Student: Jovan Mladenović 1277

Profesor: Aleksandar Stanimirović

Distribuirane baze podataka - Cassandra

Sadržaj

[1. Uvod 3](#_Toc106829544)

[2. Arhitektura 3](#_Toc106829545)

[2.1. Particioniranje podataka - konzistentno haširanje 4](#_Toc106829546)

[2.2. Konzistentno haširanje korišćenjem token prstena 4](#_Toc106829547)

[2.3. Višestruki tokeni na jednom fizičkom čvoru (vnodes) 5](#_Toc106829548)

[3. Verzioniranje podataka i podesiva kozistentnost 6](#_Toc106829549)

[3.1. Strategija replikacije 6](#_Toc106829550)

[3.2. Verzioniranje podataka 7](#_Toc106829551)

[3.3. Podesiva kozistentnost 7](#_Toc106829552)

[4. Člansto distribuiranog klastera i detekcija otkaza 8](#_Toc106829553)

[4.1. Tračarenje (gossip) 8](#_Toc106829554)

[4.2. Članstvo i detekcija otkaza 8](#_Toc106829555)

# Uvod

Apache Cassandra je započela kao projekat unutar Facbook-a za brzu i laku pretragu konverzacija i drugog sadržaja koji je zanimao korisnike. Arhitektura je bila kombinovana sa modelom po kome je baziran Amazonovog Dynamo koji dozvoljava skalabilnost. Rezultat je bio visoko skalabilana baza podataka koja je mogla da izdrži najzahtevnije slučajeve korišćenja. Danas, Cassandra, je besplatna za upotrebu pod Apache 2.0 licencom. Kasandra nudi *Cassandra Query Language* (CQL), jezik slišan SQL-u za kreiranje baza pristup podacima. CQL omogućava korisnicima da organizuju podatke unutar klastera korišćenjem:

* **Keyspace** – definiše kako se dataset replicira unutar datacentra. Replikacija je broj kopija sačuvanih unutar klastera.
* **Table** – definiše šemu kolekcije unutar particija. Tabele sadrže redove, koji sadrže kolone. Tabele mogu dodavati kolone bez prekida rada.
* **Particije** – definiše obavezni deo primarnog ključa. Svi upiti u sebi sadrže particioni ključ.
* **Red** – sadrži kolekciju kolona identifikovanih primarnim ključem i opcionim dodatnik kluster ključevima.
* **Kolona** – jedan podatak koji pripada redu

Kasandra eksplicitno bira da ne implementira operacije koje zahtevaju koordinaciju particija jer su spore i teške za globalno održavanje konzistentnosti. Na primer Kasandra ne podržava Transakcije izmedju particija, distribuirane join-ove, strane ključeve. Neke od glavnih osobina kasandre su:

* **Open** **source**
* **Fleksibilan i prepoznatljiv interfejs** – Kasandrin upitni jezik je jako sličan SQL-u, što olakšava prelazak većini programera
* **Visoke performanse** – svaki čvor može da obavlja sve operacija čitanja i upisa za razliku od tradicionalnih SQL sistema
* **Zero downtime** – ako neki čvor otkaže, korisnik se upućuje na sledeći dostupan čvor
* **Skalabilnost** – kasandra dozvoljava skaliranje prostim dodavanjem više čvorova
* **Efikasna replikacija**

## Arhitektura

Kasandra se oslanja na nekoliko tehnika preuzetih od Amazonovog Dynamo distribuiranog key-value sistema. Konkretno, Kasandra se bazirano na Dynamu oslanja na:

* Particioniranje podataka korišćenjem konzistentnog haširanja.
* Multi-master replikaciju korišćenjem verzioniranih podataka
* Distribuirano članstvo klastera i detekcija otkazivanja bazirana na protokolu tračarenja (gossip)
* Ikrementalno skaliranje na pristupačnom harveru

