UNIVERZITET U NOVOM SADU

FAKULTET TEHNIČKIH NAUKA 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**: Milorad Maksić E5-17/2024

ASISTENT: Bojana Samardžić

Novi Sad, oktobar, 2025.

Sadržaj

1. UVOD

Savremeno digitalno društvo karakteriše eksponencijalni rast količine podataka koji se svakodnevno generišu iz različitih izvora. Prema procjenam, globalna količina podataka udvostručava se svakih nekoliko godina, dostigavši impresivan nivo od preko 120 zetabajta do kraja 2023. godine. Ovaj fenomen poznat kao "Big Data" ne predstavlja samo izazov u smislu obima, već i u pogledu prirode samih podataka, odnosno njihove raznovrsnosti, brzine generisanja i kompleksnosti obrade.

Posebnu kategoriju u okviru Big Data ekosistema čine frekventni podaci, odnosno podaci koji se generišu kontinuirano i u velikim količinama u realnom vremenu. Ova kategorija obuhvata senzorske podatke, log fajlove, transakcijske zapise i podatke sa IoT (Internet of Things) uređaja. Istovremeno, preko 80% svih digitalnih podataka spada u kategoriju nestruktuiranih podataka. Ove informacije ne prate unaprijed definisani model ili šemu, što ih čini posebno izazovnim za obradu. U ovu grupu spadaju tekstualni dokumenti, slike, video zapisi, poruke sa društvenih mreža i multimedijalni sadržaj raznovrsnih formata.

Tradicionalni sistemi za upravljanje podacima, koji su dizajnirani za rad sa strukturiranim podacima u relacionim bazama, pokazali su se neadekvatnim za rukovanje ovakvim izazovima. Potreba za skladištenjem, obradom i analizom velikih količina frekventnih i nestruktuiranih podataka rezultirala je razvojem novih tehnologija, arhitektura i pristupa koji omogućavaju efikasno upravljanje ovim resursima. Distribuirani sistemi, NoSQL baze podataka i streaming platforme postali su standardni alati u modernim infrastrukturnim sistemima.

Efikasno skladištenje frekventnih i nestruktuiranih podataka postalo je kritično za brojne oblasti savremenog poslovanja i tehnologije. U infrastrukturnim sistemima, kao što su pametni gradovi, energetske mreže, transportni sistemi i industrijska automatizacija, sposobnost brzog prikupljanja, skladištenja i analize podataka direktno utiče na efikasnost sistema, pouzdanost operacija i donošenje pravovremenih odluka.

2. BIG DATA

Big Data predstavlja skupove podataka čiji su obim, brzina generisanja i raznovrsnost toliko veliki da zahtjevaju nove tehnologije i metode za njihovo skladištenje, obradu i analizu. Ovaj termin ne označava samo količinu podataka, već i novu paradigmu u načinu na koji organizacije prikupljaju, analiziraju i koriste informacije za donošenje odluka i stvaranje vrednosti. Za razliku od tradicionalnih pristupa gde su podaci bili statični i periodično ažurirani, Big Data podrazumjeva dinamičko okruženje u kome se podaci neprestano generišu, obrađuju i analiziraju.

Za razliku od tradicionalnih sistema za upravljanje podacima, Big Data pristup podrazumjeva distribuiranu obradu kroz korišćenje više računara koji rade paralelno kako bi se podaci obradili brže i efikasnije. Ovaj pristup omogućava horizontalnu skalabilnost, što znači da se kapacitet sistema može povećati jednostavnim dodavanjem novih čvorova, umesto nadogradnje postojećeg hardvera. Sistemi su dizajnirani sa tolerancijom na greške, što znači da mogu nastaviti sa radom čak i ako neki čvorovi prestanu da funkcionišu. Dodatno, Big Data sistemi su prilagođeni radu sa različitim tipovima podataka, od struktuiranih tabela do nestruktuiranih slika i video zapisa, što predstavlja značajnu razliku u odnosu na tradicionalne relacione baze podataka.

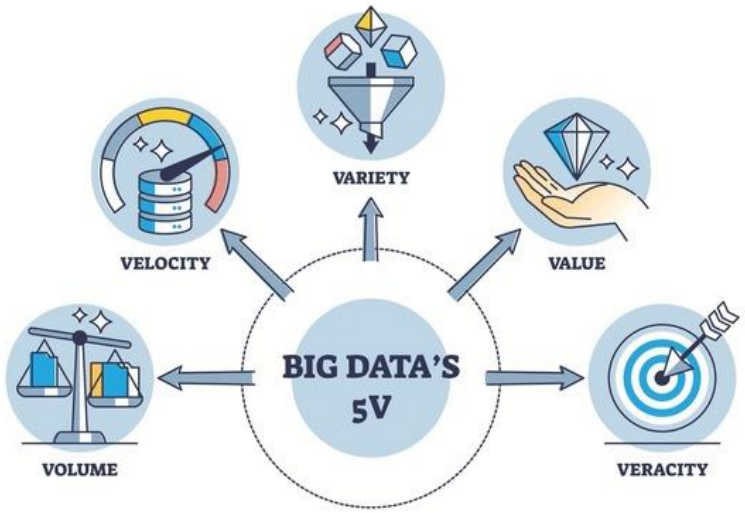
Kroz analizu podataka u realnom vremenu, omogućava se donošenje informisanih odluka koje mogu biti ključne za konkurentsku prednost. Procesi se optimizuju kroz prediktivnu analitiku i automatizaciju, što smanjuje troškove i povećava efikasnost.

