Tehnologije upravljanja podacima

Obrada podataka u realnom vremenu i analitičke tehnologije sa novim vrstama baza podataka

Daliborka Mačinković
Informacione tehnologije
Fond zdravstvenog osiguranja Republike Srpske
Banja Luka, Republika Srpska, BiH
daliborka.macinkovic@teol.net

Sadržaj—**Potreba za razvojem novih tehnologija upravljanja podacima uslovljena je okruženjem za rad sa velikom količinom podataka poznatom kao Big Data, sve većim brojem usluga u računarskom oblaku - Cloud Computing, rastućim brojem smart objekata u internetu - Internet of Things. U slučajevima kada relacione baze podataka ne mogu da osiguraju dovoljno dobre performanse pri radu sa ogromnom količinom heterogenih podataka razmatraju se NoSQL baze podataka. Nove vrste baza podataka trebaju omogućiti visoku dostupnost, brzinu obrade, sigurnost i privatnost, pouzdanost, konzistentnost, skalabilnost, distribuiranost. Sve veće izazove predstavljaju podaci nastali interakciom između uređaja (M2M – machine-to-machine interaction) i interneta stvari (IoT – Internet of Things) koje treba obraditi u realnom vremenu. Korišćenje tehnologija za brzu obradu podataka u realnom vremenu, memorijske analitike i analitike toka podataka, pruža nove mogućnosti. Poslovna inteligencija omogućila je korisnicima da iz podataka poslovanja preduzeća dobiju ključne informacije na temelju kojih će donositi poslovne odluke. Određena reaktivnost zahtijeva napredak u analitičkim tehnologijama prediktivne analitike, kontekstualno svjesnih servisa, bihevioralne analitike, kompleksnog procesiranja događaja.**

Ključne riječi- NoSQL baze podataka; memorijsko upravljanje podacima; analitika toka podataka; prediktivna analitika; kompleksno procesiranje događaja (CEP)

I. UVOD

Ovaj rad predstavlja nove tehnologije za upravljanje podacima i nove vrste baza podataka. Cilj jeste da se sagledaju pravci razvoja, trenutne tehnologije i izazovi za obradu i čuvanje podataka izazvani novim okolnostima.

"Big Data" se definiše kao količina podataka koju treba smijestiti, kojom treba upravljati i koju treba procesirati na efikasan način [1]. Big Data karakterišu 3V kao količina podataka, (Volume), brzina kojom podaci dolaze i obrađuju se (Velocity), raznorodne strukture podataka (Variety), ali i karakteristike kao što su vjerodostojnost u podacima (Veracity), neospornost podataka (Validity), nestalnost (Volatility) i dr.

Veliki pokretač primjene novih baza podatka i novih tehnologija upravljanja podacima je pojava Cloud Computinga. Unutar osnovnih usluga cloud-a izdvojila se i usluga DBaaS (Database as a Service) [2]. Neke od osnovnih zahtijeva koje bi trebale da imaju baze podatka koje treba da odgovore zahtjevima cloud okruženja su: visoka dostupnost, brzina, sigurnost i privatnost, pouzdanost, konzistentnost, skalabilnost, distribuiranost.

Internet of Things (IoT) je nova paradigma koja rapidno dobija na snazi u scenariju modernih bežičnih telekomunikacija. Osnovna ideja ovog koncepta je prisustvo oko nas različitih stvari ili objekata – kao što su Radio-Frequency IDentification (RFID) tagovi, senzori, aktuatori, mobilni telefoni itd. – koji su u mogućnosti, putem jedinsvenih adresinih šema, da ostvaruju interakciju i sarađuju za postizanje zajedničkih ciljeva [3]. Očekivanja su da će se broj smart uređaja do 2020. godine povećati na 50 milijardi, što predstavlja značajno povećanje broja smart uređaja po osobi[4]. Senzori i drugi uređaji generišu ogromnu količinu podataka koja zahtijeva obradu toka podataka u realnom vremenu. Informacije sa novih izvora podataka treba učiniti vrijednim, sa mnoštvom inteligencije i integrisati u ih u poslovne sisteme za stvaranje novih vrijednosti. Prisutni koncepti poput data collection, data processing, data mining, information sharing, information fusion, information integration ukazuju na raznorodne aktivnosti koje treba obaviti sa podacima u novonastalim okolnostima.

