenu
[2] Amazon DynamoDB - potpuno upravljana, visoko dostupna i skalabilna NoSQL baza podataka koju pruža AWS
[3] Hadoop - open-source softverski alat koji pruža distribuirano skladištenje i obradu velikih skupova podataka na klasterima




Slika 1.1. 5V model

## Veliki obim (Volume)

Veličina podataka, ili "Volume" jedan je od ključnih aspekata *big data*. Ovaj aspekt se odnosi na ogromne količine podataka koje često prevazilaze kapacitete tradicionalnih sistema za upravljanje bazama podataka. Nekoliko ključnih tačaka u vezi sa veličinom podataka u *big data* kontekstu:

- Terabajti, petabajti i exabajti
  - *Big Data* često uključuje rad s podacima na terabajtima, petabajtima ili čak exabajtima. Ove količine podataka mogu poticati iz različitih izvora, uključujući senzore, društvene mreže, e-trgovinu, medicinske podatke i još mnogo toga
- Rastući Trend
  - Količina podataka nastavlja eksponencijalno rasti s vremenom, a taj rast se ubrzava. Razvoj interneta stvari (IoT), povećana upotreba pametnih uređaja, senzora i društvenih medija doprinose ovom neprekidnom porastu veličine podataka
- Analiza u Realnom Vremenu
  - Povećanje brzine analize u stvarnom vremenu takođe dovodi do potrebe za brzim pristupom velikim količinama podataka
- Dokumenti, Slike, i Video
  - Nestruktuirani podaci, poput tekstualnih dokumenata, slika i video snimaka, često imaju značajan udeo u ukupnoj veličini podataka
- Skladištenje podataka
  - Za skladištenje takvih velikih količina podataka, koriste se posebni sistemi, uključujući distribuirane sisteme za skladištenje poput *Hadoop Distributed File System* (HDFS)[4], objektno usmerene sisteme kao *Amazon S3*[5], te tradicionalne baze podataka optimizovane za rad s velikim setovima podataka

Rukovanje velikom količinom podataka zahteva prilagođene tehnologije i alate koji su u stanju efikasno upravljati i analizirati podatke ovakvih razmera. To uključuje distribuirane sisteme za obradu podataka, napredne algoritme za kompresiju, kao i optimizovane tehnike skladištenja podataka.

[4] HDFS je distribuirani sistem za čuvanje podataka koji je deo Apache Hadoop
[5] Amazon S3 je popularna usluga za čuvanje podataka na cloud-u koju pruža AWS-a

| | |
|---|---|
| 1 byte<br>Kilobyte<br>≈ 1000 ($10^3$) bytes<br>Megabyte<br>≈ 1000000 ($10^6$) bytes | |
| Gigabyte<br>≈ 1000000000 ($10^9$) bytes | 25 gigabajta: podaci koje *Ford Fusion Energy plug-in hibrid* analizira u toku jednog sata<br>60 gigabajta: podaci koje *Google self-driving* automobil sakupi u toku jednog sata<br>140 gigabajta: podaci koje *Nokia Here Maps* aplikacija sakupi u toku jednog dana |
| Terabyte<br>≈ 1000000000000 ($10^{12}$) bytes | 30 gigabajta: podaci koje *Boing 777* prikupi u toku jednog prekookeanskog leta |
| Petabyte<br>≈ 1000000000000000 ($10^{15}$) bytes | Nekoliko petabajta: podaci o saobraćaju skladišteni na *Inrix* platformi u svrhe analize saobraćaja za npr. *Google Traffic* |
| Exabyte<br>≈ 1000000000000000000 ($10^{18}$) bytes | |
| Zettabyte<br>≈ 1000000000000000000000 ($10^{21}$) bytes | 1 zetabajt: ukupna količina vizuelnih informacija koje je ljudsko oko poslalo kao signal mozgu prikupljenih računajući sve ljude na svetu u toku jednog dana u 2013. godini<br>4,4 zetabajta: procenjena veličina digitalnog univerzuma u 2013. godini |
| Yottabyte<br>≈ 1000000000000000000000000 ($10^{24}$) bytes | |

Tabela 1.1 Prikaz podataka za poređenje

## Visoka brzina (Velocity)

Konstatovano je da je volumen podataka koji treba da se prikupe, uskladište i analiziraju veliki (u skladu sa definicijom *Big Data*). Sledeća dimenzija je brzina koja je potrebna da se ove aktivnosti sprovedu, pri tome se pod brzinom prvenstveno podrazumeva vreme koje je neophodno da se dobije krajnji rezultat: preporuka za akciju. Jasni su razlozi zbog kojih je brzina imperativ u savremenom poslovanju:

- Prvenstveno zbog konkurentske utakmice: neophodno je identifikovati problem, prepoznati šansu pre drugih. Nekada su u pitanju sekunde, čak i milisekunde.
- Podaci imaju veoma kratak rok trajanja. Brzo zastarevaju i nepredstavljaju više konkurentsku prednost.