1 byte	
Kilobyte ≈ 1000 (10^3) bytes	
Megabyte ≈ 1000000 (10^6) bytes	
Gigabyte ≈ 1000000000 (10^9) bytes	25 gigabajta: podaci koje *Ford Fusion Energy plug-in hibrid* analizira u toku jednog sata 60 gigabajta: podaci koje *Google self-driving* automobil sakupi u toku jednog sata 140 gigabajta: podaci koje *Nokia Here Maps* aplikacija sakupi u toku jednog dana
Terabyte ≈ 1000000000000 (10^{12}) bytes	30 gigabajta: podaci koje *Boing 777* prikupi u toku jednog prekookeanskog leta
Petabyte ≈ 1000000000000000 (10^{15}) bytes	Nekoliko petabajta: podaci o saobraćaju skladišteni na *Intix* platformi u svrhe analize saobraćaja za npr. *Google Traffic*
Exabyte ≈ 1000000000000000000 (10^{18}) bytes	
Zettabyte ≈ 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 4,4 zetabajta: procenjena veličina digitalnog univerzuma u 2013. godini
Yottabyte ≈ 1000000000000000000000000 (10^{24}) bytes	

Slika 1 Prikaz velicine podataka

2. 1 Karakteristike Big Data- 5V model

Big Data se tradicionalno opisuje kroz model "5V", koji obuhvata pet ključnih karakteristika koje definišu ovaj fenomen. Ovaj model pruža sveobuhvatan okvir za razumevanje složenosti i izazova koje Big Data predstavlja za moderne organizacije.



Slika 2 Ilustracija 5V

2.1.1 Veliki obim (Volume)

Obim podataka u Big Data okruženju meri se u terabajtima, petabajtima, pa čak i egzabajtima. Da bi se ovo stavilo u perspektivu, jedan terabajt je približno ekvivalent 500 sati HD videa, dok jedan petabajt predstavlja količinu podataka koju Google procesira u jednom danu. Jedan egzabajt, pak, može biti uporediv sa svim podacima koje ljudska civilizacija generiše u jednoj godini. Ove ogromne količine podataka potiču iz različitih izvora, uključujući IoT senzore koji generišu milione očitavanja u sekundi, društvene mreže sa milijardama postova dnevno, video nadzorne sisteme koji snimaju neprekidno, naučna istraživanja I eksperimenti.

Tradicionalni sistemi za skladištenje nisu dizajnirani da efikasno rukuju ovakvim količinama podataka. Relacione baze podataka, koje su dugo bile standard u industriji, pokazuju značajan pad performansi kada se suoče sa terabajtima informacija. Ovo je dovelo do razvoja distribuiranih sistema za skladištenje koji podatke raspoređuju kroz više čvorova, omogućavajući paralelnu obradu i pristup. **Hadoop Distributed File System (HDFS)** i objektna skladišta poput **Amazon S3** postali su standardna rešenja za ovakve izazove, omogućavajući organizacijama da ekonomično skladište i upravljaju petabajtima podataka.

2.1.2 Visoka brzina (Velocity)

Brzina se odnosi na tempo kojim se podaci generišu i moraju biti obrađeni. U mnogim scenarijima, vrednost podataka opada sa vremenom, što znači da analiza mora biti izvršena gotovo trenutno kako bi se maksimizovala korist. Različiti sistemi zahtjevaju različite nivoe brzine

obrade. **Batch obrada** podrazumeva prikupljanje podataka tokom određenog perioda i njihovu periodičnu obradu, na primer jednom dnevno. Ovaj pristup je jednostavniji za implementaciju, ali donosi značajno kašnjenje između generisanja i obrade podataka**. Real-time obrada**, s druge strane, zahtjeva trenutnu obradu podataka odmah nakon što stignu, što je kritično za aplikacije poput saobraćajnih sistema ili detekcije prevara. **Stream processing** predstavlja kontinuiranu obradu podataka koji pristižu kao tok, što je neophodno za finansijske transakcije ili monitoring kritičnih sistema.

Finansijske institucije ilustruju važnost brzine u Big Data sistemima. One moraju da detektuju potencijalne prevare za milisekunde kako bi sprečile neovlašćene transakcije. Svako kašnjenje može rezultirati značajnim finansijskim gubicima.

2.1.3 Velika raznovrsnost (Variety)

Raznovrsnost se odnosi na različite tipove i formate podataka koji postoje u Big Data ekosistemu. **Struktuirani podaci** su tradicionalno organizovani u tabelama sa fiksnim šemama i lako se skladište u relacionim bazama podataka. Ovi podaci obuhvataju finansijske zapise, CRM sisteme i inventar, gde je svaki podatak jasno definisan i kategorizovan. Međutim, struktuirani podaci čine samo mali procenat ukupne količine podataka u modernim organizacijama.

Polustruktuirani podaci imaju izvesnu organizaciju ali ne prate striktnu šemu. Često koriste XML, JSON ili CSV formate i obuhvataju log fajlove, API odgovore i email poruke. Ovi podaci zahtjevaju različite tehnike obrade u odnosu na potpuno struktuirane podatke, ali ipak zadržavaju izvestan stepen organizacije koji olakšava njihovu analizu**. Nestruktuirani podaci**, koji čine više od **80%** svih podataka u modernim organizacijama, predstavljaju najveći izazov. Ova kategorija obuhvata tekstualne dokumente, slike, video zapise, audio fajlove i podatke sa društvenih mreža. Nestruktuirani podaci ne mogu se direktno procesirati tradicionalnim alatima za analizu i zahtevaju specijalizovane tehnike poput Natural Language Processing (NLP) za tekst, Computer Vision za slike i video, ili Audio transcription za zvučne zapise.