U današnjem procesu donošenja odluka, dostupnost podataka u realnom vremenu za precizne informacije su od ključnog značaja. Uz sve veći obim podataka, iz sveprisutnih objekata, povezanih uređaja, društvenih mreža, potrebe povezivanja sa podacima velikih ERP sistema, ispravno donošenje odluka u velikoj mjeri oslanja se na napredak u sposobnosti analitičkih tehnologija, koje trebaju donijeti inteligenciju u podacima. Novi oblici analitike su se pojavili da bi uklonili potrebu za prethodnim modelima metapodataka, što je rezultovalo sa bržim upitima i dinamičniom obradom podataka. Razni analitički alati imaju za cilj da izdvoje relevantne informacije iz ogromnih količina sirovih podataka i omoguće bržu obradu podataka u toku. In-memory analitika kešira velike količine podataka u RAM memoriju umjesto na fizičke diskove čime se smanjuje vrijeme upita nad podacima i pojačava brzina donošenja odluka. Streaming analititika je

jedna od oblika analitike sa potrebom analize podataka u pokretu koje treba obraditi u realnom vremenu, tako da odluke mogu biti donesene u sekundama.

Rad je organizovan u četiri dijela. Nakon uvoda u drugom dijelu opisuju se in-memory analitika sa tehnologijama, in-memory data management i in-memory low-latency messaging, i streaming analitika. Ukazuje se na tehnologije za inteligentno donošenje odluka kao što su prediktivna analitika sa pristupima zasnovanim na obrascima, pravilima i kontroli statističkih procesa, bihevioralna analitika, kompleksno procesiranje događaja, konteksno-svjesni računarski servisi. Posebno se predstavlja MapReduce. Treće poglavlje opisuje karakteristike NoSQL baza podataka najčeće grupisane kao ključ-vrijednost, orijentisane prema kolonama, dokumentno orijentisane, objektno orijentisane, bazirane na grafovima. U četvrtom poglavlju daje se zaključak o novim tehnologijama za upravljanje podacima izazvane novim okolnostima.

II. NOVE TEHNOLOGIJE UPRAVLJANJA PODACIMA

Obrada podataka generisanih interakciom između uređaja (M2M – machine-to-machine interaction) i podataka sa novih izvora interneta stvari (IoT – Internet of Things) traži određena rješenja. Nedostaci trenutnih Big Data rješenja su nedostatak rješenja centralizovane brze obrade velikih količina podataka na distribuianim sistemima, paketno procesiranje (batch processing), nedovoljno poznavanje posebih znanja za analitiku podataka, visoka latentncija, troškovi za infrastruktu i energiju za čuvanje takvih podataka, nedovoljna skalabilnost za podatake.

Kod obrade toka podataka podaci stižu kontinuirano, tok podataka je nepredvidive veličine, a podaci se mogu nakon obrade sačuvati ili odbaciti. Pri obradi toka podataka potrebno je identifikovati važne događaje. Potrebno je obrađivati paralelno više tokova podataka različitih intenziteta. Sortiranje se vrši prema vremenu, te se izvode agregatne funkcije i unije nad ulaznim tokom podataka. Nad tokom podataka izvršavaju se kontinuirani upiti, dok jednokratne upite može zahtijevati korisnik.

A. Tehnologije za brzu obradu podataka (Speed of data processing technologie)[5]

1) In-memory analytics - Memorijska analitika je pristup podacima postavljanjem upita kada se podaci nalaze u memoriji sa slučajnim pristupom (RAM), za razliku od upita nad podacima koji su pohranjeni na fizičke diskove. Ovaj pristup znatno skraćuje vrijeme odziva, omogućavajući poslovnu inteligenciju (BI) i analitiku za podršku bržem donošenju odluka. Na tradicionalnoj disk-based analitičkoj platformi, metapodataci moraju biti kreirani prije samog procesa odvijanja analitike. Način gdje su metapodaci modelovani (po uzoru) zavisi od zahtijeva za analitikom. Mijenjanje načina za modelovanje metapodataka da bi se ispunili novi zahtjevi traži dobar nivo tehničkog znanja. Memorijska analitika smanjuje ili eliminiše potrebu za indeksiranjem i pohranjivanjem (unaprijed) pre-agregiranih podataka u OLAP kocke ili agregatne tablice. To omogućava developerima da uzmu u obzir sve moguće načine analize i poboljšaju relevantnost sadržaja analitike. Predviđa se da će buduće aplikacije zahtijevati brže vrijeme odziva upita, gdje in-memory analitika može biti ugrađena na nivou čipseta, dok se tradicionalna skladišta podataka mogu eventualno koristiti za podatke koji se često ne zahtijevaju.

a) In-memory analytics je omogućena sa nizom in-memory tehnologija:

- *In-memory data management* (IMDBMS): Sistem za upravljanje in-memory bazom podataka pohranjuje cijelu bazu podataka u računarski RAM, negirajući potrebu za disk I / O instrukcijama. To omogućava aplikacijama da se pokrenu u potpunosti u memoriji;
- *In-memory grid podataka (IMDG*): pruža distribuirano in-memory smiještanje podataka u kojoj višestruke, distribuirane aplikacije mogu smjestiti i dohvatiti velike količine podataka sa objekata.

b) In-memory low-latency messaging - Ova platforma pruža mehanizam za aplikacije za razmjenu poruka što je brže moguće kroz direktnu komunikaciju memorije.

