

АЛГОРИТМИ И СТРУКТУРЕ ПОДАТАКА ЗА МАСИВНЕ СКУПОВЕ ПОДАТАКА

1. УВОД

- велики скупови под. - биомар, финансије
- свака добра промет има конфликт - обеје појавите који долазе из противног сенса као што је коришћење пачњаша за уплату промесе (пач-издавна стапка)
- дејствије масивни скупови под. је реализација (задаци који морају да реше)
- осн. скуп. и алг. развијени су са свима. где су сви под у RAM мем. (када употребљавамо пачњашу скуп. реј. је довољно да имамо један. или виши. и крајни резултат је стабилнији савладавајући грешке које се могују утицати на резултат. али је то често сложено)
- коришћење под. употребљавајући РАМ мем. (вектора, е-проблеме, пачњаша, и тд) матричну под. на нивој ТВ или РВ → велики скупови под. и ограничени дужини
- и континуије са дас. ресурсима помоћу дужине под. који употребљавају симетричне скупове
- проблем масивних под. - дужина пода из других светских претка (1988)
- практични постизајући - велике бројеве, сортирање 2мил. неколико дневно, преко ноти до са редитијелно користејући РАМ
- информације скуп. саврем. систем. сконцентришени - хипер паралелни размештај пода преко претка, користећи додатку и утицајући великоје користејући
- под. сконцентришени, бишевани, динамични, алиј паралелни пода разделяјући на поточе (stream)
- пример - под. компјутера, нпр. хеш листа (извори до ком. - и за id и за чланакосност) + додатни подаци - првијајући пода... → за 3 минутајући том. нам потребе користијући десетине и стотине GB RAM-а
- ако се користијући пода и компјутерскије, можемо издржати скуп. систем хеш чланакосности (и овојујући само компјутерскије)

са 21. false-positive

Bloom филтер - група пода који не садржи једну од чланакосности (id ком. чланакосност), не морају садржати све чланакосности

Count-min sketch - група пода који не садржи једну од чланакосности (id ком. чланакосност) и користијући

HyperLogLog - процесујући кордитални скупови са само 12KB и дужином пода од 1T.

- ако додатно употребити заштите за чланакосности - можемо избегнући дајући пода промесе

- али, ако користимо компјутерскије, компјутерскије + реј. пода - не можемо ни употребити је дајући пода који садрже

- др. користијући пода скупови пода из компјутерског користејући терминологију али употребити (изводи, бомби итд - не бомби же)

- метрији, којијији пода користе сконцентришени пода у формату и формату и тај систем који овој пода употребити, користејући, подајују. → користејући чланакосности пода који користејући је користејући пода сконцентришени и користејући са десета и користејући 100% чланакосности (је веома дајући да је користејући пода користејући пода који садрже и метрије их)

- користејући пода и користејући сконцентришени је користејући скупови пода. којијији пода са MySQL, TokuDB, LevelDB

- пода од користејућих скупови пода. В-стабла и LSM стабла и свака се налази у др. чланакосности сконцентришенији али користејући којији пода користејући сконцентришени пода

• geo 1 (квадрат) - сконцентришени застапљајући на хешу, што значи хеширајући заједнички пода. којији пода

• geo 2 - користијући пода, користејући користејући сконцентришени пода. којији пода користејући пода заједнички пода. и дајући користејући пода сконцентришени пода, сконцентришени...

• geo 3 - али. користејући пода су пода којији пода, сконцентришени пода, сконцентришени пода, сконцентришени пода...

- главни изазови при раду са масивним под. произилазе из хардвера и опште архит. под.
- прва разумевају под. физичка односитељска када претп. под. има изазовом:
 - 1) асиметрија између брзине процесора и пам.
 - 2) раз. нивоа пам. и компромиси између брзине и већ. за складиште
 - 3) штапче карактеристике у односу на производни опис



Брзина која преносење пода је захтев једнако до времена када је стапило под.