Raznovrsnost podataka nameće potrebu za fleksibilnim sistemima skladištenja koji mogu efikasno rukovati sa svim tipovima informacija. NoSQL baze podataka, objektna skladišta i Data Lakes razvijeni su upravo kao odgovor na ovaj izazov, omogućavajući organizacijama da skladište podatke u njihovom nativnom formatu bez potrebe za prethodnom transformacijom.

2.1.4 Verodostojnost (Veracity)

Verodostojnost se odnosi na kvalitet, tačnost i pouzdanost podataka. S obzirom da podaci dolaze iz različitih izvora i u različitim formatima, često postoje problemi sa njihovom

konzistentnošću i tačnošću. Netačnost može proistići iz grešaka u kucanju, grešaka senzora ili neažurnih podataka. Nepotpuni podaci, gde nedostaju određene vrednosti ili atributi, takođe predstavljaju značajan problem. Dvosmislenost podataka javlja se kada različiti izvori koriste različite formate za iste informacije, na primer različite načine zapisivanja datuma. Protivrečni podaci nastaju kada različiti izvori daju različite vrednosti za isti parametar, što može dovesti do konfuzije i pogrešnih zaključaka.Rešavanje problema verodostojnosti zahteva implementaciju rigoroznih procesa validacije podataka pri unosu.

2.1.5 Vrijednost (Value)

Vrijednost predstavlja konačni cilj Big Data inicijativa - sposobnost da se iz ogromnih količina podataka izvuku korisni uvidi i kreira poslovna vrednost. Samo prikupljanje i skladištenje podataka nije dovoljno; oni moraju biti pretvoreni u akcione informacije koje doprinose poslovnim ciljevima. Stvaranje vrijednosti iz podataka podrazumeva identifikaciju relevantnih pitanja na koja podaci mogu odgovoriti, primenu odgovarajućih analitičkih metoda koje će otkriti skrivene obrasce, vizualizaciju rezultata na razumljiv način koji olakšava interpretaciju, i konačno, integraciju uvida u procese donošenja odluka.

Vrijednost Big Data sistema ne ogleda se samo u tehničkim mogućnostima već u poslovnim rezultatima koje omogućava. Kompanije koje uspešno implementiraju Big Data analitiku mogu ostvariti konkurentsku prednost kroz bolje razumevanje potreba kupaca, optimizaciju operativnih procesa, predviđanje tržišnih trendova i personalizaciju ponude.

2. 2 Big Data u infrastrukturnim sistemima

U kontekstu infrastrukturnih sistema, Big Data tehnologije igraju kritičnu ulogu u omogućavanju pametnih gradova. Milioni senzora kontinuirano prikupljaju podatke o saobraćaju, kvalitetu vazduha, potrošnji vode i energije, omogućavajući gradskim vlastima da optimizuju resurse i usluge. Pametni semafori prilagođavaju trajanje signala na osnovu trenutnog saobraćajnog opterećenja, smanjujući gužve i vreme putovanja. Sistemi za praćenje kvaliteta vazduha omogućavaju pravovremeno upozoravanje građana o opasnim nivoima zagađenja. Monitoring potrošnje vode pomaže u detekciji curenja i optimizaciji distribucije.

Energetske mreže predstavljaju drugu kritičnu oblast primene Big Data tehnologija. Pametna brojila i senzori generišu podatke u realnom vremenu koji se koriste za balansiranje ponude i potražnje električne energije. Algoritmi za predviđanje kvarova analiziraju podatke sa električnih vodova i transformatora, omogućavajući preventivno održavanje i smanjenje vremena nedostupnosti.

Transportni sistemi obilno koriste Big Data za optimizaciju ruta, smanjenje gužvi i poboljšanje bezbednosti. Podaci sa vozila, GPS uređaja i saobraćajnih senzora prikupljaju se i analiziraju u realnom vremenu. Sistemi za upravljanje javnim prevozom koriste ove podatke za dinamičko prilagođavanje reda vožnje, informisanje putnika o kašnjenjima i optimizaciju kapaciteta.

Industrijski IoT (Internet of Things) predstavlja četvrtu značajnu oblast primene. Senzori postavljeni na proizvodnu opremu neprekidno generišu podatke o temperaturi, vibracijama, pritiscima i drugim parametrima. Analiza ovih podataka omogućava predviđanje kvarova pre nego što se dogode, što je poznato kao preventivno održavanje. Ovo značajno smanjuje neplanirane zastoje u proizvodnji i produžava životni vek opreme.

3. FREKVENTNI I NESTRUKTUIRANI PODACI

3. 1 Definicija frekventnih podataka

Frekventni podaci, poznati u stručnoj literaturi kao **streaming data ili high-frequency data**, predstavljaju podatke koji se generišu kontinuirano, u velikim količinama i visokim brzinama. Za razliku od tradicionalnih statičkih skupova podataka koji se obrađuju periodično kroz batch processing, frekventni podaci zahtevaju kontinuiranu ili gotovo real time obradu.

Frekventni podaci karakterišu se kontinuiranim generisanjem, što znači da pristižu kao neprekidan tok bez jasno definisanih početka i kraja. Za razliku od batch podataka koji se prikupljaju u određenim vremenskim intervalima, streaming podaci neprestano dolaze u sistem. Visoka brzina je druga ključna karakteristika, gde tipično govorimo o hiljadama do miliona događaja u sekundi. Ova brzina varira zavisno od aplikacije, ali uvek prevazilazi mogućnosti tradicionalnih sistema za obradu. Vremenska senzitivnost predstavlja poseban izazov jer vrednost frekventnih podataka opada sa vremenom. Stariji podaci postaju manje relevantni, što nameće potrebu za brzom analizom i reakcijom. Konačno, varijabilni volumen znači da broj događaja može značajno varirati tokom vremena, što zahteva elastične sisteme koji mogu da se prilagode promenljivim opterećenjima.