2) Streaming analytics - Analitika toka je nova paradigma analize podataka koja ne zahtijeva čuvanje podataka. Ona obrađuje podatke u letu, čim stigu velikom brzinom toka, a zatim ih odbacuje kako bi se oslobodio prostor za naknadne podatake. Podaci sa IoT senzora i uređaja se stalno mijenja, i ne mogu predstavljati promjene koje su smislene, npr periodično ažuriranje informacija temperature. Streaming analitika se mora primijenjivati da izvuče smislene promjene u podacima, zatim za otkrivanje složenih obrazaca i s vremenom doći do akcije koja će imati značenje za sredinu. Neki primjeri aplikacija u realnom vremenu koje zahtijevaju streaming analitiku uključuju mrežne transportne podatke, telefonske razgovore, ATM transakcije i senzor podatake. Streaming analitika koristi složene algoritme za obradu trenutnih tokove podataka o događajima koje prima iz jednog ili više izvora. IoT zahtijeva analitiku koja će se izvoditi u realnom vremenu i omogućuje velike količine podataka koji se čuvaju za kasnije analize.

B. Tehnologije za inteligentno donošenje odluka (Intelligent decision-making technologies)[5]

1) Context-aware computing service- Kontekstno svjesni računarski servisi su računarska paradigma koja opisuje softver / hardver koji koristi kontekstualne informacije kako bi se omogućilo da sistem predviđa i djeluje u skladu sa profilom korisnika i predodređenim zahtjevima. Nakon što je sistem prepoznao "kontekst" u kojem se odvija interakcija, ove informacije mogu se koristiti za promjenu, pokretanje i prilagođavanje ponašanja aplikacije i sistema. Dakle, kontekst su sve informacije koje se mogu koristiti za opisivanje aktivnosti ili situacije entiteta, gdje je entitet osoba, mjesto ili objekat koji se smatra relevantnim od strane aplikacije za korisnika.

2) Predictive analytics- Prediktivna analitika je skup statističkih i analitičkih tehnika koje se koriste da otkriju

relacije (odnose) i obrasce (patterns) u okviru velike količine podataka, tako da se oni mogu koristiti za predviđanje ponašanja ili događaja. Postoje tri metoda kao pristupi u prediktivnoj analitici: [6]

a) Pattern-based approach (Pristup zasnovan na obrascu) - Ovaj pristup poredi performanse i konfiguracione podatke real-time sistema sa nestruktuiranim izvorima podataka koji mogu uključivati poznate neuspješne profile, istorijske neuspješne zapise i konfiguracione podatke. Cilj je da se ekstrahuju statistički obrasci u sklopu velikog višeslojnog skladišta podataka, koristeći moćne korelacijske mašine, kako bi se utvrdilo da li trenutni podaci konfiguracije i performanse ukazuju na vjerovatnoću neuspjeha.

b) Rule-based approach (Pristup zasnovan na pravilima) - Definiše se serija pravila, zasnovanih na statističkim analizama istorijskih podataka o performansama, prethodno identifikovnih neuspješnih načina i rezultata testiranja sistema. Svako pravilo može se porediti u odnosu na više izvora podataka i drugih vanjskih faktora kao što su doba dana, uslova rada i istovremene vanjske aktivnosti prema definisanim pragovima. Kršenje tih pravila zatim može biti prikupljeno i kao eskalacione rutine korišteno za utvrđivanje vjerovatnoće ozbiljnosti i uticaja na rezultate ili ispade.

c) Statistical process control-based approach(Pristup zasnovan na kontroli statističkih procesa) - Control charts su se pokazale kao neprocjenjiva pomoć u upravljanju kompleksnim, procesom vođenim sistema. Pojava retrofit - sposobnosti u realnom vremenu telemetrija i poboljšanja u prikupljanju podataka, rješenja i kapaciteta mreže za podršku velikih količina podataka, znači statističke tehnike koje počivaju na kvalitetu unutar proizvodnog prostora. Statističke anomalije mogu se lako identifikovati i koristi za pokretanje preventivnog djelovanja na odgovarajući način kako bi se osiguralo da utiču na performanse usluge.