To samo potvrđuje da se podaci moraju prikupljati, obrađivati i analizirati praktično u realnom vremenu kako bi se što pre stekao uvid u suštinu podataka. Ovo ujedno znači da se menja paradigma istraživanja: podaci se obrađuju i analiziraju od momenta kada počnu da se prikupljaju. Iz toga sledi da se proces prikupljanja podataka nikada ne završava, već se obrada i analiza vrše iznova i iznova.

## Velika raznovrsnost (Variety)

Podaci po svojoj raznovrsnosti se mogu svrstati u sledeće kategorije: **Nestruktuirani, Kvazistruktuirani, Polustruktuirani i Struktuirani**.


Slika 1.2. Prikaz podele podataka prema struktuiranosti

**Nestruktuirani podaci** su u osnovi informacije koje ili nemaju unapred definisani model podataka i/ili se dobro ne uklapaju u tradicionalnu bazu podataka:

- tekst, PDF dokument, video, slike, audio, geoprostorni podaci, internet podaci, *click streams*, [6]log fajlovi

U određenom broju slučajeva ovi podaci se mogu posmatrati kao kategorijski ili nominalni podaci (string) što je korisno sa stanovišta njihove dalje obrade i analize.

**Kvazistruktuirani** podaci predstavljaju tekstualne podatke koji su dati u nestandardnom formatu i kao takvi se mogu formatirati, što zahteva dosta znanja, alata i vremena. Tipičan primer ovakve vrste prodataka predstavljaju *web clickstream* podaci koji mogu sadržati određene nedoslednosti, pre svega u formatu, pa i sadržaju.

**Polustruktuirani** podaci se koriste za opisivanje struktuiranih podataka koji se ne uklapaju u formalnu strukturu modela podataka. Ovi podaci ne sadrže oznake koje razdvajaju semantičke elemente, nemaju zajedničku strukturu, poseduju sposobnost sprovođenja hijerarhije unutar podataka i podrazumevaju više načina predstvaljanja iste vrste podataka. Za predstavljanje polustruktuiranih podataka koristi se *XML* (eXtensible Markup Language) programski jezik koji je sličan *HTML-u* a razvijen je od strane *W3C* (World Wide Web Consortium) u cilju prevazilaženja ograničenja *HTML*-a (Internet 6). *XML* je zamišljen kao programski jezik za opisivanje podataka – podrazumevajući opis podataka, a ne njihov izgled.

[6] Clickstream podaci se odnose na informacije o redosledu interakcija koje korisnici imaju sa veb-sajtom ili aplikacijom tokom određenog vremenskog perioda.

**Struktuirani** podaci imaju jasno definisani tip, format i strukturu. Ova vrsta podataka je najčešće smeštena u kompanijskim bazama i/ili skladištima podataka. U poređenju sa "tradicionalnom" statističkom metodologijom, struktuirani podaci se mogu svrstati u kategoriju metričkih ili numeričkih varijabli čija je obrada, analiza i interpretacija veoma precizno definisana i relativno jednostavna. Problem može predstavljati skladištenje i čuvanje velike količine ovakvih podataka kako bi se oni koristili u analitičke svrhe.

Nestruktuirani, kvazistruktuirani i polustruktuirani podaci u osnovi se mogu posmatrati kao nominalne varijable koje zahtevaju različite i mnogo zahtevnije metode obrade i analize nego što je slučaj kod struktuiranih varijabli. U praksi se najčešće primenjuju tehnike napredne statističke analize (klaster analiza), metode veštačke inteligencije (mašinsko učenje), data mining itd.

Važno je napomenuti da je evidentan rast svih vrsta podataka, s tom razlikom da generisanje struktuiranih podataka prati linearan trend, za razliku od nestruktuiranih podataka čiji je rast eksponencijalan. Tradicionalne IT infrastrukture i analitičke platforme ne mogu da prate ovoliku raznolikost.

