7.1. Objasnite zašto se relacijske baze podataka ne smatraju dobrim rješenjem za pohranu velikih podataka (Big Data).

- spajanje tablica je resursno zahtjevno (potreba denormalizacije cime gubimo 3. normalnu formu), nisu dizajnirane za horizontalnu skalabilnost, ocuvanje konzistentnosti postaje usko grlo (visoka razina konzistentnosti = stroga politika zakljucivanja i kontrole pristupa), sekundarni indeksi usporavaju sustav kod cestih izmjena (uklanjanje istih moze znatno usporiti upite)

7.2. Navedite prednosti stupcano-orijentiranih baza podataka

- smanjuje se broj I/O operacija, bolja kompresija podataka, pogodno za tablice sa velikim brojem stupaca, ucinkovito upravljanje rijetkim tablicama, mogu preskociti NULL vrijednosti

7.3. Objasnite koja je od sljedecih tvrdnji tocna:

a) podaci pohanjeni u bazu podataka HBase slijede formalno definiranu shemu,

b) baza podataka HBase dizajnirana je za laku vertikalnu skalabilnost,

c) baza podataka HBase omogucuje nisku latenciju rada s podacima i brz pristup individualnim zapisima i

d) podaci pohranjeni u bazu podataka HBase u pravilu slijede trecu normalnu formu

- b) krivo (horizontalnu), a) krivo (dodaju se obitelji stupaca), d) krivo (denormaliziran oblik) c) tocno

7.4. Objasnite nacin na koji se podaci fizicki pohranjuju kod baze podataka HBase

- HBase je mapa zapravo

- tablica = skup redaka, redak = vrijednosti unutar skupa obitelji stupaca, obitelj stupaca = skup stupaca, stupac = skup parova kljuc-vrijednost

- Podaci se perzistiraju uz pomoc neizmjenjivih datoteka - HFile

- HFile = uredena mapa parova kljuc -> vrijednost poredanih leksikografski po id-u redaka, skup parova je pohranjen u niz blokova, na kraju datoteke se pohranjuje indeks blokova pomocu kojeg HBase pronalazi trazeni zapis

- kod izmjene podataka ne izmjenjuje se HFile, nego se pise prvo u HLog ili WAL (write-ahead log) a zatim se pohranjuje u memoriju (memstore), posluzivanje informacija zdruzivanjem memstora i HFile-a

- kad se memstore prepuni stvara se novi HFile uz pomoc WAL-a

8.1. Napišite psudokod osnovnog algoritma preporuke po suradnji korisnika

- for every object o that user c has no preference for yet

    for every other user c' that has a preference for o

        compute a similarity s between c and c'

        add c's preference for o, weighted by s, into a running average of o

return the top objects, ranked by weighted average

8.2. Kako osnovni algoritam preporuke temeljene na sadržaju možemo napisati u obliku matrica? Što predstavljaju elementi svake od tih matrica?

- U x S = E

[U11 ... U1m] [S11 ... S1m] [e11 ... e1m]

[..... Uix .....] x [..... Sxj .....] = [..... eij .....]

[Un1 ... Unm] [Sm1 ... Smm] [en1 ... enm]

- uix - koliko je korisniku ci koristan objekt ox

- sxj - koliko je objekt ox slican objektu oj

- eij - kolika je procijenjena korisnost objekta oj korisniku ci

8.3. Objasnite kako biste ostvarili preporucivanje temeljeno na suradnji u slucaju velikog broja korisnika.

- pri izracunu preporuke za korisnika c se uzima u obzir samo njemu slicni korisnici, tako se smanjuje broj operacija na 2nm|C'|, inace je 2mn^2

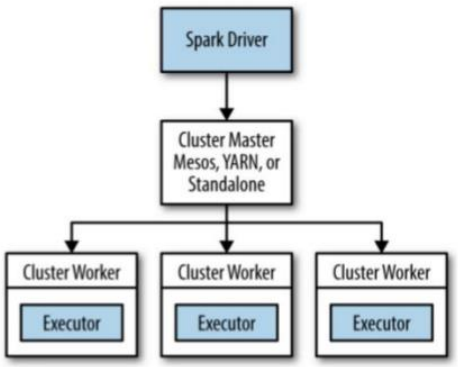
8.4. Objasnite zašto je potrebno raditi normalizaciju matrice slicnosti objekata (S') u slucaju preporuke temeljene na slicnosti objekata (item-based recommendation)

- jer zbog velike kolicine objekata je broj operacija jako velik, te ga je potrebno smanjiti (sta ja znam, lupam bzvz)

9.1. Objasnite zašto kažemo da su osnovni operatori nad tokom podataka blokirajuci. Kako ih možemo odblokirati?

- da blokiraju isporuku rezultata (koja je max vrijdnost beskonacnog slijeda?), mozemo ih odblokiratu uvevsi prozore koji racuna max vrijednost na temelju zadnjih x vrijednosti

9.2. Skicirajte i objasnite raspodijeljeno izvodenje Sparkove aplikacije



Driver je proces u kojem se izvodi glavna metoda aplikacije. Zaduzen za podjelu posla na zadatke. Iz usmjerenog aciklickog grafa operacija stvara plan izvodenja. Koordinira rasporedivanje poslova na izvodace. Vrsi rasporedivanje poslova na izvodace temeljeno na prostornoj lokalnosti.