Izvori frekventnih podataka su raznovrsni i obuhvataju brojne domene.

IoT senzori predstavljaju jedan od najznačajnijih izvora, uključujući industrijske senzore koji mere temperaturu, pritisak i vibracije u proizvodnim postrojenjima, pametne uređaje u domovima kao što su termostati, kamere i alarmni sistemi, senzore u vozilima koji prate GPS poziciju, brzinu i potrošnju goriva. Operativni sistemi generišu obimne količine frekventnih podataka kroz log fajlove aplikacija i servera, sistemske događaje i alarme, monitoring mrežnog saobraćaja i bezbednosne događaje poput pokušaja neovlašćenog pristupa.

Finansijski sistemi predstavljaju drugu značajnu kategoriju izvora frekventnih podataka. Transakcije kreditnih kartica koje se obrađuju u milisekundama, berze i trgovinski sistemi gde se izvršava veliki broj transakcija u sekundi, kao i kripto valute koje generišu podatke o transakcijama 24 sata dnevno, svi doprinose ogromnom toku informacija.

Izazovi povezani sa frekventnim podacima su višestruki i zahtevni. **Obrada u realnom vremenu** podrazumeva da sistem mora da procesira podatke odmah po prispeću, bez značajnih kašnjenja. **Upravljanje pritiskom**, poznato kao backpressure u stručnoj literaturi, predstavlja situaciju kada podaci stižu brže nego što sistem može da ih obradi, što može dovesti do preopterećenja sistema. **Tolerancija kašnjenja** mora biti pažljivo određena jer različite aplikacije imaju različite zahteve - dok finansijske transakcije zahtevaju minimalno kašnjenje, neki analitički zadaci mogu tolerisati veća kašnjenja. **Pitanje skladištenja** je kompleksno jer nije moguće niti potrebno skladištiti sve podatke dugoročno, što zahteva strategije za određivanje šta treba zadržati a šta može biti odbačeno.

3. 2 Nestruktuirani podaci- vrste i karakteristike

Nestruktuirani podaci predstavljaju informacije koje ne prate unapred definisanu šemu ili model podataka. Za razliku od strukturiranih podataka koji se lako uklapaju u tabele sa redovima i kolonama, nestruktuirani podaci dolaze u različitim formatima i ne mogu se direktno procesirati tradicionalnim relacionim bazama podataka

Tekstualni podaci čine značajan deo nestruktuiranih podataka i obuhvataju prirodni jezik u različitim oblicima. Email poruke i poslovna korespondencija, dokumenti u Word, PDF i drugim formatima, objave na društvenim mrežama, chat konverzacije, članci, blogovi i korisničke recenzije - sve ovo predstavlja tekstualne nestrukturirane podatke. Dodatno, tehnički tekstovi poput log fajlova aplikacija, sistem log-ova, programskog koda i dokumentacije, kao i XML i JSON strukture bez fiksne šeme, takođe spadaju u ovu kategoriju. Analiza tekstualnih podataka zahteva primenu naprednih tehnika Natural Language Processing (NLP) koje mogu da ekstrahuju značenje, sentiment i strukturu iz slobodnog teksta.

Multimedijalni podaci predstavljaju drugu veliku kategoriju nestruktuiranih podataka. Slike, uključujući fotografije u JPEG, PNG i RAW formatima, medicinske snimke poput rendgenskih, MRI i CT skenova, satelitske slike, skenirane dokumente i infografike, zauzimaju značajne resurse skladištenja. Jedna visokokvalitetna fotografija može zauzeti 10-50 megabajta prostora. Video zapisi, sa druge strane, predstavljaju najveće izazove u smislu skladištenja i obrade. Audio podaci, uključujući glasovne poruke i pozive, muzičke fajlove, podcast emisije, snimke sastanaka i zvučne zapise senzora, takođe spadaju u nestrukturirane podatke.

Senzorski podaci, iako često numerički, mogu biti nestruktuirani zbog nedostatka fiksne šeme. IoT senzori različitih tipova, GPS koordinate i putanje kretanja, akcelerometri i žiroskopi,

temperaturni i pritisni senzori, kao i biometrijski podaci, generišu informacije koje variraju u strukturi i formatu zavisno od tipa senzora i aplikacije.

Kontekstualna zavisnost je još jedna značajna karakteristika nestruktuiranih podataka. Značenje često zavisi od konteksta u kome se podaci koriste. Ista reč može imati različito značenje u različitim kontekstima, slika može biti nejasna bez dodatnih informacija, a zvučni zapis može biti nerazumljiv bez razumevanja situacije u kojoj je snimljen.

Raspodjela podataka u organizacijama jasno pokazuje dominaciju nestruktuiranih podataka. Prema industrijskim istraživanjima, između 80 i 90 procenata svih podataka spada u kategoriju nestruktuiranih. Polustruktuirani podaci čine 10 do 15 procenata, dok samo 5 do 10 procenata podataka ima potpuno strukturiranu formu. Još značajnije, nestruktuirani podaci rastu 10 do 50 puta brže od strukturiranih podataka, što dodatno naglašava važnost razvoja efikasnih metoda za njihovo skladištenje i obradu.