## Verodostojnost (Veracity)

S obzirom na to da podaci postoje u različitim oblicima i prikupljaju se sa mnoštva izvora – kontrolisanje tačnosti, verodostojnosti i pouzdanosti podataka predstavlja izazov za *Big Data* naučnike i istraživače. Društvene mreže, na primer, uvele su opciju *hashtags (#)* i podaci se prikupljaju u vidu skraćenica, neki podaci sadrže i greške u kucanju ili kolokvijalni govor. *Big Data* analitika omogućava rad i sa ovakvom strukturom podataka, a najčešće količina dostupnih podataka nadomesti nedostatke kvaliteta i tačnosti (Marr, 2015) i pri tome se koriste sofisticirane matematičko-statističke metode i tehnike zasnovane npr. na teoriji fazi skupova i fazi logike.

## Vrednost (Value)

U 5V modelu aspekt vrednost označava sposobnost izvlačenja korisnih informacija i smislenih uvida iz ogromnih količina podataka. Ovaj aspekt ističe značajnu svrhu i korist od rada s *big data*, naglašavajući da prikupljanje i analiza podataka ne bi trebali biti samo ciljevi sami po sebi, već bi trebali doprineti stvaranju vrednosti za organizaciju ili pojedinca. Vrednost u 5V modelu ukazuje na to da je akcenat na kvalitetu i korisnosti informacija koje se izvlače iz *big data* skupova, umesto samo na kvantitetu podataka. Ovo podvlači važnost postavljanja ciljeva analize podataka koji direktno doprinose postizanju poslovnih ciljeva i pružaju konkretnu vrednost organizaciji.

## Nestruktuirani podaci

Nestruktuirani podaci su informacije koje nisu uređene prema unapred postavljenom modelu podataka ili šemi, pa se stoga ne mogu čuvati u tradicionalnoj relacionoj bazi podataka. Tekst i multimedija su dva uobičajena tipa nestrukturisanog sadržaja. Mnogi poslovni dokumenti su nestruktuirani, kao i poruke e-pošte, video snimci, fotografije, veb stranice i audio datoteke.

Od 80% do 90% podataka koje generišu i prikupljaju organizacije su nestruktuirane, a njihov obim brzo raste - mnogo puta brže od stope rasta struktuiranih baza podataka. Nestruktuirana skladišta podataka sadrže obilje informacija koje se mogu koristiti za usmeravanje poslovnih odluka. Međutim, istorijski je bilo veoma teško analizirati nestruktuirane podatke. Uz pomoć veštačke inteligencije i mašinskog učenja, pojavljuju se novi softverski alati koji mogu da pretražuju kroz ogromne količine da bi otkrili korisnu i efikasnu poslovnu inteligenciju.

Nestruktuirani podaci su podaci kod kojih svaka instanca u skupu podataka može imati sopstvenu unutrašnju strukturu, a ova struktura nije u svakom slučaju ista. Ova vrsta podataka mnogo je češća od struktuiranih podataka. Varijacije u strukturi različitih elemenata znače da je nestrukturirane podatke teško analizirati u sirovom obliku. Korišćenjem veštačkih tehnika inteligencije, digitalne obrade signala i kompjuterskog vida, iz nestruktuiranih podataka je često moguće izdvojiti struktuirane podatke. Međutim, implementacija i testiranje ovih procesa transformacija podataka je skupa i dugotrajna. Primeri nestruktuiranih podataka su zbirke tekstove (knjige, poruke, e-mail) i zbirke zvuka, slike, muzike, videa I multimedijalne datoteke.

## Nestruktuirani podaci i struktuirani podaci

Uzmimo prvo strukturirane podatke: obično se čuvaju u relacionoj bazi podataka ili RDBMS[7]-u, a ponekad se nazivaju relacionim podacima. Može se lako mapirati u određena polja - na primer, polja za poštanske brojeve, brojeve telefona i kreditne kartice. Podaci koji su u skladu sa strukturom RDBMS-a su laki za pretraživanje, kako sa upitima koje definiše čovek, tako i sa softverom. Struktuirani podaci su jednostavni za korišćenje I trenutno upotrebljivi, međutim glavni problem je nedostatak fleksibilnosti - struktuirani podaci zahtevaju od korisnika da unapred kreiraju definicije podataka šeme. Strukturu je teško promeniti tokom vremena, a pošto postoji fiksna, unapred definisana struktura, podaci se mogu koristiti samo za njihovu namenu. Ovo ograničava slučajeve upotrebe koji se mogu opsluživati struktuiranim podacima. Čuvanje struktuiranih podataka zna biti dosta skupo - struktuirani podaci se često čuvaju u skladištima podataka, koja mogu skladištiti strukturirane podatke u velikom obimu i omogućiti brz pristup korisničkim upitima. Skladište podataka je složen sistem koji zahteva značajne resurse za rad, razvoj i održavanje. Kako organizacije rastu, broj baza podataka, tabela i polja raste eksponencijalno postaje teško upravljati struktuiranim podacima i uobičajeno je da postoje preklapanja između skupova podataka, suvišnih podataka i zastarelih podataka ili podataka niskog kvaliteta.