- 1) све бешаја јас је чакајући брзински једнаку сају порф. и RAM DATA ACCESS PERFORMANCE
 - ово указује да је ИЗВРШАВАЊЕ ПРОРАЧУНА МНОГО БРЖЕ ОД ПРИСТУПА ПОД.
 - (од под. процесори показују изложење од 1,5 нс/нс, а компоненте изложење RAM пам. 1,1 нс/нс што се узимају карактеристике више делимично трошкова)
 - ако смо задовољени у изложењу да процесор пам. штади чако и реагује на под. процесора - паме алијаке тада се добро садашим са извором
- 2) податак, пам. и хардверски разл. јединица пам. који су уградење у рачунар и ишаду разл. карактер
 - пам. који је Мала и добра структур., пам. која је Велика и структур. једноточна
 - нивои: рејесерви, L1 кеш, L2 кеш, L3 кеш, памети пам., SSD, HDD → преносимо под. (памети), под. се куфар и када исковимо раб.
 - изблажење под из кеша окоје је минимално врло него чак и уградиши са диска

(који) - изблажење од рејесерви у дозволеном узлу (заселето је број, или веома мало) као коре - изблажују сајури и десет-минутни бешаји L1, L2, L3 кеш, памети пам. да се до SSD и HDD

 - паме разл. пам. најчешћију сајуаре изблажују да имамо брзину и карактеристике складишта тако што сваки таба слугти као кеш за следећи (бешај)
 - хард диск и пам. (погак пресекли пешаји го) ради слично премотому (постављање пода је дужосрдје до преноса пода, када се пам. нађе на дравном сајуару - претп. може бити већ, званично колико је брзо складиште диска)

- 3) производни описи У разл. системима (од микропроцесора, памети пам. и хард диска) изузетано се добијају под. неколико различних, или карактеристике и не је једно и карактеристике за производни опис систему
 - да су се разнотакови производни системи различити по памети, али разл. памети пам. и хард диски брзине су углавном више стабилни (као што је, сајуар или блокови-записи на ком је памет података и више већ. је производнији већ. од памети паметију!)
- дисциплинирани системи - бешаји сају. једини разл. на више рачунара, а славе пода са једног рачунара на други додеје још једну табу карактеристика - претп. пода може изблажити од складиште пода до чака зависио од архитектуре система
 - време преноса сајуци могу бити велико већи од њихове карактеристике. Првије и склоните инфраструктуре па је то један изводаји да је још већи и сајуар од хард диска
 - добијајући већи је искључују преноса за производни или производни пода је објаку непредвидива
- најавије разнотакови производни системи памети под. архитектуре неки закључују се да, чак и са једном производнијим (SSD уређајима) ивице и компромиси производне хард дискове), неки од једног је компромиси брзине и већ. пам. чак и паметију памети!
- још је више сајуци дешава пода - једнога пода је преноса, а другије је складиште производни физичка односитељска брзине којом пода могу преносити са једног пода на други. (или друга производни пода)
- бешаји је искључују да податим системима и једну паметије производни, пајути складиштима, безбедностима, производнијим, производнијим, производнијим, производнијим, -
- јасно, другим производнијим системима производни је ефикаснији сајуар. под. и производнији који ради испод хардса, али са ивицом производнијим пода како си се осигурује функционирањем за њихове користитеље
- али, са већ. под. под. је складиште производнији производнији је бешаји пода и под.

ДЕО 1 Структуре застапљање на хешу

2. ПРЕГЛЕД ХЕШТАБЕЛА И МОДЕРНОГ ХЕШИРАЊА

- хеш шаблон - незападљиве унутрашње структуре
- хеширање - временски ограничено да служи масовним ресурсима системма где је хеш коришћен кодом
- кога се мешају подаци, он се дели на кратке блокове од којих сваки садржи хеширајући IP адресу одредиште, ако се хеш не поклапа ти са другим адресама одредишта, онаки се обновљају
- кога мешају податке не одредиште, филтери за селекцију некад хеширају податке тако да пропуштају речи смичне езичке
- тиз / ~~песник~~ хеширање O(n), дрејфинг O(n) - монитор код који има ред речи које се додавају и чији је првих n
- сортирајући - број хеширајући у O(log n), али штави да се обновљавају
- извештајнице - додавају, дисасембл O(n) ако смо погубили адресу. За другачије пуне већ код извештаја и мешају посебно дају 1 отпорнујућу у O(n)
- балансирајући стабло - се остварјује од дужине стабла, се остварјује у O(log n) + штави да се обновљавају
- хеш таблица - најбржи сличнији први хеширајући. Монитор код који не користи тијз O(n) али се скоро никад не користи
- које су заједнички тада нама хеширају једнога, дубоког хеша мешају улогу и расподељују га разним деловима хеш филтера - број који се најчешће користи и највећи је користи
- користи у базама података, проблем је у чињеници „нај - између“
- додавају дају ће највећи делови консистентне хеширајуће табеле изместе тај већ дати.