3) Complex event processing (CEP) - Kompleksno procesiranje događaja obuhvata načine za obradu događaja, u toku njihovog pojavljivanja i izvođenje obrazaca u novopridošlim podacima o događaju. To je računarski stil koji je implementiran od strane događajem vođenih, kontinuiranih inteligentnih sistema. CEP sistem koristi algoritme i pravila za obradu tokove podataka koje dobija sa jednog ili više izvora, kao što su ERP aplikacije, financijske aplikacije, web i operativna analitika za generisanje uvida. To stvara nov sažet nivo činjenica ili složene događaje, i stavlja ih u kontekst za identifikovanje prijetnji i prilika. Ove informacije se zatim koriste za odgovor u smislu poslovnih aktivnosti. Obrada kompleksno procesiranih događaja se aktivira po prijemu podataka o događajima. CEP sistemi smiještaju veliku količinu događaja u memoriski prostor, agregiraju nepovezane događaje iz više izvora i izvršavaju vrlo složene analize kada podaci o događajima stignu. Rezultat koji se dobije CEP sistemom je znanje o kompleksnim distribuiranim incidentima koji se dešavaju u sistemu. CEP je posebno koristan za IoT brojne događaje koji se proizvode svakodnevno. Događaji koje generišu RFID čitači ili senzori smatraju se primitivnim događajima. Informacije unutar primitivnog događaja je prilično ograničena. Idući naprijed sa IOT aplikacijama u realnom vremenu, informacije su sve složenije i uključuju poslovnu logiku i pravila; izvode korisne informacije, kombinaciom primitivnih događaja u kompleksne događaje. U proizvodnji, u procesima nadzora i kontrole, CEP sistemi su posebno korisni za niske latencije kod prikupljnja podataka i kako bi za senzore osigurali da se ovi procesi izvode optimalno.

4) Behavioural analytics- Bihevioralna analitika je zasnovana na ponašanju i predstavlja kombinaciju strategija i alata koja omogućuje identifikaciju npr. pojedinačnih potrošača i njihovih preferencija potrošnje i ponašanja. Može se koristiti za identifikaciju izabranog korisnika ili korisnika kroz višestruke platforme u vremenu i najefikasnija je s različitim izvorima podataka, npr. mobilne mreže ili pretplatu baze podataka. U IoT, jedan primjer bihevioralne analitike je praćenje kretanja kupaca u maloprodaji kako bi se utvrdile namjere kupovine. Trgovci mogu aktivno istraživati takve sisteme za mjerenje ponašanja kupaca i predložiti odgovarajuće preporuke za ciljane kupce. Ostala područja primjene mogu biti u sistema nadzora visokog rizika, za ograničena područja. Sposobnost analitike može pratiti obrasce ponašanja za vjerovatnost kriminalnih i terorističkih aktivnosti. Tehnike filtriranja podataka kao što su anonimnost podataka, integracija podataka i sinhronizacija podataka, koriste se da sakriju detalje infrormacija pružajući samo informacije prema zahtjevima. Uz korišćenje apstrakcije podataka, informacije se mogu izdvojiti da pruže zajednički poslovni pogled pri čemu se dobija veća agilnost u domenu. Sigurnost je od najveće važnosti, integritet podataka omogućuje pouzdano i autentično donošenje odluka.

C. MapReduce

je programski model i odgovarajuća implementacija za procesiranje i generisanje velikih skupova podataka. Korisnici definišu map funkciju koja procesira parove ključ-vrijednost i generiše, u međukoraku, skup ključ-vrijednost parova i reduce funkciju, koja obrađuje sve vrijednosti iz međukoraka, koje su vezane za isti ključ [7]. Map i Reduce funkcije se pišu imajući u vidu podatke struktuirane kao ključ-vrijednost parove. Map funkcija uzima parove iz jednog domena podataka i vraća parove iz drugog domena: *map(k, v) -> <k′, v′>*

Map funkcija se izvršava paralelno za svaki ulazni skup podataka, i kao izlaz daje listu (k', v') parova za svaki poziv. Zatim MapReduce radni okvir skuplja sve parove sa istim ključem iz svih listi, i grupiše ih zajedno, tako kreirajući po jednu listu vrijednosti, za svaki od različitih ključeva. Slijedi Reduce funkcija, koja se takođe izvršava paralelno, i kao izlaz proizvodi kolekciju vrijednosti iz istog domena podataka: reduce (k, list (v)) → list(v' ').

MapReduce paradigma je pronašla značajnu primjenu u nerelacionim bazama podataka, gde se koristi za generisanje kompleksnih izvještaja, umjesto SQL upita. Simbioza MapReduce pristupa i nerelacionih baza je prirodna, jer

MapReduce forsira paralelno izvršavanje, na više radnih stanica, a ne-relacini sistemi za upravljanje podacima su uglavnom dobro optimizovani za rad u klasteru, sa particionisanim podacima, tako da se opterećenje izvršavanja kompleksnih upita ravnomerno raspoređuje na sve čvorove.