[7] RDBMS – Sistem za upravljanje relacionim bazama podataka

Nasuprot tome, nestruktuirani podaci se ne uklapaju u ove vrste unapred definisanih modela podataka. Ne može se čuvati u RDBMS-u. A pošto dolazi u toliko formata, pravi je izazov za konvencionalni softver da unese, obradi i analizira. Jednostavna pretraživanja sadržaja mogu se preduzeti preko tekstualnih nestruktuiranih podataka uz pomoć pravih alata. Nestruktuirani podaci mogu da se čuvaju u svom izvornom formatu dok ne budu potrebni I čuvanje ovih podataka zna biti po mnogo nižoj ceni u odnosnu na strukturirane podatke. Glavni nedostatak ovakih podataka je to što zahtevaju naprednu analitiku - obično postoji potreba za veštinama nauke o podacima i naprednim algoritmima za analizu i izvlačenje uvida iz nestruktuiranih podataka. To takođe znači da nije korisno za većinu poslovnih korisnika koji nemaju veštine za obavljanje napredne analitike. Nestruktuirani podaci zahtevaju namenske alate - preuzimanje i obrada nestruktuiranih podataka zahteva specijalizovane alate i stručnost.

Osnovne razlike između struktuiranih i nestruktuiranih podataka:

1. **Format** - Obično su struktuirani podaci u obliku brojeva i teksta, predstavljeni u standardizovanim, čitljivim formatima. XML i CSV su najpopularniji formati. U modelima struktuiranih podataka format podataka je unapred određen. S druge strane, nestruktuirani podaci često dolaze u različitim oblicima i veličinama. Ne odgovara unapred definisanom modelu podataka i ostaje u izvornim (originalnim) formatima. Primeri uključuju video (tj. VMV, MPV) i audio datoteke (tj. MP3, VAV)

2. **Model podataka** - Struktuirani podaci prate unapred definisani relacioni model podataka koji opisuje odnos elemenata podataka. Nestruktuirani podaci nemaju postavljeni model podataka, ali mogu imati skrivenu strukturu.

3. **Skladištenje** - Organizacije čuvaju strukturirane podatke u relacionim bazama podataka. Skladišta podataka pomažu u centralizaciji velikih količina uskladištenih struktuiranih podataka iz različitih baza podataka. Organizacije skladište nestrukturirane podatke u sirovim formatima, a ne u bazama podataka. Data Lakes mogu da skladište velike količine nestruktuiranih podataka

4. **Tip baze podataka -** Struktuirani podaci se obično nalaze u relacionoj bazi podataka, raspoređeni u tabele sa redovima i kolonama. Oznake određuju tipove podataka. Šema tabele se sastoji od kolone podataka i konfiguracije tipa. Relacione baze podataka obrađuju podatke koristeći SQL, laku sintaksu koju korisnici mogu čitati. Nestruktuirani podaci se često nalaze u nerelacionoj (NoSQL) bazi podataka. Ovaj tip baze podataka skladišti više modela podataka bez tabela - ovo je obično dokument, baza podataka širokih kolona, grafikona i baza podataka ključnog volumena. Može da obrađuje velike količine podataka i podnosi velika opterećenja. NoSQL baza podataka sadrži kolekcije dokumenata koji liče na redove, ali ne koriste tabelarnu šemu, tako da u istoj kolekciji mogu biti različiti tipovi podataka. Nerelacioni model omogućava brže upite.

5. **Mogućnost pretraživanja i lakoća korišćenja -** Struktuirane podatke je obično lakše pretraživati i koristiti, dok nestruktuirani podaci uključuju složeniju pretragu i analizu. Nestruktuirani podaci zahtevaju obradu da bi se razumeli, kao što je slaganje pre nego što se stave u relacionu bazu podataka. Struktuirani podaci su stariji, tako da je dostupno više analitičkih alata. Standardna rešenja za rudarenje podataka ne mogu da rukuju nestruktuiranim podacima.