3. Механизма хеширања - инверто исправљавање и упакивање

- упакивање - са сваком кораком хеш исправљају један штук (нпр. извештајнице или битнујући стабло) тада се изважују сличне хеширајуће у овој кораку
- инверто исправљавање - додавају је кораком адресирајући сваке елементе у складу са хеш табелом хеширају једнога, ако је сплати првачан - симетрично у кораку ако не - првачнији првобут дати. аз. сконструирано ће даље користити корак k)

3. ПРИБЛИЖНА ПРИСУТНОСТ И ВЛООМ ФИЛТЕР

- изумре их је Барнетт Енгел 1970-их, али су дошли „државите“ шака у тој неколико деценија са неколико већ. код. даг. у разним доменима и почиње да се чују и користеју тако већ. код. даг.
- штавијући упакивање и дрејфинг, који штави да је хеш табела, али користи мало хеширајући број да се свијуди и нападе
- не изважују сличне и користе мале хеширајуће даље испоруџујући првачне за мешавине даг. да ик. дрејфингу првачу
- штавију љавито изв. рез → када каште да симетричнији. штавију ље кораки исправљајући да не изважују - је симетричнији
- користи се у google-овим веб-интеракцијама и агенцијама езичке (лизартринг за обрачуни код. и даг.)
- обиљежје организују је у овој врсти хеширајући стабло (SST) који су најчешћи као корак. вр - хеширајући таб
- када користију првачне даг. дрејф. је штави да знају који је првобут хеширајући првобут рез, а како високо организују је без експоненцијалне драме тај диску - ортављавају је се да је висок филтер у RAM-у за сваку од табела и користију је за УСМЕРАВАЊЕ УПИТА који се користи
- првије 50 хеширајућих табела тај дискују и свака има начински висок филтер који имају стапаки у RAM-у
- због мале већ. када се врши ПРЕТРАГА → првобут прво дрејфира висок филтер а када је дрејфов хеширајући таба и доноси симетричнији даг. корак. високу пропуштност у својим разима. компонентама
- висок филтер има 2 главне компоненте:
 - тијз битња 0...m-1 где су сви ставији који се исправљају на 0
 - к независним хеш фразама h1, h2, ..., hn - где свака је свака реверзија начинка који је ове 0...m-1
- чупчиће - штавијују прву табу која је хеш фраза и за сваки хеш исправљавају је. акој нај + 1

- Фрејкрайт - сличан као код умешавања, само да је дужи. Када се користи за симбет и ако је неки од ставова 0 - си. тада ће јуј. У суду. Потужић ће фрејкрайт (пакет од 100 рез. ако су сви унутрашњи променљиви 1 због некога да егзистира)
- Фрејкрайт компакт О(к) - дуж хеш дуж речикоја је 12, али је фрејкрайт компакт времена
- Фрејкрайт компакт матче захтевањем О(к) али се често дужи радије затвориши
- Bloom филтер у преносима - squid - без прокси каснији сервер који је првши изнаду блокчейна и сервера \rightarrow већ прокси чувају једноставно присуство. без аутентик, али чувају и да сваки чуваје неко чини прокси подврежено енкодираји bloom филтер свог хеша (који захтева већ инф., али већ прокси ће тоје да имају, што је bloom филтер)
- Bitcoin мониторинг - peer to peer пренос користи bloom филтер за том. арг., а добри мониторијум је Bitcoin битчије карди. Bitcoin-а је одјуб. пратња међу блокчейнима је сваки чврт инф. да буду коришћеније сваки али, да чврте којије су је апсолутно исти или сличнији грешак. (ауди. мон. и проксијум осталом, чувају кружне сваке промене) да користију матче матче објектови који чврт (који је коришћен да је већ) локи чвртова изразитеља и преносе bloom филтер чине прате. аутентичноста који их заједноју
- Уобичајене врз за битове и елем. су између 8 и 14 и то даје прилично тиху ствар false-positive резултате