### Particioniranje podataka - konzistentno haširanje

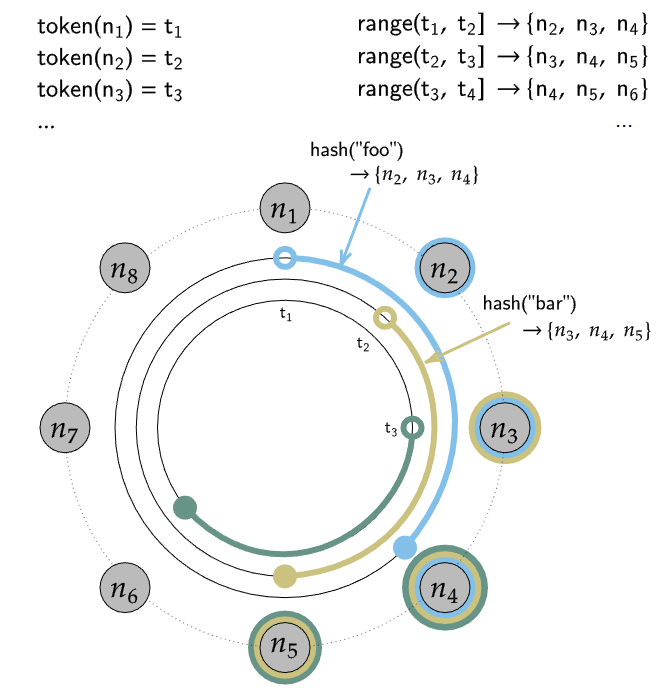
Kasandra ostvaruje horizontalnu skalabilnost particioniranjem svih podataka sačuvanim unutar sistema korišćenjem haš funkcije. Svaka particija se replicira na više fizičkih čvorova unutar klastera. Kako svaka replika prihvata izmenu podataka koju poseduje, svaki ključ mora biti verzioniran. Kasandra za razliku od Dynamo-a, koji je koristio vektorske časovnike, ima prostiji sistem – poslednji upis pobeđuje, gde svaki ključ ima svoji timestamp i poslednja verzija podatka je pobednik.

### Konzistentno haširanje korišćenjem token prstena

Kasandra particionira podatke unutar čvorova korišćenjem specijalne vrste haširanja koje se zove conzistentno haširanje. Kod obišnog data haširanja tipično se alociraju ključevi particiji uzevši haš datog ključa po modulu broja čvorova.

Kasandra umesto toga mapira svaki čvor na jedan ili više tokena na kontinualnom haš prstenu, i odlučuje o vlasništvu nad ključem haširanjem ključa na prsten i „kružeći“ po prstenu u jednoim smeru. Glavna razlika izmedju konzistentnog i naivnog haširanja je kada se broj čvorova promeni, konzistentno haširanje mora da preseli samo mali deo ključeva.

Na primer, ako imamo klaster sa osam čvorova sa podjednako distanciranim tokenima i replikacioni faktor od tri, da bi našli koji čvorovi poseduju neki ljuč mi prvo generišemo haš tog ključa i zatim idemo po prsteu u smeru kazaljki na satu dok ne dodjemo do tri različitih čvorova koji sadrže replike traženog ključa. Na slici 3.2.1 vidimo da u Dynamo sistemu, raspon ključeva (poznat kao token ranges), se mapiraju na isti skup fizičkih čvorova. U ovom primeru svi ključevi koji ne spadaju u rasponu tokena 1 i spadaju u raspon tokena 2 (grange(t1, t2]) se pamte na čvorovima 2, 3 i 4.



Slika 3.2.1 – primer klastera sa osam čvorova i replikacionim faktorom tri

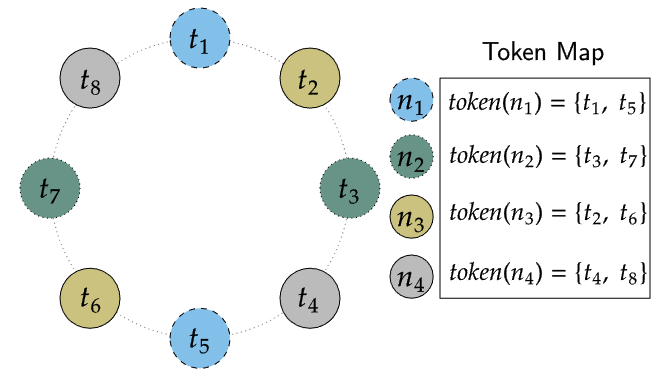
### Višestruki tokeni na jednom fizičkom čvoru (vnodes)

Konzistentno haširanje nad jednim tokenom funkcioniše dobro ako imamo mnogo fizičkih čvorova za raspodelu podataka, međutim ako imamo mali broj fizičkih čvorova sa podjednako distanciranim tokenima inkrementalno skaliranje (dodavanje samo par čvorova) je teško jer nema tokena za izbor koji bi ostavili prsten da bude balansiran. Kasandra izbegava nebalansirane tokene jer nejednak balans tokena vode do nebalansiranog opterećenja zahtevima. Na primer u primeru sa slike 3.2.1 dodavanje novog tokena bez izazivanja nebalansa.