Slika 3 Prikaz omjera tipova podataka

4. Izazovi obrade i skladištenja

Skladištenje nestruktuiranih podataka predstavlja fundamentalni izazov za tradicionalne sisteme za upravljanje podacima. Relacione baze podataka, koje su dizajnirane za strukturirane podatke, nisu efikasne za skladištenje nestruktuiranih informacija. BLOB (Binary Large Object) polja, koja relacione baze koriste za čuvanje binarnih podataka, mogu tehnički skladištiti nestrukturirane podatke, ali to nije efikasno. Pretraživanje i indeksiranje takvih podataka je izuzetno sporo, performanse drastično opadaju sa povećanjem veličine, a troškovi skladištenja postaju prohibitivno visoki. Dodatno, relacione baze nisu dizajnirane za paralelnu obradu velikih binarnih objekata, što ih čini nepraktičnim za Big Data aplikacije.

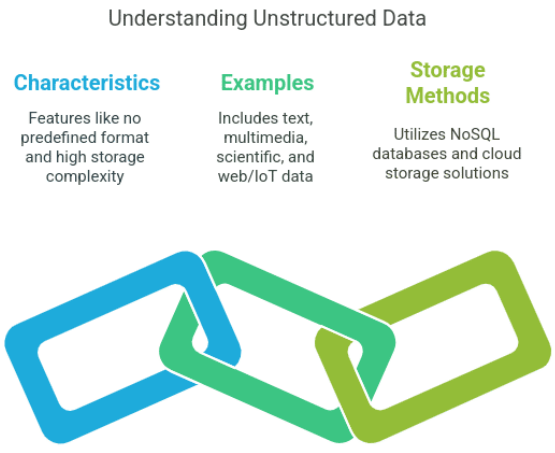
Alternativni pristupi razvijeni su kao odgovor na ova ograničenja. **File sistemi** predstavljaju najjednostavnije rešenje, gde se podaci čuvaju kao fajlovi na disku. Ovaj pristup je jednostavan

za implementaciju, ali nije skalabilan za velike količine podataka. Pretraživanje je teško, a upravljanje metapodacima postaje kompleksno kada se broj fajlova meri u milionima ili milijardama.

NoSQL baze podataka razvijene su specifično za fleksibilno skladištenje različitih tipova podataka. Document stores, kao što je MongoDB, skladište podatke u JSON ili BSON dokumentima bez potrebe za fiksnom šemom. **Key-value stores** poput Redis-a omogućavaju brz pristup podacima kroz jednostavne parove ključ-vrednost**. Wide-column stores** kao što je Cassandra pružaju kombinaciju strukture i fleksibilnosti, omogućavajući skladištenje različitih tipova podataka u istoj tabeli.

Performanse i skalabilnost postaju kritični faktori sa porastom količine frekventnih i nestruktuiranih podataka. Sistemi moraju biti dizajnirani za horizontalnu skalabilnost, što znači dodavanje novih čvorova umesto nadogradnje postojećeg hardvera. Ovaj pristup omogućava praktično neograničeno povećanje kapaciteta dodavanjem dodatnih mašina u klaster. Distribuirana obrada predstavlja srž modernih Big Data sistema, gde se posao paralelizuje na više mašina koje rade istovremeno.

Caching strategije igraju ključnu ulogu u poboljšanju performansi. Često korišćeni podaci se čuvaju u brzoj memoriji (RAM) ili SSD diskovima, omogućavajući pristup u milisekundama umesto sekundi ili minuta potrebnih za pristup podacima na HDD-u ili u udaljenim cloud skladištima. Indexing tehnike, kao što su inverzni indeksi za tekstualne podatke ili geospacijalni indeksi za lokacijske podatke, omogućavaju brzo pretraživanje kroz ogromne količine informacija.



Slika 4 Razumjevanje nestruktuiranih podataka

5. TEHNOLOGIJE ZA SKLADIŠTENJE PODATAKA

5. 1 Distribuirani file sistemi

Distribuirani file sistemi predstavljaju osnovu za skladištenje velikih količina podataka kroz više računara, omogućavajući skalabilnost, pouzdanost i visoke performanse koje tradicionalni centralizovani sistemi ne mogu postići. Ovi sistemi dijele podatke kroz više čvorova, omogućavajući paralelnu obradu i pristup, što je ključno za efikasno rukovanje Big Data skupovima.

5. 2 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) predstavlja jedan od najpoznatijih i najšire korišćenih distribuiranih file sistema u Big Data ekosistemu. HDFS je dizajniran za skladištenje izuzetno velikih fajlova, mjerenih u gigabajtima ili terabajtima, preko grupa standardnih računara. Za razliku od tradicionalnih file sistema koji zahtevaju skup hardver, HDFS može raditi na losijem hardware-u, što značajno smanjuje troškove implementacije.

Arhitektura HDFS-a zasniva se na master-slave modelu, gdje NameNode igra ulogu master čvora, dok DataNodes funkcionišu kao slave čvorovi. NameNode upravlja metapodacima i imenskim prostorom file sistema. Ovaj centralni čvor održava kompleto stablo direktorijuma, prati na kojim DataNode-ovima se nalaze pojedini blokovi podataka, i koordinira sve operacije pristupa. Važno je napomenuti da NameNode ne skladišti stvarne podatke već samo informacije o njihovoj lokaciji i organizaciji. DataNodes, s druge strane, čuvaju stvarne blokove podataka, izvršavaju read i write operacije na zahtev klijenata ili NameNode-a, i redovno šalju heartbeat signale NameNode-u kako bi potvrdili svoju dostupnost.