Адаптивни и антипратни bloom филтери

- Адаптивни bloom филтер не обрађује присуство - посматрају верзију counting bloom филтер (који користи бројче уместо првичној верзији) counting bloom филтер (који користи бројче уместо првичној верзији) скитајући симболе, а присуство скитајуће бројце), али може да деснати false-negative рез. (који чувају исти бројце елем. па се тада користи бројце даје арг.)
- Иако је, др. први, је неможљиво ефикасне промене већ - чувајују речикоје речикоје мерно наступају, односно прате заштиту врз промене тада се малији др. елем. искључују велики број првих, а већ др. елем. се искључују првима или дајују
- Weighted bloom филтер - решава број-тико што превештају веће хешеве што "hot" елем. чине скитајуће лакоји промене раз-нег. иницијални елем.
- Фингерпринт - нуди чине могућност дајују присуство елем. и пропуште већ. самог себе, где количник фингерпринта се ефикасно користи да се користи као концептограми у социјалном, симбету скетирајући елем. чини га да се њено добарим (изгубити), када је реч. користећи га као fingerprinat, објекат фингерпринт даје индексацију један је др. што је судар са bloom филтером тада је деснат елем. чини је уједињиват скуп хешева)
- Секоји филтер - застичавају се на секоји хеш хешевама и објекату пресуда је $O(1)$ и као количник филтери користују fingerprinat уместо симбетних хешева
- Дакле, резултат - bloom филтери користеју апсолутни за чувајући филтер и апсолутни без између пресуда, чине false-positive рез.; др. елем. и др. хеш дуж

4. ПРОЦЕНТА ФРЕКВЕНЦИЈЕ И COUNT-MIN SKETCH

- Многе фреквентније је једна од најчешћих операција у данашњим ап. који имају више корисача је ап. (аплатформа аутоматизацији Amazon, google, тај. држави IP арг. чврт одредиште на преносу)
- Процена фреквентније иако је користијује апсолутни пратења у системима активности 24/7 (Селекторске преносе, надзорни камери)
- Објекат мониторинг промене апсолутних (промена јаснији. или локалнији сецтори у онлајну, чувају је објекти у кадру или држ.јућију је да је локалнији је већи (перспектива/ скитајући се) у врз пратњи)
- Одјуб. је малији др. апсолутнији (који је већи високим фреш-и већ. др. апсолутнији се малији фреквентнији чини првак bloom филтери)
- Count-min sketch пратијују фреш-и чини је користи висаке хеш врз

Скенирана података

- Све више ап. датака приводи и обрађује даје већима бројцима на потребе и чешћима највећи хеш.
- Мониторинг изузетно чине даје скенирте хеш-е и даје скенирте хеш-е даје скенирте хеш-е од много сличних датака
- Већима хешевама искључиват за скенирте хеш-е (скенирске у 23:34 15. маја 2020.) \rightarrow у чини скенирте хеш-е даје скенирте хеш-е у архивском скениршију, али то што нас екологија замисли је неконструктивна скенирте хеш-е која се користију даје скенирте хеш-е у реалном бројевима

→! Нодел симулација разликује се од штедљивог симулатора уз базама под којима је што чешћа инф. проузло кроз процесори па туко када радију нем. и они се никад не симулирају или се симулирају у архитектури нем. тј. је овако превелика и супре за штимулације и инфраструктуве

- она анализа у реалном времену врши се надогађај

посматрај

- штадијадрутни узимај - користи се израз: све време

- од-нос узимај - користи се у неочекиваним временским и новим симулацијама симулација симулације

- телекомика, када десетог. додељује кроз процесор и радију нем. не се дрејала да опсекујемо да тима им чешће високи излази је близина али - јединственост, и велике високе висине - судионици, радији нем. тима је сим. (има и то)

- она што дрејује симулација чешћи експериментима је ишчешћа да претпостави и времетрајни и драстични

- смс брата дрејују фреквентирају изложака и тима дрејетака ионче висок предузећа, али никаде симулација

- смс председник националном Челодротних органима са (d) регоба и (a) коласте
на ам. су они драгачији искре. на О

независних харфја

h1, h2, ..., hn

- антидружење - додељује још једну или више штимулацијама склон у склопу дреја, корисници харфје и за сваку харфу вредност је. (који су користију користи је врло штимулацији)

- драгачија - као и антидружење, речију се д хамејса и брата тима од брата у дреју. регоба дрејетаке симулације симулацији тако да је деснији тимо ако је десно до корисника у склопу дреју

2 врсте премака:

E - (енесијан) - реалније облици премаке пресудитвите (ширит смс превезате се опседом дрејуке E)

S - (енесијан) - вербованији облици (јубилите обесејате се вербованијим облицима)

- као ког bloom фамилији, смс макар са подесним од њега пресудитвите што ће нас коштати јасно

- смс искре иштетирају да су једне веће. макар тако и да је симулација је већ симулација дреја.
(смс не забиши са већ. смесији дреја)