D. Dosadašnja rješenja i izazovi

Hadoop-ov HDFS (Hadoop Distributed File System) distribuirani sistem podataka i MapReduce tehnologija su osmišljeni kako bi se efikasno mogle obraditi velike količine podataka korišćenjem više računara vezanih u klaster. Trenutni izazovi su razvoj novih tehnologija za distribuirano izvršavanje analiza online procesiranja umjesto paketnog (batch processing) centraliziranog procesiranja kao efektivni način brze obrade velikih senzorskih podataka za stvarnu real time obradu. Procesiranje događaja u velikim tokovima sa ugrađenim tehnikama mašinskog učenja ima za cilj omogućavanje brzog učenja i odlučivanja bez potrebe za čuvanjem i agregiranjem dolazećih podataka. Razmatraju se mogućnosti brzih SQL upita bez pokretanja suvišnih MapReduce poslova, zatim identifikacija važnih događaje, za brzu obradu i odluke. Rješenja in-memory baza podataka, koji mogu obavljati visoke analitičke i transakcijske obrade predstavljaju velike kompanije kao što su SAP HANA, SAP Sybase Stream [8], Oracle Exalytics In-Memory Machine, SQL Connector for Hadoop [9], Microsoft[10], IBM[11], Teradata[12]. Neke od platformi za obradu toka podataka u realnom vremenu su: Aurora [13], Storm [14], Dryad [15], StreamCloud [16]. Primjeri upotrebe NoSQL baza o kojima se govori u narednom dijelu su Google i Amazon sa svojim BigTable i Dynamo DB, zatim Facebook Cassandra DB, Linkedin Voldemort DB.

III. NOVI OBLICI BAZA PODATAKA

U slučajevima kada relacione baze podataka ne mogu da osiguraju dovoljno dobre performanse pri radu sa ogromnom količinom heterogenih podataka koriste se novi oblici baza podataka. NoSQL definicija [17]: "Sledeća generacija baza podataka koja se odnosi na svojstva: ne-relacione, distribuirane, otvorenog koda i horizontalno skalabine". Često se navode i sljedeće karakteristike: slobodne sheme (schema-free), jednostavna podrška replikaciji, jednostavni API, BASE (ne ACID, eventually consistent), ogromne količine podataka itd. Termin NoSQL je dosta širok i obuhvata raznovrsne baze podataka, zasnovane na različitim arhitekturama i tehnologijama, za koje je ipak moguće izvući zajedničke karakteristike, koje ih diferenciraju od relacionih baza. Ono što je zajedničko svim nerelacionim bazama podataka jeste da ne počivaju na relacionom modelu i prilagođene su radu sa velikom količinom nestruktuiranih podataka [18].

A. Relacione baze podataka [19] podržavaju ACID svojstva transakcija:

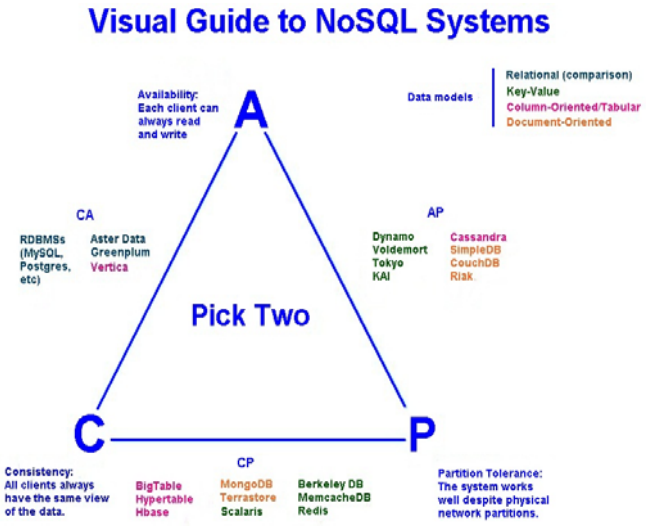
1) *Atomnost (Atomicity),*
2) *Konzistentnost(Consistency),*
3) *Izolaciju (Isolation),*
4) *Trajnost (Durability).*

B. NoSQL baze podataka prate BASE svojstva [20]:

1) *Basically Available* (raspoloživost- većina podataka je dostupna veći dio vremena)
2) *Soft state* (ne mora biti konzistentna u svakom trenutku)
3) *Eventually consistent.* (teži se vremenskoj tački u kojoj će svi čvorovi imati konzistentne podatke)

C. Za NoSQL baze podataka prema CAP teoremi (2000: Eric Brewer) samo dva od sljedeća tri aspekta mogu biti garantovana u isto vrijeme u distribuiranom sisemu [21]:

1) *Consistency (Konzistentnost)* - Postoji uređeni redosljed u kojem se sve operacije izvršavaju, tako da svaka operacija proizvodi efekt kao da se izvršava momentalno, odnosno u jednom trenutku vremena. Ovo je ekvivalentno zahtjevu da se distribuirana, dijeljena memorija ponaša kao da se nalazi na jednom čvoru, i da obrađuje zahtjeve jedan za drugim. Bitna karakteristika ovakve dijeljene memorije je da će svaka operacija čitanja, koja se izvršava nakon operacije upisivanja, vratiti rezultat koji je proizvela operacija upisivanja.
2) *Availability (Dostupnost)*- Podaci uvijek moraju biti dostupni. Pod dostupnošću se smatra da servis mora biti u svakom trenutku operativan i u stanju da obrađuje zahtjeve klijenata. Pored toga, svaki zahtjev, koji primi neki čvor sistema, mora rezultovati odgovarajućim odgovorom, odnosno svaki algoritam, koji obrađuje zahtjev, mora imati ograničeno vrijeme izvršavanja i na kraju mora proizvesti adekvatan odgovor.
3) *Partition Tolerance (Otpornost na particionisanje)*- Baza podataka radi normalno i u slučaju ispada u mreži ili računaru. Otpornost na particionisanje se odnosi na sposobnost sistema da funkcioniše u uslovima mrežnih otkaza, odnosno kada postoje problemi u komunikaciji između čvorova sistema. Prema [21], Gilbert i Linč definišu otpornost na particionisanje na sledeći način: “Nijedan skup otkaza, osim otkaza kompletne mreže ne smije da prouzrokuje nepravilno funkcionisanje sistema”.



Slika 1. CAP teorema i NoSQL rješenja [22]

Suština Bruverove teorme je da ne mogu istovremeno biti zadovoljena sva tri zahtjeva u potpunosti, odnosno nešto se mora žrtvovati, kako bi se dobila poboljšanja na nekom drugom polju, i tu premisu treba imati na umu prilikom projektovanja sistema. Na sl. 1. je prikazan odnos CAP teoreme i primjera NoSQL baza podataka.

D. Kategorije NoSQL (nerelacionih) baza podataka

1) Key-Value stores (Ključ-vrijednost) - Ovaj model možemo uporediti sa tabelom u relacionom modelu koja ima dvije kolone, ključ i vrijednost. Podaci se čuvaju u distribuiranim heš mapama, gde ključ najčešće predstavlja neki string, a vrijednost može biti neki od tipova koje podržavaju svi programski jezici, kao što su stringovi, brojevi, nizovi ili objekti. Ove baze podataka čuvaju raznorodne podatke, ali ne rade nikakve dodatne pretrage podataka po više kriterijuma. Najpopularniji predstavnici: Riak, Redis, Memcached, Amazon DynamoDB, Voldemort DB.

2) Column-oriented databases (Orijentisane prema kolonama.) - Za razliku od relacionih baza podataka, u kojima se podaci grupišu kao redovi, ovdje se podaci grupišu kao kolone, čime se dobijaju bolje performanse kada postoji potreba za upitima koji treba da vrate samo određene atribute, a ne kompletne entitete. U ovim bazama podataka imamo pojmove column i super column. Kolone imaju ime vrijednost i timestamp, znači nemamo potrebu definirati shemu. Najveće prednosti ove vrste baze podataka su brzina i skalabilanost. Ograničenja su složeni upiti, transakcije, postavljanje ograničenja. Najpopularniji predstavnici: BigTable, Hbase, Sybase IQ, Cassandra DB.

3) Document-oriented (Dokumentno orijentisane) - Podaci su organizovani kao kolekcije dokumenata, koji mogu imati različitu strukturu, čime je podržano jednostavno dodavanje i izbacivanje atributa. Ove baze čuvaju XML, JSON, BSON formate dokumenata. Podaci nisu normalizovani. Najpopularniji predstavnici: Apache CouchDB, MongoDB, OrientDB, Terrastore DB.

4) Graph databases (Bazirane na grafovima)- Podaci se predstavljaju u formi grafova, tako što su entiteti predstavljeni čvorovima, a njihove relacije ivicama grafa. Svaka veza i čvor nose određenu informaciju na osnovu kojih se mogu postavljati brzi upiti. Pretraživanje podataka po vezama pokazuju velike prednosti performansi. Ograničenja mogu predstavljati količine podataka odnosno maksimalan broj čvorova. Najpopularniji predstavnici: Neo4J, FlockDB, AllegroGraph, InfiniteGraph, VertexDB.

5) Object oriented databases (Objektne baze)- Podaci se čuvaju kao objekti, što je potpuno u skladu sa objektno-orijentisanom filozofijom. Na ovaj način se eliminiše potreba za konvertovanjem podataka iz objektnog u relacioni model, svaki put kada se podaci čitaju ili upisuju u bazu. Najpopularniji predstavnici: db4o, ObjectStore, GemStone/S. Postojanje određenih razlika između graf baza podataka s jedne i ključ-vrijednost, dokument, kolone orjentisane s druge iniciralo je u nekim radovima da graf baze podataka nisu navedene među NoSQL baze podataka.

Prisutne su podjele NoSQL baza [17] i na:

6) Višemodelne baze podatka (Multimodel Databases) - ArangoDB, OrientDB, Datomic, FatDB, AlchemyDB.

7) Multidimenzione baze podataka (Multidimensional Databases)- Intersystems Cache, GT.M, SciDB, MiniMDB, rasdaman.