Replikacija podataka predstavlja ključnu karakteristiku HDFS-a koja obezbeđuje visoku dostupnost i pouzdanost. Sistem automatski kreira više kopija svakog bloka podataka, pri čemu je podrazumevani faktor replikacije tri. Ovo znači da svaki blok postoji na tri različita DataNode-a, što obezbeđuje da podaci ostanu dostupni čak i ako dođe do kvara jednog ili dva čvora. Replikacija takođe omogućava paralelizaciju read operacija, jer klijenti mogu čitati podatke sa najbližeg ili najmanje opterećenog čvora koji ima potreban blok. Dodatno, replikacija pruža zaštitu od hardverskih kvarova koji su neizbežni u velikim klasterima sa stotinama ili hiljadama čvorova.

Prednosti HDFS-a su brojne i dobro dokumentovane u praksi. Visoka tolerancija na greške omogućava sistemu da nastavi sa radom čak i kada dođe do kvara pojedinačnih čvorova, što je kritično za pouzdanost u produkcijskim okruženjima. Skalabilnost sistema je gotovo linearna, jer se kapacitet i performanse mogu povećati jednostavnim dodavanjem novih čvorova u klaster.

Ekonomičnost je značajna prednost jer HDFS ne zahteva skup specijalizovani hardver, već može raditi na standardnim serverima.

Međutim, HDFS takođe ima određena ograničenja. Sistem nije efikasan za skladištenje velikog broja malih fajlova, jer svaki fajl zauzima prostor u NameNode memoriji za metapodatke, bez obzira na svoju veličinu.

5. 3 NoSQL baze podataka

NoSQL (Not Only SQL) baze podataka razvijene su kao odgovor na ograničenja relacionih baza u kontekstu Big Data aplikacija, posebno za nestrukturirane i polustruktuirane podatke. Termin "NoSQL" ne označava potpuno odbacivanje SQL-a ili relacionog modela, već označava širu kategoriju baza podataka koje nude alternative za različite slučajeve korišćenja.

5. 4 Document Stores- MongoDB

MongoDB predstavlja jedan od najpopularnijih document store sistema, koji skladišti podatke u BSON (Binary JSON) formatu. BSON predstavlja binarnu reprezentaciju JSON dokumenata sa dodatnom podrškom za tipove podataka koji ne postoje u standardnom JSON-u, kao što su datumi i binarne vrednosti. Svaki dokument u MongoDB-u je nezavisna jedinica koja može imati potpuno različitu strukturu od drugih dokumenata u istoj kolekciji, što pruža fundamentalnu fleksibilnost u modeliranju podataka.

Fleksibilna šema predstavlja jednu od ključnih prednosti MongoDB-a. Za razliku od relacionih baza gde se šema mora definisati unapred i gde je promena šeme skupa i kompleksna operacija, MongoDB dozvoljava da različiti dokumenti u istoj kolekciji imaju različita polja.

Prednosti MongoDB-a uključuju eliminaciju potrebe za unapred definisanom šemom, što omogućava brže razvijanje i lakše prilagođavanje promenama. Struktura podataka se može lako menjati bez potrebe za migracijom. Odlične performanse za read-heavy workloads postižu se kroz indeksiranje i mogućnost distribuiranja read operacija. Rich query language pruža ekspresivne mogućnosti za složene upite.

Nedostaci sistema uključuju manju konzistentnost u poređenju sa relacionim bazama, jer MongoDB koristi eventual consistency model u distribuiranim konfiguracijama. JOIN operacije nisu efikasne u MongoDB-u, jer je sistem dizajniran oko denormalizovanih podataka. Skladištenje može zauzeti više prostora nego u relacionim bazama zbog denormalizacije i BSON overhead-a.

5. 5 Key-Value Stores- Redis

Redis (Remote Dictionary Server) predstavlja najjednostavniji oblik NoSQL baza, gde je svaki podatak organizovan kao par ključa i vrednosti. Međutim, Redis je daleko od toga da bude samo

jednostavan key-value store. Sistem podržava različite strukture podataka i pruža moćne funkcionalnosti koje ga čine pogodnim za širok spektar aplikacija.

In-memory storage je fundamentalna karakteristika Redis-a. Svi podaci se čuvaju u RAM-u, što omogućava ekstremno brze performanse sa latencijom merenom u submilisekundama. Ovo Redis čini idealnim za slučajeve korišćenja gde je brzina kritična. Međutim, Redis nije ograničen samo na memory storage. Sistem pruža opcije za persistence podataka na disk kroz RDB (Redis Database) snapshots koji periodično snimaju stanje cele baze, ili kroz AOF (Append Only File) koji loguje svaku write operaciju, omogućavajući replay u slučaju restarovanja.

5. 6 Streaming platforme za frekventne podatke

Streaming platforme predstavljaju specijalizovane sisteme dizajnirane za rukovanje kontinuiranim tokovima podataka. Za razliku od batch sistema koji obrađuju statičke skupove podataka, streaming platforme moraju upravljati podacima koji neprestano pristižu, često u ogromnim količinama i visokim brzinama.

5. 7 Apache Kafka

Apache Kafka se nametnula kao de facto standard za streaming platformu u Big Data ekosistemu. Kafka predstavlja distribuirani sistem koji omogućava publish-subscribe messaging, trajno skladištenje event stream-ova i obradu stream-ova u realnom vremenu. Originalno razvijena u LinkedIn-u za rukovanje njihovim ogromnim količinama log podataka, Kafka je postala open-source projekat koji danas koriste hiljade organizacija širom sveta.