6. **Kvantitativno naspram kvalitativnog -** Struktuirani podaci su kvantitativni, što znači da imaju prebrojive elemente. Lakše je analizirati klasifikovanjem stavki na osnovu zajedničkih karakteristika, istraživanjem odnosa između varijabli ili grupisanjem podataka u grupe zasnovane na atributima. Nestruktuirani podaci su kvalitativni, što znači da su informacije koje sadrže subjektivne, a tradicionalni analitički alati i metode ne mogu da se nose sa njima. Na primer, povratne informacije kupaca na društvenim medijima mogu da generišu podatke u tekstualnom obliku, zahtevajući naprednu analitiku da ih obradi. Tehnike uključuju razdvajanje i slaganje volumena podataka u logička grupisanja, rudarenje podataka i detekciju obrazaca.

## Koji su primeri nestruktuiranih podataka?

Nestruktuirane podatke mogu kreirati ljudi ili ih generisati mašine. Evo nekoliko primera koje su generisali ljudi:

- E-pošta: Polja za poruke e-pošte su nestrukturirana i ne mogu se raščlaniti tradicionalnim analitičkim alatima. Međutim, metapodaci e-pošte daju im određenu strukturu i objašnjavaju zašto se e-pošta ponekad smatra polustruktuiranim podacima.
- Tekstualne datoteke: Ova kategorija uključuje dokumente za obradu teksta, tabele, prezentacije, e-poštu i datoteke evidencije.
- Društveni mediji i veb-sajtovi: podaci sa društvenih mreža kao što su Tvitter, LinkedIn i Facebook, i veb-sajtova kao što su Instagram, sajtovi za deljenje fotografija i YouTube.
- Mobilni i komunikacioni podaci: Za ovu kategoriju ne tražite dalje od tekstualnih poruka, telefonskih snimaka, softvera za saradnju, ćaskanja i trenutnih poruka.
- Mediji: Ovi podaci uključuju digitalne fotografije, audio i video datoteke.

Evo nekoliko primera nestruktuiranih podataka koje generišu mašine:

- Naučni podaci: Ovo uključuje istraživanja nafte i gasa, istraživanje svemira, seizmičke slike i atmosferske podatke.
- Digitalni nadzor: Ova kategorija sadrži podatke kao što su izviđačke fotografije i video snimci.
- Satelitski snimci: Ovi podaci obuhvataju vremenske podatke, oblike zemljišta i vojna kretanja.

## Kako se skladište nestruktuirani podaci?

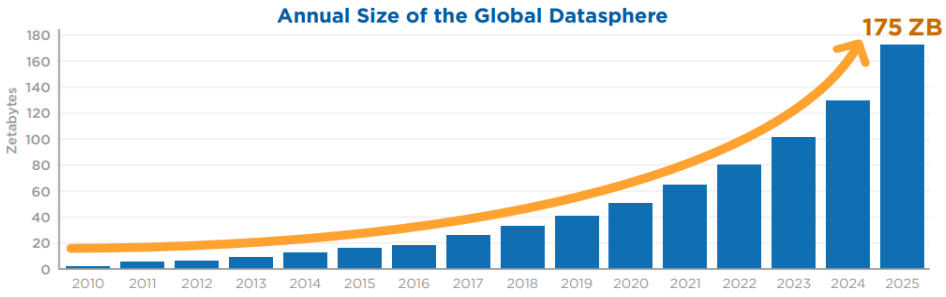
Nestruktuirani tipovi podataka zapravo mogu imati unutrašnje strukturne elemente. Smatraju se "nestruktuiranim" jer njihove informacije nisu pogodne za formatiranje tabele koje zahteva relaciona baza podataka. Kao što je ranije pomenuto, nestruktuirani podaci mogu biti tekstualni ili netekstualni (kao što su audio, video i slike) i generisani od strane ljudi ili mašina. Nerelacione baze podataka kao što je MongoDB su preferirani izbor za skladištenje mnogih vrsta nestruktuiranih podataka. Nestruktuirani podaci se mogu čuvati na više načina: u aplikacijama, NoSQL[8] (nerelacionim) bazama podataka, data lakes i data warehouses. Platforme kao što je MongoDB Atlas su posebno pogodne za smeštaj, upravljanje i korišćenje nestruktuiranih podataka.