8) Viševrijednosne baze podataka (Multivalue Databases) U2, OpenInsight, TigerLogic PICK, Reality, OpenQM, Model 204 Database, ESENT, jBASE.

9) Baze podatka nad mrežom i u oblaku (Grid & Cloud Database)- GigaSpaces, GemFire, Infinispan, Queplix, Hazelcast.

10) XML Databases - EMC Documentum, xDB, eXist, Sedna, BaseX, Qizx, Berkeley DB XML.

E. Prednosti i nedostaci relacionih i nerelacionih baza podataka

Kada se radi sa ogromnom količinom podataka nerelacione baze podataka često pokazuju bolje performanse nego relacione, sa istim hardverom i količinom podataka [23]. Nerelacione baze podataka bolje skaliraju, jer su prilagođene za rad u distribuiranom okruženju. Ukoliko se pojavi potreba, korisnici mogu lako dodavati nove, jeftine mašine u klaster, dok bi kod relacionih baza povećanje performansi prouzrokovalo značajno veću investiciju (dodavanje memorije, diskova). Značajno je teže osigurati konzistentnost i integritet podataka kod NoSQL baza (eventually consistency model), nego što je slučaj kod relacionih baza zahvaljujući ACID karakteristikama. Kod relacionih baza podataka struktura i podaci su rigidni, dok kod NoSQL baza podataka shema je fleksibilna (shema free) što značajno olakšava i pojednostavljuje dodavanje novih atributa ili promjenu tipa atributa. Relacione baze su dugo prisutne i dobro provjerene na tržištu te postoji veliki broj specijalista, kao i alata za njihovo održavanje. NoSQL baze podataka koriste specifične jezike za operacije, koji su često prilagođeni domenu i modelu kojeg baza podataka podržava. NoSQL baza podataka u nekim slučajevima nisu najbolje prilagođeni za kompleksne upite. Evaluacija postojećih NoSQL baza podataka podrazumijeva dobro poznavanje karakteristika i funkcionisanja pojedinačnih vrsta baza, zajedničkih karakteristika ovih baza, kao i specifičnosti koje ih čine različitim. U zavisnosti od korisničkog zahtjeva mogu se razmotriti grupe NoSQL baza podataka koje će najbolje odgovarati traženom zadatku. Pravci daljeg razvoja baza podataka idu u skladu sa zahtjevima okruženja.

IV. ZAKLJUČAK

Novo okruženje, sa uslugama u računarskom oblaku (cloud services), web podaci i korišćenje društvenih mreža u poslovne svrhe, smart objekti na internetu (Internet of Things), uslovili su razvoj novih rješenja upravljanja podacima. Sistemi za Big Data, memorijske baze podataka, obrada toka podatka, kompleksni algoritmi za procesiranje podataka, NoSQL baze podataka, sistemi za inteligentno poslovno odlučivanje daju nove mogućnosti. Potreba da se radi efikasno sa velikom količinom podataka zahtijeva nove tehnologije i standarde.

Informacije sa novih izvora podataka kao što su senzori treba učiniti vrijednim, sa mnoštvom inteligencije i integrisati ih u poslovne sisteme za stvaranje novih vrijednosti. U ovom radu predstavljeni su neki od prepoznatih i grupisanih tehnologija za upravljanje podacima u realnom vremenu kao što su upravljanje memorijskim podacima, obrada podataka u toku, streaming analitika, bihevioralna analitika, prediktivna analitika, kompleksno procesiranje događaja. Zatim su predstavljene NoSQL baza podataka prema podjelama koje se najčešće susreću u literaturama. NoSQL baze podataka su nastale iz potreba za čuvanjem i obradom velikih količina heterogenih podataka pri čemu garantuju konstantne performanse. Na osnovu predstavljenih karakteristika i prednosti NoSQL baza podataka, zajedničkih karakteristika i specifičnosti koje ih čine različitim, moguće je izvršiti evaluaciju ponuđenih baza te razmotriti najbolji izbor u skladu sa korisničkim zahtjevom. Kako i jedne i druge vrste baze podataka (relacione i NoSQL) pokazuju prednosti i nedostatke jedno od mogućih rješenja jeste integracija heterogenih baza podataka. Ovo rješenje zahtijeva detaljno planiranje načina i troškova, razvijanja, održavanja, neophodnih resursa za uspješno funkcionisanje ovakvog rješenja. Za krajnjeg korisnika rješenje bi trebalo biti predstavljeno kao jedinstvena baza podataka, dok se zapravo na nižim slojevima sastoji od više raznorodnih baza. Dalja istraživanja odnose se na predstavljanje rezultata korišćenja nove tehnologije za obradu podatak u realnom vremenu na izabranoj platformi, bazi podataka i analitičkim alatima u određenom poslovnom domenu.