Arhitektura Kafka se zasniva na producer-consumer modelu, gde produceri šalju poruke (events) u Kafka klaster, a consumeri čitaju te poruke. Ključna razlika u odnosu na tradicionalne message queue sisteme je što Kafka trajno skladišti sve poruke na disk, omogućavajući multiple consumers da čitaju iste podatke, i replay istorijskih podataka kada je potrebno.

Ključni koncepti Kafka arhitekture uključuju topics, partitions i brokers. Topics predstavljaju kategorije ili feed names gde se poruke objavljuju. Svaki topic je logička kategorija koja može sadržati milione poruka. Svaki topic se deli na jedan ili više partitions, gde svaka particija predstavlja ordered, immutable sequence poruka.

Brokers su serveri koji čine Kafka klaster. Svaki broker čuva subset particija i servira read i write zahteve za te particije. Kafka klaster tipično se sastoji od više brokera za visoku dostupnost. Particije su replikovane između brokera, obezbeđujući da podaci ostanu dostupni čak i ako neki broker postane nedostupan.

Skalabilnost se postiže jednostavnim dodavanjem novih brokera ili particija. Za razliku od mnogih distribuiranih sistema, dodavanje kapaciteta u Kafka često ne zahteva prekid servisa.

Prednosti Kafka uključuju centralizovanu platformu za sve streaming podatke u organizaciji, visoki throughput i nisku latenciju, trajno skladištenje stream-ova što nije tipično za message queue sisteme, i mogućnost replay-ovanja koja je kritična za mnoge slučajeve korišćenja. **Nedostaci** uključuju složeniju konfiguraciju i održavanje u poređenju sa jednostavnijim message queue sistemima, zavisnost od ZooKeeper-a za koordinaciju (mada novije verzije Kafka uklanjaju ovu zavisnost), i potrebu za dodatnim alatima kao što je Kafka Streams ili Apache Flink za kompleksnu stream obradu.

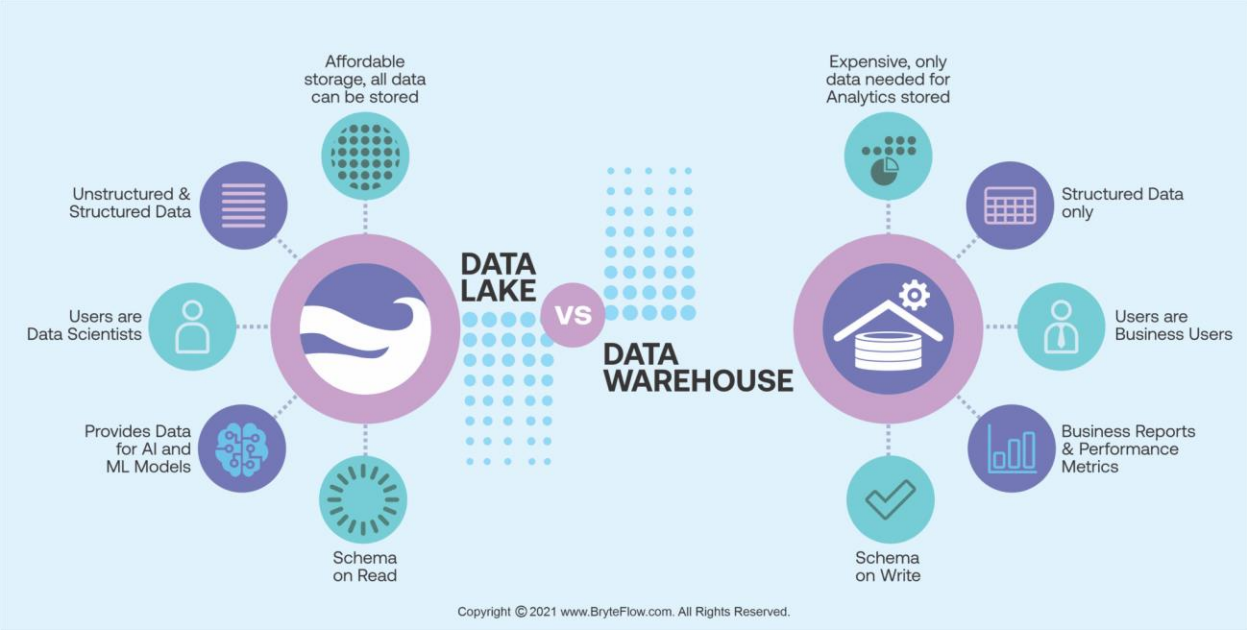
5. 8 Data Lakes i hibridna rešenja

Data Lakes predstavljaju paradigmu skladištenja koja omogućava organizacijama da sačuvaju sve svoje podatke u nativnom formatu, bez potrebe za prethodnom transformacijom ili strukturiranjem. Ovaj pristup fundamentalno se razlikuje od tradicionalnih Data Warehouse sistema.

5. 9 Koncept Data Lake

Data Lake je centralizovano skladište koje omogućava skladištenje svih struktuiranih, polustruktuiranih i nestruktuiranih podataka bez obzira na njihov obim. Podaci se čuvaju u njihovom nativnom formatu, a transformacija se primenjuje samo kada postoji konkretna potreba za analizom, što se naziva schema-on-read pristupom. Ovo se razlikuje od Data Warehouse pristupa gde se podaci transformišu pre skladištenja (schema-on-write).