Dynamo pristup zahteva koršćenje „virtuelnih čvorova“ kako bi se rešio ovaj problem. Virtuelni čvorovi rešavaju ovaj problem dodeljivanjem više tokena po jednom fizičkom čvoru. Dodavanje višestrukih tokena jednom fizičkom čvoru unutar prstena male klastere činimo većima i dodavanjem samo jednog fizičkog čvora deluje kao da smo dodali mnogo. Kasandra definiče nekoliko stavki kako bi se shvatio ovaj koncepti:

* **Token** – jedna pozicija na dynamo haš prstenu
* **Endpoint** – jedan fizički IP i port na mreži
* **Host ID** – jedinstveni identifikator za jedan fizički čvor , nalazi se na jednom endpointu i sadrži nekoliko Tokena
* **Virtuelni čvor** – token na haš prstenu u vlasništvu jednog fizičkog čvora

Mapiranje tokena se nalazi unutar **Token Map** mape gde Kasandra prati koja se pozicija na prstenu mapira na koji fizički čvor. Primer sa slike 3.3.1 predstavlja klaster sačinjen od četiri fizička čvora dodeljivanje svakom po dva tokena.



Slika 3.3.1 – klaster sačinjen od četiri fizička čvora dodeljivanje svakom po dva tokena

Višestruki tokeni po fizičkom čvoru imaju sledeće benefite:

* Kada se doda novi čvor on prihvata približno istu količinu podataka od drugih čvorova što dovodi do podjednake distribucije podataka unutar klastera.
* Kada se isključi čvor gubi količinu podataka približno jednakoj količini na drugim čvorovima, ponovo održavajući podjednaku raspodelu
* Ako švor postane nedostupan, opterećenje upita se distribura po više drugih čvorova

Međutim, ima i svojih nedostataka

* Svaki token dodaje i do *2 \* (Replikacioni faktor - 1)* dodatnih „komšija“ na prsten, što dovodi do većeg broja kombinacija otkaza čvorova gde gubimo pristup delu token prstena.
* Održavanje na nivou klastera je usporeno. Na primer ako se poveća broj tokena po čvoru, broj diskretnih operacija popravke koje klaster mora da obavlja se povećavaju
* Operacije koje zahvataju veliki opseg tokena mogu biti usporene

# Verzioniranje podataka i podesiva kozistentnost

Kasandra vrši replikaciju svake particije na više čvorova unutar klastera kako bi održala dostupnost i izdržljivost. Kada dođe do promene podataka, vrži se kaširanje particionog ključa kako bi se odredio raspon tokena kome podatak pripada, zatim vrši replikaciju promene na osnovu strategije repliciranja.

Sve strategije replikacije imaju faktor replikacije (RF), koji govori Kasandri koliko kopija particije treba da postoje. Na primer ako je RF = 3, podaci se upisuju u tri različite replike. Replike su uvek takve da se nalaze na različitim fizičkim čvorovima. Strategije replikacije mogu takođe da odluče da preskoče čvorove koji se nalaze u istom domenu otkaza kao što su na primer isti rafovi ili data centri.

## Strategija replikacije

Svaki keyspace ima svoju strategiju replikacije. Sve produkcije bi trebale da koriste *NetworkTopologyStrategy* dok za testiranje klastera dok raspored data centara nije poznat bi trebala da se koristi strategija *SimpleStrategy.*

* **NetworkTopologyStrategy** - zahteva da se specificira faktor replikacije za svaki data centar unutar klastera, čak i ako klaster ima samo jedan data centar. NetworkTopologyStrategy je preporučen jerolakšava dodavanje novih fizičkih ili virtuelnih data centara u klaster. Takođe ako je broj rafova veći ili jednak faktoru replikacije NetworkTopologyStrategy garantuje kreiranje replika untar data centra na različitim rafovima specificiranom od strane izdajice (Snitch-a).
* **SimpleStrategy** - se odnosi prema svim čvorovima identično, ne uzimajući u obzir konfigurisane data centre ili rafove.

## Verzioniranje podataka

Kasandra koristi mutaciono timestamp verzioniranje za garantovanje eventualnu konzistentnost podataka. Kada god izmena udje u sistem ima sa sobom timestamp klijentskog časovnika, ako ne onda od dobija timestamp od šasovnika koordinatorskog čvora. Ažuriranja se razrešavaju po principu poslednji upis pobeđuje.