LITERATURA

[1] S. Kaisler, F. Armour, J. Espinosa, J. and G.Washington,"Big Data: Issues and Challenges Moving Forward", 46th Hawaii Conference on System Sciences, IEEE Computer Society, pp 995–1004, 2013.

[2] M. Hogan,: Cloud Computing & Databases, ScaleDB Inc., 2008.

[3] D. Giusto, A. Iera, G. Morabito, L. Atzori (Eds.), The Internet of Things, Springer, 2010. ISBN: 978-1-4419-1673-0.

[4] D. Evans. The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. 2011., [Online]: http://postscapes.com/ cisco-internet-of-things-white-paper-how-the-next-evolution-of theinternet - is-changing-everything%20 [Accessed 10. 2014].

[5] InternetOfThings [Online]. https://www.ida.gov.sg/~/media/Files/ Infocomm%20Landscape/Technology/TechnologyRoadmap/ InternetOfThings.pdf [Accessed 06. 2014.]

[6] R. Addy. Emerging Technology Analysis: Predictive Support Services. [Online] Available from: http://www.gartner.com/id=1875816 [Accessed 10. 2014].

[7] J. Dean, S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google Inc., 2004.

[8] SAP. SAP HANA redefines In-Memory. [Online] Available from: http://www.sap.com/solutions/technology/in-memory-computing-platform/index.epx [Accessed 10. 2014].

[9] M. Rittman , In-Memory Big Data Analysis with Oracle Exalytic, Oracle Openworld 2012, San Francisco, 2012.

[10] Microsoft Research-Database Group Projects [Online] Available from: http://research.microsoft.com/en-US/groups/db/projects.aspx [Accessed 10.2014]

[11] IBM. IBM solidDB – Fastest Data Delivery. [Online] Available from: http://www-01.ibm.com/software/data/soliddb/ [Accessed 09. 2014].

[12] Teradata. Teradata Database. [Online] Available from: http://www.teradata.com/products-and-services/database/teradata-14/ [Accessed 09.2014].

[13] D. Abadi, D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, M. Stonebraker, N. Tatbul, S. Zdonik, Aurora: a new model and architecture for data stream management. The VLDB Journal 2003;12 (2):120–139.

[14] Storm Project S. Storm: Distributed and fault-tolerant realtime computation. http://storm-project.net/2012.

[15] M. Isard, M. Budiu, Y. Yu, A. Birrell, D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks." 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems (EuroSys), EuroSys '07, 2007;59–72

[16] V. Gulisano, R. Jimenez-Peris, M. Patino-Martinez, C. Soriente, and P. Valduriez, "Streamcloud: An elastic and scalable data streaming system," Parallel and Distributed Systems, IEEE Transactions on, vol. 23, no. 12, pp. 2351–2365, 2012.

[17] NoSQL datebase, http://nosql-database.org/ [Accessed 08. 2014]

[18] N. Leavitt, "Will NoSQL Databases Live Up to Their Promise?", Technology News, IEEE Computer Society, 2010.

[19] B. Lazarević, Z. Marjanović, N. Aničić, S. Babarogić, "Baze podataka", Fakultet organizacionih nauka, Beograd, 2003.

[20] P.J. Sadalage, M. Fowler, (2012). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Crawfordsville, Indiana: Pearson Education.

[21] S. Gilbert, N. Lynch, "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services", ACM SIGACT (Volume 33 Issue 2), 2002.

[22] The CAP theorem and the design of large scale distributed systems: Part I, Silvia Bonomi University of Rome "La Sapienza", www.dis.uniroma1.it/~bonomi 2012/13.

[23] U. Bhat, S. Jadhav, "Moving Towards Non-Relational Databases", International Journal of Computer Applications (Volume 1 – No. 13), 2010.

ABSTRACT

The need for development of data management new technologies is conditioned by several factors: working environment with large amounts of data (known as "Big Data"), constantly increasing number of services in the computer cloud (Cloud Computing) and growing number of "smart" objects in the Internet (Internet of Things). In cases when relational databases cannot provide good performance when they operate with vast amount of heterogeneous data, another types of databases are considered, such as NoSQL database. New types of databases should provide high availability, processing speed, security and privacy, reliability, consistency, scalability and distribution possibility. Data which are created in an interaction between devices (M2M - machine-to-machine interaction) and the Internet of Things (IoT - Internet of Things), which need to processed in real time, represent the increasing challenges for the future period. New opportunities can be found in using technology for fast data processing in real time, memory analytics and data flow analytics. Business intelligence has enabled users to obtain crucial information from the enterprise business data and use them as a base for making business decisions. Certain reactivity requires advances in analytical technologies of predictive analytics, contextually-aware services, behavioral analytics and complex processing of events.

DATA MANAGEMENT TECHNOLOGIES
Data processing in real-time and analytical technologies with new types of databases

Daliborka Macinkovic