Izvodaci izvode dodieljene zadatke i u memoriji pohranjuju podatke neophodne za izvodenje aplikacije.

9.3. Objasnite razliku izmedu dviju vrsta operacija nad RDD-om (Resilient Distributed Dataset).

- transformacije - pretvaranja jednog toka u drugi

- izlazne operacije - operacije nad tokom podataka koje daju neki rezultat

9.4. Objasnite nacin na koji Spark obraduje tok podataka prividno stvarnovremeno. Koji parametri su bitni za definiranje vremenskog prozora u Sparkovoj aplikaciji?

- stvarnovremenost je prividna jer se podaci iz tokova obraduju u mikroskupinama

- windowDuration - koliko zadnjih prozora "pamti"

- slideDuration - koliko svakih prozora se pomice

10.1. Navedite i objasnite dva moguca nacina zapisa podataka o strukturi grafa te navedite koji je pogodniji za zapis velikih mreža te objasnite zašto.

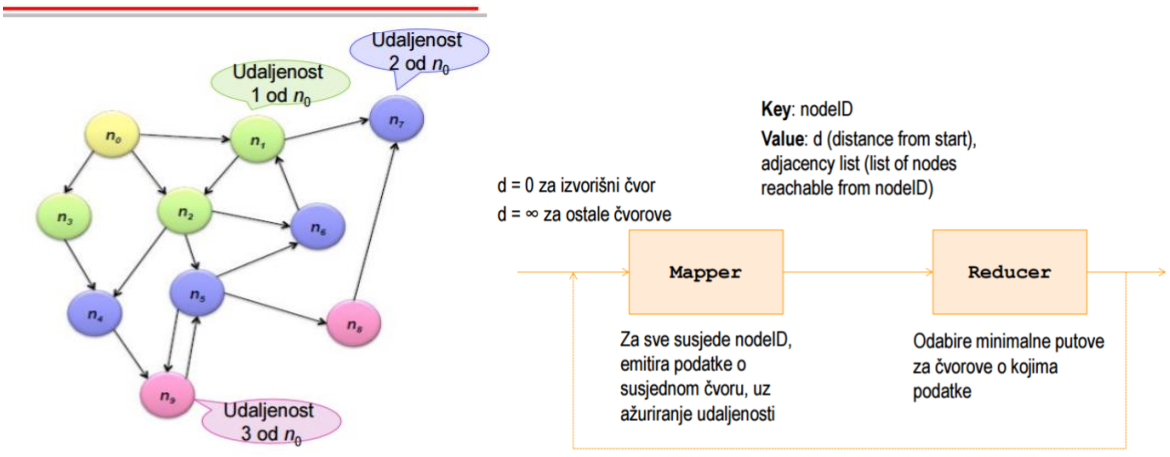
- matrica susjedstva i lista susjedstva, lista je pogodnija jer nemora svaki cvor u svojoj listi imati sve ostale cvorove nego samo one sa kojima je povezan dok je matrica full velicine NxN

10.2. Navedite kako možete odrediti izlazne i ulazne grane pojedinog cvora grafa ako za zapis koristite listu susjedstva

- id cvora ciju listu gledamo je cvor iz kojeg ide strelica, a cvor u njegovoj listi je cvor prema kojem ide strelica

10.3. Skicirajte algoritam za paralelno izvodenje Dijkstrinog algoritam u okolini MapReduce

- u jednoj iteraciji MapReduce analiziraju se cvorovi na udaljenosti 1 od izvorista, zatim na udaljenosti 2 itd



10.4. Objasnite ideju mjeru popularnosti PageRank. Navedite kako je moguce izracunati PageRank na iterativni nacin u okolini MapReduce.

- pageRank - karakterizira kolicinu vremena koju korisnik provede na nekoj web stranici

- mjera popularnosti web-resursa neovisna o txt sadrzaju dokumenta

- ako stranica i ima hiperlink na stranicu j, tada i implicitno prenosi vaznost j

- xi = sum(1/Nj)*xj -> rang stranice i = rang stranice j kroz broj izlaznih linkova iz j

11.1. Što je strojno ucenje?