## Za šta se koriste nestruktuirani podaci?

Jednostavne pretrage sadržaja mogu se izvršiti na tekstualnim nestruktuiranim podacima. Tradicionalni alati za analitiku su optimizovani za visoko strukturirane relacione podatke, tako da su od male koristi za nestrukturirane izvore kao što su bogati mediji, interakcije sa korisnicima i podaci društvenih medija.

Veliki podaci i nestruktuirani podaci često idu zajedno procenjuje se da je 90% ovih izuzetno velikih skupova podataka nestrukturirano. Novi alati su nedavno postali dostupni za analizu ovih i drugih nestruktuiranih izvora. Pokrenute AI i mašinskim učenjem, takve platforme funkcionišu brzinom skoro u realnom vremenu i obrazuju se na osnovu obrazaca i uvida koje otkrivaju. Ovi sistemi se koriste protiv velikih nestruktuiranih skupova podataka kako bi omogućili aplikacije koje nikada ranije nisu bile moguće kao:

- Analiziranje komunikacija radi usklađenosti sa propisima
- Praćenje i analiza razgovora i interakcija korisnika na društvenim mrežama
- Sticanje pouzdanog uvida u široko rasprostranjeno ponašanje i preferencije kupaca




Slika 1.3 Rast količine podataka u vremenskom razdoblju od 2010. do 2025. godine

[8] NoSQL (Not Only SQL) - je vrsta baze podataka koja se razlikuje od relacionih baza podataka. Ova vrsta baze podataka je projektovana za efikasno upravljanje velikim volumenima podataka.

# Skladištenje podataka

Da biste donosili poslovne odluke na osnovu informacija, potrebni su vam **podaci**. Sposobnost vaše organizacije da prikuplja podatke iz različitih izvora i da iz njih obezbedi smislene uvide može pomoći u povećanju poslovne vrednosti. Da biste postavili osnovu za uspešnu strategiju podataka, počinjete tako što ćete odrediti kako će se vaši podaci čuvati, analizirati i ispitivati. Ovde ćemo pričati o različitim opcijama skladištenja, razlikama između njih i kako funkcionišu. Takođe ćemo pokriti koju opciju da odaberete na osnovu vaše strategije podataka, infrastrukture i ciljeva.

## Baze podataka

Baza podataka je sistematska kolekcija srodnih informacija ili podataka koji se čuvaju elektronski na način kome se lako pristupa, preuzima, njime se upravlja i ažurira. Baze podataka pomažu da se olakša skladištenje, preuzimanje, modifikacija i brisanje podataka. Oni obično prikupljaju informacije o ljudima, mestima ili stvarima. Postoje različite vrste namenski izgrađenih baza podataka, koje se obično biraju da podrže različite modele podataka na osnovu organizacionog pristupa koji je izabran. One mogu uključivati relacione, NoSQL, orijentisane objektima, in-memory, GIS, vremenske serije, široke kolone i ledger databases. Za kompanije je važno da definišu svoju strategiju podataka koja je u skladu sa njihovim poslovanjem. Odluka o tome koju bazu podataka koristiti je takođe odluka o dizajnu/modeliranju. To može biti jednostavno kao procena vaših zahteva za transakcijskim podacima i korišćenje toga kao okvira za izbor prave baze podataka. Baze podataka se mogu koristiti za različite slučajeve korišćenja kao što su bankarske transakcije, upravljanje sesijama, otkrivanje prevara, veb aplikacije sa velikim prometom, sistemi e-trgovine itd. Na primer, platforme društvenih medija koriste baze podataka dokumenata i grafikona za skladištenje korisničkih informacija kao što su imena, adrese e-pošte i ponašanje korisnika. Podaci se zatim koriste za preporuku sadržaja korisnicima i poboljšanje korisničkog iskustva. SQL baze podataka su savršeno prikladne za skladištenje struktuiranih podataka, dok su NoSQL baze podataka najbolje za rad sa nestruktuiranim i polustruktuiranim podacima.