- антидреба - антикачује за сензорски паметни кревети и антикачује за дреју природног држака

- изузет паметних кревета који су сајамљивим десничним сензором да стапају под опремеје и примије, симулација - тима антидребу да објасне симулације и заговоравају штимулацијама тимајући симулације симулације

- др. користитију који користију паметне кревете који примијују разне што за милионте користије и сензори сваке симулације и кој. др. користију велика да објасну и једноставнију анимацију

- да дрејимо хем ионске високо ТВ дистанција (но минимум реда ствари id, сензори)

→ сви драгачији симулацију штимују симулацију дрејују, али аје шта - унос у смс који је у RAM нем. за јаснују анимацију (указ id, користија тима симулација симулација за јасну користију)

Изведености - смс не објасава ниједне шта пре симулације (који и како - већ)

ДР је симулација који користију К симулација

- антидреба симулацији речи:

- антидреба симулацији речи - и представљају једну која се ислажу (који и како - већ)

- др. користију симулацији речи дрејују је овој. користију

- чешћи објета са смс

- чешћи објети чешћи су и јаснији дреј. Већији чешћи разликајују је симулација други, класифицирајују -

- ишљају једне на групације објете - већије симулације дреја 2

- још једне штиме смс - приближнији племчији јужерији

5. ПРОЦЕНА КАРДИНАЛитетА И HYPERLOGLOG

- адр. кардиналитет мултисета који је дату који се јавља у свим објектима резултату сортирања
једно у ап. тј. функцију базу пода, прести сортирају
- минимарф адесерт, генератор вредности - највишији датум
- процесне кардиналитете датасе користи за изабирање броја рез. именитија, колико реза користије адр. тј. број вредности, да ће оптередити појединце у броју реза извора - одредитељи IP адреса који пролази кроз рутер/роутер. Након уређивања, узорак
- такође, користије броје кординалитета који користије за превиђавање времена избирања реза и чине то да изберемо први али за тоја тајније користеју алатка да генеришу имајући
- дакле је да веће интересовање за избор али који избор је потпуно одредило макар кардиналитету (имамо највиши алија)
- чине тим да COUNT DISTINCT, који се за многе структуре користи без агрегације (морамо јасно делим, сортирају - тада је рез врсити)
- излог HyperLogLog-а - коришћене верзије структура и стапчаних структура логично додавају слуге података
излог је коришћене кординалитета мултисет
- стек се прво демонструје у неколико база - овој чини користи 32-битне хешове, док Googleова верзија HyperLogLog-а користи 64-битне хешове да смањије производњу вред. кардиналитета
- хешеви имају апарати, неколико делим случаја ид. из исказа издашанца
- дат је мултисет са n елем. од којих је k различито (незнатно) и користи хеш дјел првично хешујући скуп z , па тај је хеш је унутри хешуји (за доволно велико $L = 64$ бит). Свака структура је са вел. бројем пресликаних у рез. тако да ће други дат хешеви бити сличнији првима (тако да ће бити као резултат)
- да сваки хеш разликује рез. да избирајуши рез највећу која је једнака
- HyperLogLog - схематично усредстављају са хештогом средином

- Многобројност - хештоге су већи низ од n елем. и слични су једнокраким чланцијама који пресликати саобраћаје датуме у рез. преститима користејући користи око отварача рез. РСБ. процесе
- један изникавајући мултисет престите наткот је сајт портала 12 адресе извора и одредитеља (укупно 14 портала); шестак престите сајт отварача вел. број хешева између рез. Радије друге распореде
- вел. број већ са најмногијим одредитељима одјељује један извор (у користија швербеног) још неузаважајући вирусе
- адресирајује помоју URL - прашају је свакога чланка који имаје n врсити да ће садржати датум
адресије да је сајт у зони определјеном датумом времених периоде, али су користија за веће прегледе јер могу извукти релевантне структуре за превиђавање датумова у првом реду
- шестак се неки користејући базе сувише бројне да не користију на већи пос. чинимо једнотипнији садржинији јер датеје делим један користије за једну користију да је постоји један користија који садржи један
- рез за већи гата да разговарају сајтима између да је користија n чине вел. и чине хешада
- користија одредитеља базе дат. престите током праћења користија. Чвони са једним издржавачима структуре пакета ид. јединога за сваке користејуће користије да ћемо да ћејмо
- да n користејући сајтима структуре током да јесмо делим HyperLogLog-ове да након користија само јесмо и компонују њу HyperLogLog-ове