- strojno ucenje je polje racunalne znanosti koje se bavi specificnim nacinom programiranja u kojem ocekujemo da racunalo samostalno dode do odredenih spoznaja na osnovu dostupnih skupova i odabrane metode ucenja

11.2. Objasnite metodu linearne regresije.

- pretpostavka je da izmedu prediktora i predikcija postoji linearna veza

- trazi se pravac kod kojeg ce zbroj najmanjih kvadrata odstupanja ciljeva od pravca biti najmanji

- pogreska procjene = srednja vrijednost kvadrata odstupanja od pravca

11.3. Ukratko navedite korake u procesu dubinske analize podataka

- prikupljanje i integracija podataka iz razlicitih izvora

- definicija problema kojeg se zeli rijesiti provedbom dubinske analize

- priprema, ciscenje i transformacija dobivenog podatkovnog skupa

- eksploratorna analiza podataka

- odabir metoda za stvaranje prediktivnih i/ili deskriptivnih modela

- treniranje i podesavanje modela

- evaluacija i usporedba modela

- priprema izvjestaja o rezultatima analize

11.4. Navedite prednosti i nedostatke korištenja Spark + MLib arhitekture

- prednosti: veliki podatkovni skup a postojeca analiticka platforma nema dovoljno kapaciteta, snazno integracijski orijentirani, transparentno obavlja zadane zadatke na velikom skupu podataka kao da se radi o lokalnom radu, moze posluziti kao alternativni prirucni stroj za obradu podataka kada se lokalni rezim pokaze nedovoljan, integracija sa sparkom otvara pristup novim funkcionalnostima kao sto su Spark streaming ili stvarnovremensko izvodenje upita nad velikim skupom podataka (Spark SQL)

- nedostaci: nova tehnologija, tesko procijeniti dugorocnu stabilnost API-ja, integracijs-ka rjesenja i dalje nisu na visokoj razini glede jednostavnosti, stabilnosti i transparentnosti, nepotpune funkcionalnosti, problem dugorocne stabilnosti i odrzivosti napisanog programskog koda, integracija sa postojecim rjesenjima se moze pokazati disruptivnom

12.1. Opišite postupak dobivanja malog svijeta iz rešetke modelom Watts-Strogatz. Objasnite koja dva parametra i kako moramo pri tome promatrati da mali svijet ne bi postao slucajna mreža.

- model Watts-Strogatz - veze u resetki se nasumicno prespajaju (vjerojatnost prespajanja p)

- precaci drasticno smanjuju prosjecni najkraci put

- mali svijet - mreza s visokim koeficijentom grupiranja i malim prosjecnim najkracim putom

- moramo promatrati vjerojatnost prespajanja i koeficijent grupiranja (valjda, ne znam)

12.2. Skicirajte proizvoljnu mrežu od 4 cvora te na primjeru jednog cvora navedite formulu objasnite što je koeficijent grupiranja cvora (local clustering coefficient). Kao iz koeficijenta grupiranja cvora možemo dobiti koeficijent grupiranja društvene mreže (network average clustering coefficient)?

- koeficijent grupiranja cvora i, Ci, predstavljaja omjer broja medusobnih veza medu susjedima cvora i, E, i ukupnog broja njihovih mogucih medusobnih veza Ci = 2E / (z^i * ((z^i)-1))

- koeficijent grupiranja nam kaze koliko su gusto povezani prvi susjedi nekog cvora

- koeficijent grupiranja mreze je prosjecni koeficijent grupiranja cvorova u mrezi C = (1/N) * sum(Ci)

12.3. Objasnite zašto je centralnost cvora bitna za širenje informacija u mreži. Kako definiramo centralnost po položaju (betweenness centrality)?

- centralnost pokusava identificirati najvaznije cvorove u grafu

- centralnost po polozaju, centralnost po blizini, po svojstvenom vektoru, po stupnju...

12.4. U RDD-u prVertices se nalazi graf s pagerank-om cvorova društvene mreže (id cvora tipa Long i pagerank cvora tipa Double). Nadopunite sljedeci programski kod da ispišete razdiobu vrijednosti pagerank-a cvorova od veceg prema manjem kao niz redaka u obliku: pagerank cvora, broj cvorova s tim pagerank-om.

```
JavaRDD<Tuple2<Long, Double>> prVertices = getVertexPagerank();
JavaPairRdd<Double, Long> prCounts = prVertices
        .mapToPair(pair -> new Tuple2<>(pair._2, 1))
        .reduceByKey((x, y) -> x+y)
        .sortByKey();
prCounts.foreach( d -> System.out.println("\n" + d._1 + ", " + d._2));
```