Preduzeća se oslanjaju na baze podataka i prikupljene podatke kako bi donosili informisane odluke koje bi mogle biti ključne za profitabilnost. Baze podataka takođe pomažu organizacijama da obezbede prilagođeni skup proizvoda i usluga kupcima, poboljšavajući njihova iskustva i povećavajući zadržavanje. Na primer, organizacije koriste svoje baze podataka za poboljšanje poslovnih procesa prikupljanjem podataka kao što su prodaja, obrada porudžbina i korisnička podrška. Preduzeća analiziraju te podatke kako bi poboljšala ove procese, proširila svoje poslovanje i povećala prihod. Drugi primer je da se obezbede lične zdravstvene informacije: zdravstveni radnici koriste baze podataka za bezbedno skladištenje ličnih zdravstvenih podataka kako bi poboljšali negu pacijenata. Kada razmišljamo o izboru prave baze podataka za naše aplikacije, to je uglavnom izbor između dva tipa sistema za obradu podataka: analitička obrada na mreži (OLAP) i obrada onlajn transakcija (OLTP). Glavna razlika između ova dva je ta što se OLAP koristi za sticanje vrednih uvida dok je OLTP čisto operativan.

Analitička obrada na mreži (OLAP) se obično koristi za obavljanje analize velikih skupova podataka pri velikim brzinama, dok obrada transakcija na mreži (OLTP) prikuplja i održava podatke o transakcijama u bazi podataka. Odabir pravog sistema za vašu aplikaciju, uglavnom zavisi od vaših ciljeva. OLAP može pomoći u otključavanju vrednosti iz velikih skupova podataka. U međuvremenu, u situaciji kada treba da upravljate dnevnim transakcijama možete da koristite OLTP da biste pomogli u upravljanju velikim brojem transakcija u sekundi.

## Skladišta podataka (Data Warehouses)

Skladište podataka je centralno skladište informacija koje se mogu analizirati da bi se donele bolje informisane odluke. Podaci mogu da pristižu u skladište podataka iz transakcionih sistema, relacionih i drugih izvora. Skladište podataka centralizuje i konsoliduje velike količine podataka iz više izvora. Zbog ovih mogućnosti, skladište podataka se može smatrati „jedinstvenim izvorom istine organizacije". Struktuirani podaci iz više izvora se često čuvaju u skladištima podataka.

Skladište podataka može da sadrži više baza podataka. Unutar svake baze podataka, podaci su organizovani u tabele i kolone. Unutar svake kolone možete definisati opis podataka, kao što je ceo broj, polje podataka ili string. Tabele se mogu organizovati unutar šema, koje možete zamisliti kao fascikle. Kada se podaci unesu, oni se čuvaju u različitim tabelama koje opisuje šema. Alati za upite koriste šemu da bi odredili kojim tabelama podataka treba pristupiti i analizirati ih.

Neke prednosti korišćenja skladišta podataka uključuju: informisano donošenje odluka, konsolidovane podatke iz mnogih izvora, analizu istorijskih podataka, kvalitet podataka, doslednost i tačnost i odvajanje analitičke obrade od transakcionih baza podataka, što poboljšava performanse oba sistema.

## Jezero podataka (Data Lakes)

Jezero podataka skladišti strukturirane, polustrukturirane i nestrukturirane podatke, podržavajući mogućnost skladištenja neobrađenih podataka iz svih izvora bez potrebe za njihovom obradom ili transformacijom u tom trenutku u bilo kojoj skali. Tipično, primarna svrha jezera podataka je analiza podataka kako bi se stekao uvid. Objedinjavanjem svih vaših podataka na jednoj lokaciji, u mogućnosti ste da obavljate analitiku kao što su transformacije podataka i mašinsko učenje preko izvora podataka kao što su datoteke evidencije, podaci iz tokova klikova i društvenih medija uskladištenih u jezeru podataka. Još jedna prednost skladištenja podataka u jezeru podataka je mogućnost skladištenja podataka u različitim formatima uključujući *JSON, ORC* i *Parquet*. Mogućnost da se iskoristi više podataka, iz više izvora, za manje vremena, i osnaživanje korisnika iz više linija poslovanja da sarađuju i analiziraju podatke na različite načine, vodi ka boljem i bržem donošenju odluka. Primeri gde jezera podataka mogu da dodaju vrednost uključuju: pomoć u interakciji sa klijentima tako što će se kombinovati podaci za osnaživanje poslovanja, povećanje operativne efikasnosti što zauzvrat pomaže u smanjenju operativnih troškova i povećanju kvaliteta.

Nestruktuirani podaci se mogu čuvati u različitim tipovima rešenja za skladištenje podataka, uključujući jezera podataka, baze podataka i skladišta podataka. Izbor mesta za skladištenje nestruktuiranih podataka zavisi od specifičnih potreba, slučajeva upotrebe i karakteristika podataka.