Kako repike unutar Kasandre mogu nezavisno da prihvataju izmene, moguće je da neke replike imaju novije verzije od drugih. Kako bi se garantovala eventualna kozistentnost Kasandra implementira **anti-entropy repair <repair>** gde replike kalkulišu hijerarhijska haš stabla na osnovu svojih data setova (Markle stabla) koja se mogu upoređivati između replika kako bi identifikovali podatke koji se ne poklapaju, kao što je u Dynamo-u kod Kasandre replike generišu periodično Markle stabla i razmenjuju ih međusobno i sinhronizuju podatke koji se ne poklapaju. Za razliku od Dynamo-a, Kasandra, implementira takođe **sub-range repair** i **incremental repair.** Sub range repair dozvoljava Kasandri da poveća rezoluciju haš stabla (čak do nivoa particije) kreirajući viče stabla koja obuhvataju porcije celog skupa podataka. Incremental repair dozvoljava Kasandri da popravli particije koje su se promenile od poslednje popravke.

## Podesiva kozistentnost

Kasandra podržava „po poeraciji“ razmenu između konzistentnosti i pristupašnosti pomoći **Nivoa konzistentnosti.** Nivoi konzistentnosti su verzija Dynamo-ovog *R + W > N* mehanizma konzistentnosti koji određuje da broj čvorovia koji izvršavaju čitanje (R) i broj švorova koji izvrčavaju upis (W) mora biti veći od faktora replikacije (N). U Kasandri se međutim bira iz niza nivoa konzistentnosti koji dozvoljavaju operatoru da bira ponašanje R i W bez poznavanja faktora replikacije. Postoje:

* ONE – samo jedna replika mora da odgovori
* TWO – dve replike moraju da odgovore
* THREE – tri replike moraju da odgovore
* QUORUM – većina (n/2 + 1) replika moraju da odgovore
* ALL – sve replike moraju da odgovore
* LOCAL\_QUORUM – većina replika u lokalnom datacentru (u kome god datacentru se nalazi koordinator) moraju odgovoriti.
* EACH\_QUORUM – većina replika u svakom data centru moraju da odgovore
* LOCAL\_ONE – samo jedna replika mora da odgovori, unutar višestrukih data centar klastera ovo garantuje da se zahtev ne šalje na remote datacentre

Operacije upisa se uvek šalju svim replikama, bez obzira na nivo konzistentnosti.

# Člansto distribuiranog klastera i detekcija otkaza

Protokoli replikacije i particioniranje podaataka se oslanja na informacije koji čvorovi su živi a koji mrtvi unutar klastera kako bi se operacije čitanja i upisa optimalno rutirale. U Kasandri informacije od ovome se dele distribuirano kroz mehanizam detekcije otkaza baziranom na protokolu tračarenja (gossip).

## Tračarenje (gossip)

Tračarenje je kako Kasandra propagira osnovne informacije o klasteru. U kasandri čvorovi razmenjuju informacije ne samo o svom stanju već i o čvorovima za koje oni imaju informacije. Ove informacije se verzioniraju časovnim vektorom *[generation, version],* gde je generacija je timestamp i verzija je logički sat koji se uvečava svake sekunde. Ovaj par časovnika dozvoljava Kasandrinom tračarenju da ignoriše stare verzije stanja klastera. Svaki čvor unutar klasterra izvršava zadatak trtačarenja nezavisno i periodično. Svake sekunde svaki čvor unutar klastera:

* Ažurira lokalno heartbeat stanje (verziju) i konstruiše čvorov lokalni pogled na klastera.
* Bira nasumično drugi čvor sa kojim razmenjuje trač.
* Pokušava da komunicira se nedostupnim čvorem, ako takav postoji

## Članstvo i detekcija otkaza

Tračarenje formira osnovu članstva u prstenu, ali **detektor otkaza** na kraju odlučuje da li su čvorovi operativni ili ne. Svaki čvor unutar Kasandre pravi nezavisne provere i odluke da li je neki od drugih čvorova aktivan. Odluka se primarno bazira na otkucaj srca susednih čvorova. Na primer ako čvor ne vidi inkrementalni otkucaj srca drugog čvora, posle nekog vremena detektor otkaza odlučuje da je čvor otkazan i Kasandra više neće rutirati zahteve prema tom čvoru. Ako čvor ponovo počne da šalje otkucaje, Kasandra će pokušati da kontaktira taj čvor i ako uspe da otvori komunikacioni kanal smatraće čvor ponovo aktivnim.

Kasandra nikad neće ukloniti čvor iz tabele stanje tračarenja osim ako ne dobije eksplicitno naređenje od operatora ili ako se doda novi čvor sa *replace\_address\_first\_boot* opcijom.