Razlika između Data Lake i Data Warehouse je fundamentalna. **Data Warehouse** skladišti obrađene, strukturirane podatke koji su transformisani u skladu sa unapred definisanom šemom. Ovi podaci su optimizovani za specifične analitičke upite i izvještaje. Data Lake, s druge strane, skladišti sirove podatke u svim formatima. Data Warehouse koristi schema-on-write pristup gde se struktura podataka definiše pre skladištenja, dok Data Lake koristi schema-on-read gde se struktura primenjuje tokom čitanja. Data Lake je izuzetno skalabilan i ekonomičan jer koristi pristupacnog hardware ili cloud storage, dok je Data Warehouse često skuplji sa specijalizovanim hardverom. Use case za Data Lake je eksplorativna analiza i machine learning, dok je Data Warehouse fokusiran na strukturirane izvještaje i business intelligence.



Slika 5 Razlike Data Lake - Data Warehouse

5. 10 Lambda arhitektura

Lambda arhitektura predstavlja pristup koji kombinuje prednosti batch i stream processing-a. Problem koji Lambda arhitektura rešava je kako obezbediti kompletnu istorijsku analizu (što zahtjeva batch processing svih podataka) dok istovremeno pružaju real-time insights (što zahtjeva stream processing novih podataka).

Lambda arhitektura se sastoji od tri sloja. **Batch Layer** predstavlja master dataset koji skladišti sve podatke u raw formatu i periodično izvršava batch processing za kreiranje batch views. Ovaj layer koristi tehnologije kao što su HDFS za skladištenje i Apache Spark ili MapReduce za obradu. Batch layer garantuje kompletnost i tačnost podataka jer procesira sve dostupne informacije.

Speed Layer procesira samo nove podatke koje Batch Layer još nije obradio, kreirajući real-time views. Ovaj layer koristi streaming tehnologije kao što su Kafka, Flink ili Storm. Speed layer omogućava sistemu da pruži trenutne insights bez čekanja da batch job završi.

Serving Layer kombinuje rezultate iz batch i speed layera, omogućavajući korisnicima da dobiju odgovore koji uključuju i istorijske podatke i najnovije događaje. Ovaj layer koristi baze podataka optimizovane za brzo čitanje kao što su Cassandra, HBase ili ElasticSearch.

Prednosti Lambda arhitekture uključuju kompletnu istoriju podataka kroz batch layer, real-time insights kroz speed layer, i fault tolerance jer svaki layer može raditi nezavisno.

Nedostaci uključuju složenu arhitekturu koja zahteva održavanje dva odvojena sistema za obradu, duplikaciju logike jer isti algoritmi moraju biti implementirani i za batch i za stream processing, što može dovesti do neslaganja u rezultatima ako implementacije nisu identične.

6. ZAKLJUČAK

Skladištenje velike količine frekventnih i nestruktuiranih podataka predstavlja jedan od najznačajnijih tehničkih i organizacionih izazova savremenih informacionih sistema. Kroz ovaj rad, analizirane su ključne karakteristike Big Data fenomena, specifičnosti frekventnih i nestruktuiranih podataka, kao i tehnologije i pristupi koji omogućavaju efikasno upravljanje ovim resursima. Eksponencijalni rast količine podataka, kombinovan sa njihovom raznovrsnošću i brzinom generisanja, zahtjeva fundamentalno drugačiji pristup od tradicionalnih metoda zasnovan na relacionim bazama podataka i centralizovanim sistemima.

Analiza tehnologija za skladištenje pokazala je da ne postoji univerzalno rešenje koje bi bilo optimalno za sve scenarije. Svako rješenje ima svoje prednosti i mane.

Data Lakes su se nametnuli kao standard za skladištenje heterogenih podataka, omogućavajući organizacijama da čuvaju sve svoje podatke u nativnom formatu bez potrebe za prethodnom transformacijom. Lambda arhitektura ilustruje kako se mogu kombinovati batch i stream processing pristupe kako bi se postigla i kompletnost i aktuelnost podataka. Moderna cloud rešenja dodatno pojednostavljuju implementaciju kroz managed services koji eliminišu mnoge operativne složenosti, iako donose nove izazove kao što je vendor lock-in.

Uspješna implementacija sistema za skladištenje frekventnih i nestruktuiranih podataka zahtjeva pristup koji balansira tehničke mogućnosti sa poslovnim potrebama. Izbor tehnologija mora biti vođen specifičnim zahtevima organizacije umesto praćenja trenutnih trendova. Planiranje za rast je kritično jer sistemi moraju biti dizajnirani za horizontalno skaliranje od samog početka. Fokus na operativnu izvrsnost kroz monitoring, alerting, backup i disaster recovery nije opcija već neophodnost. Kontinuirano učenje i prilagođavanje su ključni jer se tehnologije brzo razvijaju i ono što je bilo optimalno rešenje može postati zastarelo.

Na kraju, uspjeh u ovoj oblasti nije samo stvar izbora pravih tehnologija već i izgradnje tima sa odgovarajućim veštinama, uspostavljanja jasnih procesa i donošenja odluka. Podatke često nazivaju novom naftom, ali ova analogija je nepotpuna - za razliku od nafte, vrednost podataka ne opada njihovim korišćenjem. Naprotiv, što više organizacija koriste svoje podatke, to više mogu naučiti i što bolje mogu optimizovati svoje operacije. Ključ je u posedovanju infrastrukture i sposobnosti da se ti podaci efikasno skladište, obrađuju i transformišu u akcione informacije koje pokreću poslovnu vrednost.

7. 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] Predavanja iz predmeta “Big data u infrastrukturnim sistemima” za 2024/25 god.