Ukratko, dok su jezera podataka posebno dizajnirana za rukovanje nestruktuiranim i sirovim podacima, baze podataka i skladišta podataka takođe mogu da prime nestrukturirane podatke, posebno sa pojavom NoSQL baza podataka i evolucijom tehnologija skladištenja podataka. Izbor između ovih rešenja za skladištenje zavisi od faktora kao što su priroda podataka, zahtevi za performansama, potrebe skalabilnosti i ukupni ciljevi obrade i analize podataka organizacije.

| Parameters | Data Lake | Data Warehouse |
|---|---|---|
| Data type | Raw (all types, no matter source of structure) | Processed (data stored according to metrics and attributes) |
| Data purpose | To be determined | Currently being used |
| Process | Extract Load Transform (ELT) | Extract Transform Load (ETL) |
| Schema position | After data storage, to offer agility and easy data capture | Before data storage, to offer security and high performance |
| Users | Data scientists, those who need in-depth analysis and tools (such as predictive modeling) to understand it | Business professionals, those who need it for operations |
| Accessibility | Accessible and easy to update | Complicated to make changes |
| History | Relatively new for big data | The concept has been around for decades |

Tabela 2.1 Razlike između data lake I data warehouse

# ZAKLJUČAK

U zaključku, skladištenje velikih količina čestih i nestruktuiranih podataka je višestruki izazov koji zahteva inovativna rešenja u oblasti informacionih tehnologija. Tokom ovog seminarskog rada, ušli smo u različite strategije i tehnologije koje se koriste za efikasno upravljanje takvim količinama podataka u eri velikih podataka.

Jezera podataka se pojavljuju kao ključne komponente u ovom pejzažu, obezbeđujući centralizovano skladište sposobno da prihvati različite formate i da rukuje ogromnim količinama podataka u njihovom sirovom, neobrađenom stanju. Fleksibilnost svojstvena jezerima podataka omogućava organizacijama da skladište i naknadno obrađuju česte i nestrukturirane podatke, postavljajući osnovu za naprednu analitiku i izvlačenje uvida.

NoSQL baze podataka, skladište objekata i distribuirani sistemi datoteka značajno doprinose okruženju skladištenja, od kojih svaki nudi jedinstvene prednosti u rukovanju nestruktuiranim podacima. Njihova prilagodljivost i skalabilnost čine ih integralnim u arhitekturi organizacija koje se bore sa zahtevima za skladištenje koje postavljaju velike količine različitih podataka.

Dok istražujemo tehnologije koje se razvijaju, postaje očigledno da se tradicionalno razgraničenje između struktuiranih i nestruktuiranih podataka zamagljuje. Savremene baze podataka i skladišta podataka, posebno oni koji prihvataju NoSQL paradigme, pokazuju spremnost da se prilagode nestruktuiranim formatima podataka, premošćujući jaz između struktuiranih i nestruktuiranih podataka za skladištenje.

Model “3V” - zapremina, brzina i raznovrsnost - dugo je povezan sa velikim podacima, obuhvatajući suštinu izazova koje postavljaju veliki i raznovrsni skupovi podataka. Međutim, kako proširujemo naše razumevanje, model “5V” uvodi dve dodatne dimenzije - verodostojnost i vrednost — naglašavajući važnost tačnosti podataka i krajnji cilj izvlačenja smislenih uvida.

U navigaciji kroz prostor skladištenja čestih i nestruktuiranih podataka, organizacije moraju ne samo da se pozabave tehničkim razmatranjima već i da usklade svoje strategije sa specifičnim poslovnim potrebama. Potraga za obradom u realnom vremenu, efikasnim mogućnostima pretraživanja i bezbednim, pouzdanim skladištenjem zahteva holistički pristup.

U zaključku, skladištenje velikih količina frekventnih i nestruktuiranih podataka nije samo tehnološka zagonetka već i strateški imperativ za organizacije koje imaju za cilj da iskoriste puni potencijal svojih podataka. Dok se nalazimo na raskrsnici tehnoloških inovacija i rastućih količina podataka, put ka efikasnim rešenjima za skladištenje nastavlja da se razvija, obećavajući nove mogućnosti za stvaranje uvida i informisano donošenje odluka u dinamičnom pejzažu velikih podataka.

# LITERATURA

[1] Hedli Vikam : R ZA STATISTIČKU OBRADU PODATAKA

[2] Viktor Mayer-Schonberger, Kenneth Cukier : Big Data

[3] https://www.mongodb.com/unstructured-data/database

[4] https://www.geeksforgeeks.org/what-is-unstructured-data