# Konfiguracija Kasandre

Konfiguracioni fajlovi za Kasandru se nalaze u folderu **/etc/cassandra.** Glavni konfiguracioni fajl koji sadrži najbitnija podešavanja, o kome će biti reč u ovom dokumentu, je **cassandra.yaml**. Glavne tačke za podešavanje su:

**cluster\_name** – ima klastera, svi članovi klastera moraju imati isto ime, može biti bilo šta.

Num\_tokens – broj virtuelnih čvorova unutar kassandra instance, najčešće korišćena vrednost je 256.

Seeds – ip adrese klaster seed čvorova. Seed čvorovi služe kao poznata tačka za dobijanje informacija o klasteru (kao što je na primer lista čvorova). Oni su poznate lokacije na koje se može osloniti. Preporučeno je da postoje barem tri seed čvorova unutar jednog data centra.

Listen\_adress – ip adresa koju će Kasandra osluškivati za unutračnju komunikaciju (Kasandra prema Kasandri). Kasandra će nagađati adresu ako se polje otavi prazno ali je preporučeno da se definiše ručno. Ovaj podatak je specifičan za svaki čvor.

Rpc\_adress – ip adresa koju će Kasandra osluškivati za komunikaciju sa klijentom.

Enpoint\_snitch – predstavlja izdajicu koju Kasandra koristi, ona joj govori koji čvorovi pripadaju kojim data centima i rafovima. Postoje nekoliko vrsta o kojima neće biti govora u ovom dokumentu.

U nastavku su primeri podešavanja cassandra.yaml fajla tri čvorova unutar data centra.

Čvor 1:

cluster\_name: 'Test Cluster'

num\_tokens: 256

seed\_provider:

    - class\_name: org.apache.cassandra.locator.SimpleSeedProvider

        - seeds: 10.0.0.1, 10.0.0.2

listen\_address: 10.0.0.1

rpc\_address: 10.0.0.1

endpoint\_snitch: GossipingPropertyFileSnitch

Čvor 2:

cluster\_name: 'Test Cluster'

num\_tokens: 256

seed\_provider:

    - class\_name: org.apache.cassandra.locator.SimpleSeedProvider

        - seeds: 10.0.0.1, 10.0.0.2

listen\_address: 10.0.0.2

rpc\_address: 10.0.0.2

endpoint\_snitch: GossipingPropertyFileSnitch

Čvor 3:

cluster\_name: 'Test Cluster'

num\_tokens: 256

seed\_provider:

    - class\_name: org.apache.cassandra.locator.SimpleSeedProvider

        - seeds: 10.0.0.1, 10.0.0.2

listen\_address: 10.0.0.3

rpc\_address: 10.0.0.3

endpoint\_snitch: GossipingPropertyFileSnitch

Da bi obezbedili pouzdano repliciranje podataka moramo da kažemo svakom čvoru gde se nalazi, za to menjamo fajl **/etc/cassandra/cassandra-rackdc.properties**, možemo koristiti sopstevene standarde za imenovanje lokacija. Slede redom podešavanj za čvorove 1, 2 i 3:

dc=dc1

rack=rack1

dc=dc1

rack=rack1

dc=dc1

rack=rack2

Zatim možemo pokrenuti kasandru i uključiti da se pokreće sa sistemom komandama

systemctl enable cassandra

systemctl start cassandra

Kako bi videli tsanje klastera koristimo komandu **nodetool status.**

root@scw-cassandra:~# nodetool status

Datacenter: datacenter1

=======================

Status=Up/Down

|/ State=Normal/Leaving/Joining/Moving

--  Address        Load       Tokens       Owns (effective)  Host ID                               Rack

UN  10.0.0.3  119.29 KiB  256          60,7%             c9b13a33-147f-4293-8aaf-21ace6d1b756  rack2

UN  10.0.0.2   170.88 KiB  256          65,3%             2a100701-3da4-444a-892d-164d2222009c  rack1

UN  10.0.0.1    15.47 KiB  256          74,1%             93feee5d-3de8-4c0a-908d-2432f26a1a1e  rack1

Na kraju da bi pristupili klasteru možemo iskoristiti **cqlsh** alat koji je automatski instairan na svim čvorovima. Za ip biramo jednu od rpc adresa čvorova

cqlsh 10.0.0.1

Connected to scw-cassandra01 at 10.0.0.1:9042.

[cqlsh 5.0.1 | Cassandra 3.2.1 | CQL spec 3.4.0 | Native protocol v4]

Use HELP for help.

cqlsh>